Corpus oraux, guide des bonnes pratiques 2006

Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, Laurence Goury, Michel Jacobson, Isabelle de Lamberterie, Christiane Marchello-Nizia, Lorenza Mondada

▶ **To cite this version:**

Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, et al.. Corpus oraux, guide des bonnes pratiques 2006. Presses universitaires d'Orléans; CNRS Éditions, 203 p., 2006, 2-913454-30-5 : 2-271-06425-2. hal-00357706

HAL Id: hal-00357706
https://hal.science/hal-00357706

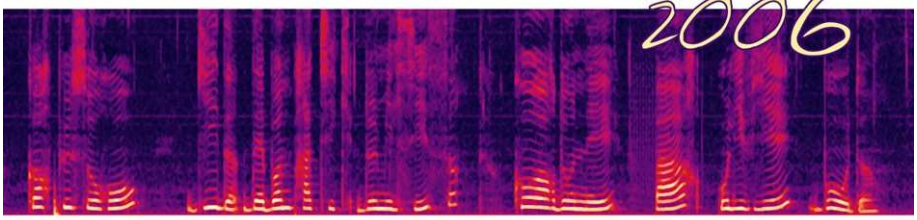
Submitted on 1 Feb 2009

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 ORAUX

Guide des bonnes pratiques 2006



coordonné par **Olivier BAUDE**

PRESSES UNIVERSITAIRES ORLÉANS

CNRS EDITIONS

CORPUS ORAUX

Guide des bonnes pratiques

2006

CORPUS ORAUX

Guide des bonnes pratiques 2006

coordonné par **Olivier BAUDE**

PRESSES UNIVERSITAIRES ORLÉANS

CNRS EDITIONS

Délégation générale à la langue française et aux langues de France
6, rue des Pyramides75001 PARIS
http://www.dglflf.culture.gouv.fr

ISBN 2-271-06425-2 (CNRS ÉDITIONS)
ISBN 2-913454-30-5 (PUO)
EAN 9782271064 257 (CNRS ÉDITIONS)
EAN 9782913454 309 (PUO)

© Presses Universitaires d'Orléans / CNRS ÉDITIONS

Cet ouvrage est le résultat des travaux d'un groupe de réflexion
réuni autour d'Isabelle **de LAMBERTERIE**.
Il a été coordonné par Olivier **BAUDE**.

Olivier **BAUDE** (*DGLFLF et CORAL – Université d'Orléans*)
Claire **BLANCHE-BENVENISTE** (*EPHE et Université de Provence*)
Marie-France **CALAS** (*DMF*)
Paul **CAPPEAU** (*Université de Poitiers*)
Pascal **CORDEREIX** (*BnF)*
Laurence **GOURY** (*CNRS – CELIA*)
Michel **JACOBSON** (*CNRS – LACITO*)
Isabelle **de LAMBERTERIE** (*CNRS-CECOJI*)
Christiane **MARCHELLO-NIZIA** (*CNRS-ILF et ENS-LSH-Lyon*)
Lorenza **MONDADA** (*ICAR, CNRS, Université Lyon2*)

Avec la collaboration de :

Gilles **ADDA** (*pour le collectif COPTE LIMSI-CNRS*), Michel **ALESSIO** (*DGLFLF*), Alain **CAROU** (*BnF*), Ibrahim **COULIBALY** (*CDF – Université de Grenoble*), Valérie **GAME** (*BnF*), Fabrice **MOLLO** (*CNRS-CECOJI*), Michel **RAYNAL** (*INA*), Jean **SIBILLE** (*DGLFLF*), Dominique **THERON** (*BnF*), Luc **VERRIER** (*BnF*).

PRESENTATION DES AUTEURS

OLIVIER BAUDE
Maitre de conférences en sciences du langage à l'Université d'Orléans, membre du Centre Orléanais de Recherche en Anthropologie et Linguistique (EA-3850). Secrétaire du conseil scientifique de l'Observatoire des pratiques linguistiques, Délégation générale à la langue française et aux langues de France.

CLAIRE BLANCHE-BENVENISTE
Professeur émérite, École Pratique des Hautes Études à Paris et à l'Université de Provence. Recherche dans le domaine de la linguistique française : langue écrite et langue parlée, syntaxe, morphologie, constitution de corpus de langue parlée.

MARIE-FRANCE CALAS
Conservateur général du Patrimoine. Inspecteur général des musées, Direction des Musées de France. Spécialiste du domaine sonore, compris comme un vaste domaine pluridisciplinaire incluant l'histoire, la gestion, la conservation et la valorisation des enregistrements parlés, musicaux, des sons de l'environnement, aujourd'hui partie intégrante du patrimoine immatériel.

PASCAL CORDEREIX
Conservateur en chef des bibliothèques. Chef du service des documents sonores au département de l'Audiovisuel de la Bibliothèque nationale de France ; par ailleurs vice-président de l'Association française des détenteurs de documents audiovisuels et sonores (AFAS). L'essentiel de son activité est orienté vers les questions d'archivistique du son.

LAURENCE GOURY
Chargée de Recherche à l'IRD (Institut de Recherche pour le Développement), membre du CELIA (Centre d'Etude des Langues Indigènes d'Amérique), linguistique de terrain et typologie (en particulier langues créoles).

MICHEL JACOBSON
Ingénieur informaticien au laboratoire des « Langues et Civilisations à Tradition orale » du Centre National de la Recherche Scientifique. Co-responsable du programme « Archivage ». Spécialiste de la gestion de corpus oraux.

ISABELLE DE LAMBERTERIE
Directrice de recherche au CNRS, responsable de l'équipe « Normativité et société de l'information » du Centre d'études sur la coopération juridique internationale (CECOJI – UMR 6224), membre du Comité d'éthique du CNRS.

CHRISTIANE MARCHELLO-NIZIA
Professeur en Sciences du langage à l'ENS-LSH (Lyon), Directrice de l'Institut de Linguistique Française (CNRS) : Linguistique historique, histoire du français, théories de l'évolution des langues.

LORENZA MONDADA
Professeur en Sciences du Langage à l'Université Lyon 2 et membre du Laboratoire ICAR (UMR CNRS 5191). Travaille en linguistique interactionnelle sur les corpus de langue parlée en interaction ainsi que sur l'analyse multimodale de corpus vidéo.

PREFACE DE XAVIER NORTH,
DELEGUE GENERAL A LA LANGUE FRANÇAISE ET AUX LANGUES DE FRANCE

Rares sont les moments, dans l'histoire des sciences ou des politiques culturelles, où un ensemble de données brutes et de matériaux incertains se convertit en objet de savoir. La publication de ce guide est de ceux-là, puisqu'il offre à tout chercheur les outils, les « bonnes pratiques » qui lui permettront de procéder à cette métamorphose : la transformation de productions verbales en un corpus oral, susceptible d'être étudié et conservé, et par conséquent de prendre place dans le patrimoine culturel de la nation.

Sans doute les productions langagières dans leur forme *écrite*, fixe et définitive, d'œuvres littéraires ou de documents d'histoire, n'ont-elles jamais cessé d'être au cœur des politiques mises en œuvre par le Ministère de la Culture, qu'il s'agisse du livre ou des archives. Mais ce n'est que tout récemment qu'on s'est avisé de porter intérêt à l'aspect vivant du langage dans son jaillissement spontané, dans son énonciation quotidienne, ordinaire, et dans l'extraordinaire variété de ses parlers… Pour la première fois, s'esquisse ainsi la possibilité de constituer, sur des bases assurées, de véritables archives de la parole. Le progrès des technologies devrait y contribuer.

Un corpus oral, en effet, n'est pas une simple collection d'enregistrements de la parole humaine, c'est un objet « construit » : le traitement des données (numérisation, transcription, indexation) permet non seulement de les conserver, mais les fait passer à un statut nouveau, matière de recherche et de valorisation. Encore faut-il pouvoir s'appuyer sur des prescriptions de méthode, cohérentes et faciles à mettre en œuvre.

Grâce au « Guide des bonnes pratiques », c'est un nouveau et vaste domaine qui s'offre désormais à la curiosité des chercheurs. Par l'intermédiaire de son *Observatoire des pratiques linguistiques,* la Délégation générale à la langue française et aux langues de France a donné l'impulsion de départ, puis s'est attachée à regrouper et à coordonner les énergies et les ressources diverses qui ont produit ce travail, qu'elles proviennent du monde de la recherche, ou des différents horizons du Ministère de la culture concernés par cette initiative.

Assurer le développement des corpus oraux, leur diffusion et leur conservation, c'est aussi rendre accessible, donner à entendre le patrimoine linguistique français dans sa diversité, sa richesse et sa vérité. C'est encore se donner un outil précieux de connaissance des pratiques langagières nécessaires à la définition des politiques de la langue, mais aussi des politiques éducatives et sociales.

Pendant plusieurs mois, cette démarche a rassemblé juristes, linguistes, conservateurs et informaticiens dans la volonté de concilier les chemins nouveaux de la culture et de la recherche avec le respect du droit. C'est le résultat d'un effort commun de pensée que nous présentons aujourd'hui, avec l'espoir qu'il féconde à son tour de nombreux travaux.

PREFACE DE BERNARD MEUNIER, PRESIDENT DU CNRS

L'oral et l'écrit. Ces deux mots possèdent une force évocatrice puissante. Nous pensons à la manière dont les civilisations se sont structurées par les pratiques orales et ensuite par la création d'écritures permettant de mieux transmettre dans l'espace et dans le temps les paroles des uns et des autres.

Mon regard de chercheur sur le rôle respectif de l'oral et de l'écrit dans la diffusion des connaissances scientifiques ne me fait pas oublier que, bien au-delà du rôle primordial de l'écrit, la présentation orale devant ses pairs, ou un large public, est toujours essentielle pour diffuser, convaincre, faire partager des idées. L'oral garde une force de conviction, permettant d'atteindre le plus grand nombre dès lors qu'il peut être enregistré et transmis à l'aide des moyens audiovisuels actuels.

La collecte et l'utilisation des corpus oraux doivent se faire selon le respect de « bonnes pratiques », comme cela se fait pour celles des corpus écrits. Nous savons tous combien une phrase, sortie de son contexte et diffusée sans retenue, peut devenir dangereuse pour son auteur, un groupe de personnes ou une communauté.

Les auteurs de ce remarquable travail ont abordé en profondeur tous les aspects juridiques de la collecte et de l'usage des corpus écrits. Je souhaite que cet ouvrage bénéficie de la meilleure diffusion auprès des acteurs et des utilisateurs des corpus oraux que nous sommes tous, à un moment ou à un autre.

PREFACE DE JEAN-NOËL JEANNENEY, PRESIDENT DE LA BIBLIOTHEQUE NATIONALE DE FRANCE

La Bibliothèque nationale de France est heureuse d'avoir contribué à l'élaboration de ce *Guide*. Elle entretient, en effet, un rapport ancien et étroit avec les langues parlées, leur préservation et leur diffusion. Son département de l'Audiovisuel est l'héritier des *Archives de la Parole* de Ferdinand Brunot, créées dès 1911. Depuis cette date, notre établissement s'est constamment préoccupé d'assurer les meilleures conditions de captation et de conservation des expressions orales de toute sorte, comme de leur diffusion auprès du public le plus large.

Aujourd'hui, les technologies numériques renforcent ce lien historique et scientifique. En matière de conservation, un plan ambitieux de numérisation de nos collections a été engagé dont les documents sonores et audiovisuels bénéficient en particulier. D'autre part la diffusion de ces richesses dans nos murs et à distance est servie par l'essor spectaculaire de notre bibliothèque numérique en ligne, « Gallica », qui permet à chaque internaute, où qu'il se trouve et quel que soit l'objet de sa recherche ou de sa curiosité, d'accéder à ces sources fondamentales de la connaissance.

Fruit d'une confiante collaboration, ce *Guide* témoigne de la complémentarité des savoirs entre linguistes, juristes, conservateurs, informaticiens, techniciens du son et de l'image : je me réjouis que la Bibliothèque nationale de France ait contribué à cette entreprise novatrice et féconde.

Cet ouvrage applique les rectifications de l'orthographe, étudiées par le Conseil supérieur de la langue française (1990), et approuvées par l'Académie française et les instances francophones compétentes.

Les mots signalés par un astérisque renvoient au glossaire juridique situé en fin d'ouvrage.

1 Présentation
- 1.1 Les objectifs
- 1.2 Les conditions d'élaboration
- 1.3 Les aspects juridiques
- 1.4 Les autres aspects
- 1.5 La méthode
- 1.6 Le cadre juridique français
- 1.7 Un « guide des bonnes pratiques » ?
- 1.8 Quelques questions fréquentes

2 Le contexte
- 2.1 La linguistique et les corpus oraux
- 2.2 Cadres politiques de la diffusion de la recherche
- 2.3 Cadres juridiques

3 La démarche
- 3.1 Expliciter la démarche
- 3.2 Éléments de la situation en jeu
- 3.3 Pratiques de terrain
- 3.4 Anonymisation
- 3.5 Transcription

4 Les corpus oraux, objets de patrimoine ?
- 4.1 Rappel de la situation
- 4.2 Les initiatives privées
- 4.3 L'accès aux collections

5 Annexes
- Fiches juridiques
- Fiches techniques
- Institutions
- Travaux

1 PRESENTATION

1.1 LES OBJECTIFS

Il existe actuellement quantité de recherches fondamentales ou appliquées, qui se fondent sur l'exploitation de « corpus oraux » (collections ordonnées d'enregistrements de productions linguistiques orales et multimodales). Issu de la prise de conscience de linguistes soucieux d'assurer la pérennité des sources et un accès diversifié aux documents oraux qu'ils produisent, ce *Guide des bonnes pratiques* aborde en priorité les « corpus oraux », créés et utilisés par et pour des linguistes. Mais les questions soulevées par la création et l'exploitation documentaire de ces corpus se retrouvent dans de nombreuses disciplines, l'ethnologie, l'anthropologie, la sociologie, la psychologie, la démographie, l'histoire orale notamment utilisent l'enquête orale, le témoignage, l'interview, le récit de vie. Fondé sur la démarche des linguistes, ce *Guide* recoupe toutefois les préoccupations d'autres chercheurs qui utilisent des corpus oraux (par exemple en synthèse et reconnaissance de la parole), même si les besoins spécifiques de ceux-ci ne sont pas systématiquement abordés dans le présent document.

Le *Guide* que nous vous proposons s'est fixé pour premier objectif de fournir les *informations* nécessaires à la constitution de corpus de données orales ou multimodales, et d'offrir des *propositions* utiles concernant non seulement les aspects juridiques, mais aussi les aspects matériels touchant aussi bien à la collecte, qu'à la structuration et la mise en forme des données, qu'à l'exploitation, la communication et la conservation de ces données.

Le second objectif de ce *Guide* est d'aider les chercheurs qui constituent ou enrichissent des corpus oraux à *anticiper* certaines « difficultés à retardement » qui risquent de grever lourdement l'exploitation puis le devenir de leur corpus. Certains choix initiaux, certains manques peuvent révéler leur importance à des étapes ultérieures du processus, alors qu'il est trop tard pour modifier quoi que ce soit.

Le troisième objectif est de favoriser l'émergence de *pratiques communes*, afin de satisfaire aux exigences actuelles de conservation et d'interopérabilité des corpus, d'évaluation, et d'éthique tant dans la constitution que dans l'usage des données.

1.2 LES CONDITIONS D'ELABORATION

Le conseil scientifique de l'Observatoire des pratiques linguistiques (Délégation générale à la langue française et aux langues de France) a souhaité encourager fortement les actions de conservation, de constitution et de valorisation des corpus oraux et multimodaux pour les raisons suivantes :

- permettre la sauvegarde d'un riche patrimoine sur les pratiques linguistiques en France ;
- aider à la constitution de grands corpus de référence, pour la recherche, l'enseignement, les industries de la langue mais aussi le patrimoine ;
- aider au développement des outils informatiques de traitement, d'enrichissement et de valorisation des corpus ;
- favoriser la mise à disposition de ces corpus.

1.3 LES ASPECTS JURIDIQUES

Très vite il est apparu que les aspects juridiques liés à la constitution et à l'utilisation des corpus oraux constituaient un obstacle récurrent et capital.

Ces aspects juridiques concernent principalement les questions de droits moraux et patrimoniaux et de propriété des données, que l'on retrouve à chacune des quatre grandes étapes du travail sur corpus :

- le recueil des données et l'enregistrement (droit à l'image, à la voix, situation d'enquête, autorisations…) ;
- l'utilisation et l'exploitation informatisée des données (archivage, base de données à des fins de recherche, d'enseignement, d'ingénierie…) ;
- la diffusion et la mise en circulation des données (droits, droit de citation, diffusion en ligne…) ;
- la conservation des données.

Au vu du grand nombre de domaines concernés, la DGLFLF a suscité la création d'un comité composé d'experts de diverses disciplines. Ce comité a instauré un *groupe de travail* ayant pour objectif d'aider les équipes de recherche à normaliser les pratiques de recueil et d'exploitation de corpus au regard de la législation en tenant compte de l'ensemble des contraintes liées à la recherche. Le guide que nous présentons ici est le résultat d'une quinzaine de mois de travail de ce groupe.

Ce groupe de travail devait évidemment comprendre des juristes spécialistes du droit de la recherche, mais pas seulement : la nécessité de compétences en termes de constitution des corpus, d'utilisation et de conservation ont conduit à adjoindre aux juristes des linguistes pratiquant de la « linguistique de corpus » et travaillant sur des données orales, des représentants des grandes institutions de conservation patrimoniale (INA, INSI, BnF) et des informaticiens spécialistes en gestion de corpus.

Pour remplir sa mission, ce groupe de travail s'est donné pour objectifs notamment de :

- recenser les pratiques actuelles et définir en priorité les contraintes méthodologiques et théoriques liées à la recherche ;
- diffuser une synthèse sur la législation existante ;
- établir des recommandations ;
- et, le cas échéant, en cas de vide ou de flou, formuler des propositions pour l'élaboration de normes et règles juridiques (notamment européennes).

Il fallait pour cela tout d'abord :

- recenser les domaines juridiques concernés ;
- identifier et quantifier les risques ;
- repérer les réponses existantes ;
- et ensuite construire ces réponses sous la forme d'une série de recommandations de bonnes pratiques (juridiques et éthiques).

Pour cela le groupe a décidé de travailler en étroite relation avec plusieurs équipes témoins pratiquant ou ayant pratiqué le recueil de données orales ou audio-visuelles.

Le but était de parvenir ainsi à une « typologie des situations », et de faire le tour de toutes les pratiques et solutions déjà utilisées, tant en France qu'ailleurs.

1.4 LES AUTRES ASPECTS

Chemin faisant le groupe de travail s'est aperçu que proposer uniquement une série de recommandations ou de solutions de nature juridique ne permettrait pas de répondre de façon satisfaisante aux difficultés rencontrées.

Il est en effet apparu que bien souvent la difficulté ou la solution étaient liées au type de pratique de collecte ou d'utilisation ; que certaines solutions passaient par des voies techniques qui avaient un retentissement sur les données elles-mêmes (anonymisation ou floutage) ; qu'il n'était pas indifférent de résoudre tel ou tel problème juridique à tel moment plutôt qu'à tel autre. Bref, proposer des solutions à des questions juridiques revenait à évoquer le processus même de collecte ou de mise en forme, de transmission ou d'utilisation de ce type de données.

Enfin, au-delà du respect dû aux droits des personnes enregistrées, s'est posée la question du « droit d'auteur » de ce type de données : quels sont les droits des collecteurs de ces données ? Qui en est juridiquement responsable, qui a le droit de les transmettre ? Sous quelles formes ? Comme on le voit, les aspects juridiques liés à la propriété scientifique ou à la responsabilité pénale étaient, eux aussi, indissociables de la pratique de recueil et d'utilisation des données.

Dès lors, ne valait-il pas mieux élargir la compétence du « Guide » projeté, et évoquer non seulement les pratiques juridiques, mais aussi l'ensemble des pratiques mises en jeu dans ce type de corpus ? C'est le choix qui a été fait, car cela permettait de maintenir liés tous les aspects, tels qu'ils le sont dans la réalité.

1.5 LA METHODE

La méthode à laquelle s'est rallié le groupe de travail se caractérise par les traits suivants :

- la conviction qu'il ne faut pas laisser croire qu'il existe des réponses toutes faites à tout type de situation ;
- la volonté de ne pas « brider » les chercheurs (en interdisant certaines pratiques par exemple) ;
- le respect de la méthodologie du chercheur et des contraintes liées à l'observation (les chercheurs souhaitent enregistrer des situations sans que les contraintes, notamment techniques et juridiques, les modifient).
- la nécessité d'élaborer et de rédiger ce guide en mettant en commun les compétences requises aux différentes étapes (linguistes, juristes, conservateurs) ;
- l'affichage d'une démarche fondée sur le respect de la loi et de l'éthique ;
- la nécessité de fournir à travers ce *Guide* un outil d'expertise des risques (repérage, mais aussi évaluation).

1.6 LE CADRE JURIDIQUE FRANÇAIS

Un bon nombre de questions et de solutions tournent autour de la notion de *consentement* des enquêtés mais aussi de la responsabilité des instances *propriétaires*. C'est certes un point nodal. Mais il est loin d'être le seul en cause, et par ailleurs les réponses à une telle question se sont révélées complexes.

Les pratiques actuelles de recueil de consentement et d'autorisation sont très variées. Il n'existe pas de normes reconnues, et les difficultés sont multiples.

Tout d'abord, le consentement doit être *éclairé* (cadre, finalités, « risques » pour l'enquêté).

Mais le recueil de consentement a priori peut parfois gêner l'enquête (paradoxe de l'observateur) en formalisant une situation alors qu'on souhaite obtenir des données « naturelles » proches de la conversation familière.

Ainsi, par exemple, une pratique qui s'est révélée intéressante et efficace consiste (en plus du recueil de l'autorisation) à laisser aux enquêtés un document expliquant le cadre, les finalités, les risques, l'accessibilité, et les coordonnées permettant de retrouver ultérieurement les références des publications et des résultats.

La difficulté provient également d'une *contradiction* entre l'obligation d'indiquer les finalités de l'enquête pour éclairer le consentement, et l'impossibilité de prévoir à l'avance l'ensemble des finalités *et les possibilités futures d'utilisation des données, étant donné le souci actuel de parvenir à* une interopérabilité *maximale*.

Il faut noter enfin que certaines cultures orales (et pas seulement à l'autre bout du monde) n'offrent pas la possibilité de proposer et de garder une trace écrite du consentement.

Et toutes les autres questions de nature juridique offrent la même complexité : anonymat, cryptage, floutage, définition des responsabilités, dépôt, communications, etc., toutes pratiques nécessairement liées à la constitution et à l'existence d'un corpus oral. Aucun de ces aspects ne repose sur une pratique unique, définie clairement et partout reconnue.

Et chacune de ces étapes se retrouve intimement liée à des choix techniques, à des pratiques sociales ou scientifiques, tout cela étant difficilement dissociable.

D'où le choix du groupe de travail, d'offrir un Guide qui ne soit pas seulement un « mémento juridique », mais aussi une aide pratique et fiable envisageant tous les aspects du processus.

1.7 UN « GUIDE DES BONNES PRATIQUES » ?

Prenant en compte les cadres juridiques existant en France (et plus généralement dans un certain nombre de points en Europe), ce guide s'appuie sur les questionnements des chercheurs qui ont participé à son élaboration. Ceux-ci ont cherché à comprendre les fondements des règles juridiques applicables et les enjeux liés à leur respect et à leur mise en œuvre. C'est donc une *vision dynamique de la régulation juridique* qui sert de trame à ce guide, à travers la démarche que suivent les chercheurs. Les auteurs du guide, eux-mêmes impliqués sur les terrains de recherche dont il est question ici, ont eu le souci de proposer des pratiques et usages

respectueux des droits existants. Pour cela, la démarche du chercheur doit consister à connaitre l'existence de ces droits et des contraintes qui en découlent. Il s'agira ensuite de tirer les conséquences de ces contraintes tant dans la phase du recueil des données que dans celle de leur valorisation.

Pour présenter de façon rigoureuse et crédible une telle démarche, il faut tout d'abord la situer dans son contexte, que celui-ci soit scientifique, politique, juridique ou institutionnel. Les usages et pratiques proposés seront tout au long « éclairés » par ce contexte, de façon à mieux comprendre quels sont les enjeux du respect ou du non respect de ces usages ou pratiques.

1.8 QUELQUES QUESTIONS FREQUENTES

Le premier objectif de ce guide est d'apporter des informations et des éléments de réponse aux questions qui se posent à tous chercheurs ou responsables de la constitution, de l'exploitation, de la conservation et de la diffusion de corpus.

Pour répondre à cet objectif, le guide a été conçu avec de nombreux renvois qui forment autant de parcours de lecture possibles. Les questions suivantes représentent les interrogations qui se posent traditionnellement au commencement d'un projet de recherche et proposent ainsi un premier exemple de parcours.

FOIRE AUX QUESTIONS

1. *Quelles autorisations dois-je faire signer aux locuteurs que j'enregistre pour pouvoir ensuite exploiter ce corpus et pouvoir :*
   a. le citer dans un travail universitaire ;
   b. le citer dans un article publié dans une revue scientifique ;
   c. le citer dans un ouvrage à diffusion commerciale ;
   d. le mettre à disposition sur un site ;
   e. le diffuser sur CD.

   Ces différents types d'exploitation sont-ils soumis aux mêmes règles ?

   Réponse : Les questions a, b et c relèvent du droit de citation (voir fiche *Droit de citation*). Les éléments de réponses aux questions d et e sont notamment présentés dans les chap. 2.1.5, 2.3 et 3.5. (voir fiche *Consentement* et *les exemples d'autorisations*).
2. *J'ai fait un enregistrement de personnes que je connais bien.*
   a. A quelles conditions puis-je l'exploiter ? (exploiter est pris au sens de la question 1)
   b. Peuvent-elles revenir sur leur autorisation ?

   Réponse : Tout le chapitre 3.4. est une réflexion sur les conditions de recueil des données, qui veut sensibiliser aux problèmes nombreux qui peuvent se poser au cours du recueil. Le fait de bien connaitre les personnes concernées ne diminue pas les exigences juridiques (qu'il faut a à leur égard), au contraire (il pose des questions de confiance qui peuvent donner lieu à des situations assez complexes). Voir fiche *Consentement*.
3. *Lorsque j'enregistre des enfants,*
   a. qui peut donner son consentement ?
   b. lorsque l'enfant sera majeur peut-il revenir sur ce consentement ?
   c. si l'enregistrement a lieu dans le cadre scolaire, faut-il des autorisations particulières ?

   Réponse : Ce cas rejoint le cas plus général des personnes pour lesquelles il faut demander une autorisation supplémentaire des responsables et tuteurs (parents et institution scolaire, dans ce cas) (voir chap. 3.3.2 catégorie des participants).

4. *Dans le cadre d'un travail au sein d'un laboratoire,*
   a. Qui est considéré comme l'auteur du corpus ?
   b. Quel(s) droit(s) ce travail donne-t-il au chercheur ?
   Réponse : Voir chap. 2.3 (droit d'auteur) et la fiche Droit d'auteur.
5. *Qui est considéré comme « responsable » de la diffusion et du traitement d'un corpus ?*
   Réponse : Voir chap. 2.3 et la fiche Responsable du traitement.
6. *Si je masque les noms propres de personnes, cela suffit-il pour que je puisse utiliser librement une transcription ?*
   Réponse : L'anonymisation ne consiste pas simplement en un effacement des noms propres. Voir chap. 3.5 Anonymisation et la fiche *Données personnelles et anonymisation*.
7. *Sous quelles conditions puis-je archiver mon corpus sous la forme de fichiers informatiques ?*
   Réponse : Il faut prendre en compte les aspects juridiques (protection de la vie privée, loi informatique et liberté, demande d'autorisation, voir les fiches *Données personnelles et anonymisation*, *Responsable du traitement* et les aspects techniques de conservation (voir fiches techniques).
8. *Si les personnes que j'ai enregistrées (dans les médias ou en privé) sont décédées, ai-je une liberté d'exploitation de ces enregistrements ?*
   Réponse : Les droits des auteurs survivent 70 ans après leur mort ! Quant à la protection au titre de la vie privée, elle ne peut être invoquée après la mort de la personne sauf si de son vivant la personne a interdit la diffusion. Par ailleurs, les membres de la famille du défunt peuvent invoquer leur droit personnel à la protection de la vie privée. Voir chap. 2.3.1 et fiche *Données personnelles et anonymisation*.
9. *Je découvre dans une armoire des enregistrements. Je voudrais pouvoir les exploiter. Je n'ai plus la trace de qui a enregistré ou qui a été enregistré.*
   a. Puis-je me servir de ces documents ?
   b. Quelles précautions (quelles garanties) dois-je prendre ?
   Réponse : On ne saurait trop inciter à la prudence et il est nécessaire de faire des recherches pour identifier les documents, y compris pour des raisons de rigueur scientifique. Voir chap. 2.3 et chap. 3.5.
10. *J'enregistre une émission à la radio (ou à la télévision).*
    a. Puis-je utiliser librement la transcription ?
    b. Puis-je utiliser la version sonore ?
    c. Du point de vue des autorisations, y a-t-il une différence entre émissions des radios publiques et des radios privées ?
    d. Y a-t-il une différence entre enregistrer des personnalités connues et enregistrer des « anonymes » (personnes qui témoignent, s'expriment en libre antenne, auditeurs qui posent des questions, etc.) ?
    e. les droits d'exploitation sont-ils différents si j'achète une cassette, un dévédé ou un cédé de l'émission ou si j'enregistre moi-même l'émission lorsqu'elle est diffusée ?
    Réponse : Les émissions radio sont protégées, qu'elles soient publiques ou privées. Voir 3.3.1 sur la reprise d'enregistrements médiatiques, et plus particulièrement la notion de documents d'actualité.
11. *J'aimerais constituer un corpus de données authentiques. Quelles sont les précautions que je dois prendre ?*
    Réponse : Voir chap. 3. où est proposée une réflexion articulant méthodologie de recherche sur le terrain et problèmes éthico-juridiques rencontrés au fil de la démarche.

Beaucoup d'autres questions encore…

2 LE CONTEXTE

SCIENTIFIQUE, POLITIQUE, JURIDIQUE ET INSTITUTIONNEL

Qui dit contexte dit « *mise en perspective* ». Telle est la finalité de ce chapitre qui présente ce qu'est le travail scientifique du linguiste sur l'oral. La mise en perspective se devait d'être aussi *politique* et *juridique*. Le contexte institutionnel a une importance grandissante compte tenu des besoins d'assurer, sur la durée, la « traçabilité » et la poursuite des recherches. En garantissant la pérennisation tant des données qui ont permis à un chercheur de travailler que des résultats obtenus, le chercheur comme l'institution participent au développement des connaissances dans un avenir proche ou plus lointain.

2.1 LA LINGUISTIQUE ET LES CORPUS ORAUX

Depuis une vingtaine d'années, les études sur les corpus de langues parlées ont complètement renouvelé les sciences du langage. Il suffit, pour s'en convaincre, de consulter les bibliographies récentes, en France et hors de France (par exemple la *Revue Française de Linguistique Appliquée* ou les *Recherches sur le Français Parlé)*. Ces études ont permis de formuler de nouvelles hypothèses sur le fonctionnement normal et pathologique du langage et elles sont devenues une composante essentielle du dialogue entre les linguistes et les informaticiens. En France, jusqu'à cette période encore récente, l'intérêt pour les langues parlées était essentiellement réservé aux domaines où il s'exerçait « par défaut » : en premier lieu les études sur les aspects proprement sonores de la langue (phonétique, phonologie et prosodie), le parler des jeunes enfants, ou tout ce qu'on classait parmi les « langues sans traditions écrites », en France les langues régionales et parlers locaux et, hors de France, tout ce qu'on nommait « langues exotiques ». A cela s'ajoutaient quelques essais isolés, dans les années 1950-1960, pour rassembler des modèles de français parlé afin d'enseigner le français en tant que langue étrangère, notamment le *Français Fondamental* et le *Corpus d'Orléans*.

Les représentations de la langue française, en particulier dans les grammaires, restaient fondées sur des données de langue écrite, littéraire ou non, les « grapholectes », comme les nommait Ong (1988), ou sur des données fournies par l'intuition. Cette mise à l'écart des données de langue parlée a entrainé deux conséquences majeures, d'une part l'image très négative que les Français ont de leur propre langue et d'autre part une influence considérable sur les théories linguistiques les plus courantes. Les nouvelles données révélées par les corpus de langue parlée n'ont sans doute pas encore fait évoluer l'image de la langue dans le grand public, mais elles ont déjà beaucoup fait évoluer les théories parmi les spécialistes.

De nouveaux domaines, abordés dès les années 1970 en Grande-Bretagne (Sinclair & Coulthard, 1975 pour l'École de Birmingham), ont émergé en France, comme l'essor des modèles de l'interaction et l'analyse conversationnelle (article fondateur de Sacks, Schegloff, Jefferson aux États-Unis en 1974, articles de Bange et de Quéré en France, en 1983 et 1984).

Les données de langue parlée collectées avant l'ère de l'informatique ne peuvent pas être comparées à ce qu'on appelle aujourd'hui « corpus de langue parlée ». Chacune

des collections anciennes, dispersées au gré des recherches, suivait ses propres règles de choix, d'enregistrement, de transcription et de conservation, de sorte qu'il est difficile maintenant d'y accéder et de les mettre en commun (les enregistrements du *Français Fondamental* ont été effacés, ceux du *Corpus d'Orléans* doivent être aujourd'hui retranscrits). Aucune ne pouvait atteindre de très grandes dimensions (il s'agissait généralement de quelques heures d'enregistrements seulement) et, dans ces données, la recherche d'informations ne pouvait se faire que manuellement. A partir des années 1980-90, le développement de l'informatique a permis de créer des corpus modernes de langue parlée dans le monde entier, en premier lieu dans les pays anglo-saxons. Une nouvelle discipline est née, celle des linguistiques de corpus (G. Kennedy en a donné une description en 1998 pour l'anglais et Habert et ses collaborateurs pour le français en 1997), qui intéressent les universitaires et les industries de la langue et qui, au titre de *Language Resources*, font maintenant partie des patrimoines nationaux. La France, qui était en avance pour la mise au point des corpus de langue écrite (en particulier pour FRANTEXT qui est à la source du *Trésor de la Langue Française*), a pris un grand retard dans la constitution des corpus de langue parlée.

Il existe de nombreux types de corpus de langue parlée, prévus pour divers objectifs, dans plusieurs disciplines. Il s'agit toujours d'enregistrements de données sonores, éventuellement accompagnées de données visuelles (prises en vidéo, ou à la télévision), presque toujours accompagnées de transcriptions et de traitements informatisés. Sans prétendre tout exposer ici, on en présentera quatre aspects : les types de données et de locuteurs, la dimension des corpus, les transcriptions, et un bref panorama des exploitations et des résultats.

2.1.1 TYPE DE DONNEES ET DE LOCUTEUR

Certaines données sont « sollicitées ». On fait par exemple venir dans des laboratoires de phonétique des locuteurs qui, agissant en tant que « cobayes », fournissent des types de prononciations et d'intonations, dans de très bonnes conditions d'enregistrement. On leur fait prononcer des mots et des listes de mots, des nombres et des listes de nombres, ou on leur fait lire des textes ou fragments de textes. Ces documents servent à différentes exploitations, soit pour consigner et étudier les prononciations en tant que telles, comme le font J. Durand, B. Laks et Ch. Lyche pour étudier la prononciation du français contemporain (projet PFC), soit pour tester un comportement langagier (comme on le fait dans des services hospitaliers qui étudient des phénomènes d'aphasie), soit pour établir des analyses qui servent à la synthèse de la parole ou à la lecture automatique de textes écrits (*Text-to-Speech data)* ou aux dialogues homme-machine (c'est l'objectif de *SpeechDat Exchange*, qui stocke de 500 à 5 000 enregistrements téléphoniques pour 28 langues). Dans toutes ces situations, les locuteurs savent généralement qu'ils sont enregistrés et ils ont une idée, précise ou approximative, de la finalité de leur prestation.

D'autres données sont dites « de parole continue », avec divers degrés de spontanéité (la notion a été spécialement étudiée dans un numéro de la *Revue Française de Linguistique Appliquée*). Certaines sont recueillies dans des situations qui n'ont pas été provoquées par le chercheur et qui auraient eu lieu de toute façon sans lui. D'autres,

plus ou moins « sollicitées », sont orchestrées et organisées par le chercheur. L'idéal du spontané total serait d'enregistrer les locuteurs sans qu'ils s'en doutent (micros cachés, enregistrements pirates), en le leur disant ensuite ou sans le leur dire, l'objectif étant de saisir leur langage « en toute liberté », avec un minimum de contrôle. Les dispositions juridiques limitent cette possibilité. La présence de l'enquêteur et des appareils apporte de toute façon un frein à cette liberté (c'est la question du « paradoxe de l'observateur » popularisée par W. Labov). Dans la pratique, divers degrés de contrainte peuvent être identifiés, selon qu'il s'agit de parole privée ou de parole publique, devant des familiers ou des étrangers, avec diverses formes de complicité ou non, selon qu'il s'agit de parole en face-à-face ou de parole transmise par un canal comme le téléphone, le répondeur, la radio, la télévision ou d'autres dispositifs techniques. Une bonne approche ethnographique (enregistrements répétés) permet de résoudre le problème de la sensibilité au micro. Mais cela demande qu'on y consacre beaucoup de temps pendant la phase de recueil des données.

Il est rare que les corpus modernes soient composés de paroles « de tout venant ». Le choix des locuteurs et des situations d'enregistrement est généralement fixé en fonction des objectifs donnés au départ. Les chercheurs proposent de collecter des conversations entre adultes, des négociations professionnelles, des entrevues (préparées ou non), des prises de parole dans des organismes publics, des discours électoraux, des explications entre services publics et utilisateurs, des cours publics, des sermons, des discours politiques, des conférences (spécialisées ou de vulgarisation), des témoignages historiques, des récits de faits-divers, des récits de vie (produits par des individus, des groupes, des représentants de groupes, des porte-paroles), des dialogues entre mères et jeunes enfants, des enfants enregistrés dans un contexte scolaire ou en dehors (dans leurs jeux ou dans leurs récits, en réponse à des tests ou en dehors, dans des situations scolaires ou non, dans des jeux libres ou contraints, avec parodie et jeux de rôles), des malades dans les hôpitaux, etc. Un exemple : une banque de données CLAPI (Corpus de Langue Parlée en Interaction) est constituée actuellement à Lyon (laboratoire ICAR) afin de réunir des corpus de « parole en interaction » les plus diversifiés possibles, dans des situations non provoquées par les chercheurs : conversations à table, concertations entre notaires, appels à des centres d'aide sociale d'urgence et à des consultations thérapeutiques, etc. Cette banque de données comporte 300 h. d'enregistrements audio et vidéo, des transcriptions et des « métadonnées » décrivant les caractéristiques des locuteurs.

De nombreuses disciplines cherchent à étudier les corrélations entre les productions de langue parlée et d'autres phénomènes. Les corrélations entre langage et paramètres socio-économiques ont été à la base des recherches de sociolinguistique. Aux Etats-Unis, W. Labov avait produit de célèbres études sur les Noirs des grandes villes américaines de l'Est, en enquêtant dans les domiciles, dans les rues ou dans les grands magasins (avec des conditions d'enregistrement souvent défectueuses). Les études sur le développement du langage se font en fonction de l'âge des enfants, des activités observées, des consignes fournies et des données familiales. La prise en compte des « genres » (tels que les conçoit D. Biber pour l'anglais) amène à faire des corrélations avec les lieux de prise de parole, les sujets dont il est question, les types

d'interlocuteurs et le type d'échanges (monologues, dialogues, conversations à plusieurs). Pour pouvoir mesurer ces corrélations, le contenu et la taille des corpus sont généralement définis à l'avance : tant de types de situations et de locuteurs (comme l'avait fait l'équipe Sankoff-Cedergren dans les années 1970 pour étudier la variation sociale dans la ville de Montréal). Dans d'autres cas, les chercheurs découpent, à l'intérieur de corpus existants, des sous-corpus représentatifs adaptés à leur étude (c'est ce qu'a proposé D. Biber pour faire des échantillonnages dans le grand *British National Corpus*). Il s'agit en ce cas de corpus « fermés » et « échantillonnés ».

Les linguistes, de leur côté, ont souvent collecté des corpus « ouverts », qu'ils modifient au gré de l'avancement de leur travail, sans délimiter à l'avance un objet de recherche pré-déterminé, parce qu'ils sont certains de découvrir des phénomènes nouveaux, impossibles à prévoir au départ : répartition du langage formel et informel, relations entre grammaire et lexique, liens entre degrés de complexité de la syntaxe et type de situations de parole, utilisation de la morphologie orale, rôle des contextes dans la construction du sens des énoncés, rôle de la prosodie dans la structuration des textes, etc.

La qualité technique des enregistrements dépend bien évidemment des équipements techniques utilisés, mais aussi des types de situations et de locuteurs choisis (lieux bruyants, locuteurs trop nombreux, locuteurs affectés d'un défaut de parole). Ces situations diverses influent également sur le consentement des locuteurs : il est plus facile d'obtenir l'autorisation d'enregistrer la parole publique que la parole privée, les propos d'un locuteur sûr de lui-même plutôt que d'un locuteur inquiet et sensible à ce que l'on a pu appeler « l'insécurité linguistique ».

Dans tous les cas, il est bien difficile de justifier les enregistrements par l'étude de la langue. Si on explique cette finalité, les locuteurs français ont immanquablement l'impression qu'ils parlent mal et que l'étude va les ridiculiser. Peu d'entre eux sont détendus sur cette question. Presque tous les chercheurs ont mis au point des stratégies pour aborder le problème de biais : en disant qu'ils s'intéressent au contenu, aux témoignages, aux explications, au savoir particulier des locuteurs (qui peut être un savoir de langage, dans le cas des recherches sur les régionalismes). Dans les travaux sur la parole en interaction, les choses sont un peu différentes : les chercheurs peuvent dire qu'ils s'intéressent précisément à la manière dont les participants interagissent entre eux, à leur coordination, aux ajustements remarquablement précis auxquels ils recourent, par la parole, les gestes, les mimiques, les regards et l'ensemble des attitudes (ressources multimodales, difficilement contrôlables dans leur ensemble même par des locuteurs qui se surveillent).

2.1.2 *DIMENSIONS*

La dimension utile des corpus et des unités qui les constituent varie selon l'étude prévue. Les études de phonétique, de phonologie et de prosodie peuvent donner de bons résultats avec des unités sonores de durée assez limitée. Mais, si l'on veut étudier des corrélations entre le langage et d'autres phénomènes, ou si l'on veut étudier le lexique, il y faut des unités beaucoup plus développées, en quantité plus importante et dans des domaines d'activité plus diversifiés. La dimension des corpus

de langue parlée et des éléments dont ils sont composés se mesure avec deux sortes d'unités. On utilise des unités de temps lorsqu'on s'intéresse prioritairement à l'enregistrement sonore, en faisant abstraction de la transcription. On classe par exemple comme très petits éléments de corpus ceux qui durent entre quatorze et trente secondes (quatorze secondes étant la moyenne pour une information à la radio). Mais on tient compte de sous-unités encore plus petites quand on observe les chevauchements de parole entre les locuteurs ou quand on mesure les pauses (jusqu'au dixième de seconde). Les petites unités sont utilisées par exemple par les compagnies de téléphone qui construisent actuellement des services européens de renseignements par téléphone dans toutes les langues de l'Europe (EuroSpeech 2003). On classe comme petits éléments ceux qui durent dix minutes et comme très grands éléments ceux qui ont une durée de soixante ou quatre-vingt-dix minutes. En totalisant l'ensemble de ces éléments, on dira par exemple qu'on dispose de réserves de 100 ou 500 heures d'enregistrements.

Mais ces mesures sont peu fiables pour les grands composants de corpus, parce que la densité des enregistrements dépend du débit des locuteurs. En français, on estime que les locuteurs qui parlent lentement prononcent 110 mots par minute et que ceux qui parlent très vite en prononcent 350 par minute (dans certains types d'aphasie, et sous l'influence des neuroleptiques, le débit tombe au-dessous de 100 mots par minute, ce qui est pénible à écouter. Au-dessus de 350 mots par minute, l'écoute et la transcription deviennent très difficiles). La densité varie donc de un à trois, ce qui est considérable. Selon les deux débits extrêmes qui viennent d'être cités, une heure d'enregistrement peut correspondre à 6 600 ou à 21 000 mots. On a donc intérêt à évaluer les grands corpus en fonction du nombre de mots graphiques que comporte la transcription. Les grands corpus de langue parlée collectés aujourd'hui dans le monde sont de l'ordre de grandeur de la dizaine de millions de mots transcrits pour l'anglais, américain ou britannique. Malheureusement, les corpus actuels de français parlé sont de l'ordre de grandeur du million de mots. Avec une taille aussi limitée, il n'est guère possible de faire des recherches lexicales, ni d'établir des statistiques fiables sur les usages.

2.1.3 TRANSCRIPTIONS

Les transcriptions de langue parlée qui ont cours aujourd'hui sont tellement différentes les unes des autres qu'il est difficile de les rassembler sous une même étiquette. Dans certains cas, lorsqu'on ne retient que le contenu des enregistrements, en en changeant librement la forme, les termes de *transposition* ou d'*adaptation* conviendraient mieux. C'est ce que font souvent les journalistes, lorsqu'ils rapportent les propos de personnes interviewées, en résumant ces propos et en leur donnant généralement une tournure plus normative (là où un homme politique important dit *ça, je sais pas, pour pas que…,* ils rétablissent *cela, je ne sais pas, pour que …ne pas…*). Les historiens et les sociologues ont parfois des pratiques voisines, lorsqu'ils s'intéressent avant tout au contenu informatif : ils font un tri dans les données, coupent les passages qui ne les intéressent pas et suppriment les particularités de la production orale qui leur paraissent gênantes, répétitions, hésitations ou retouches.

Certains secteurs d'activité, comme les transcriptions de débat parlementaire, ont même codifié ces tâches, en établissant plusieurs degrés d'adaptation.

Lorsqu'il s'agit de s'intéresser au langage lui-même, le choix d'un type de transcription dépend des finalités de l'étude (des projets européens et internationaux se sont donné des consignes d'édition de corpus) et, comme le signalait déjà E. Ochs en 1976, la transcription engage toujours une théorie. Certaines études nécessitent de disposer de transcriptions phonétiques ou phonologiques. Le standard Unicode, synchronisé sur la norme ISO-10646, comporte déjà dans sa version 4.0 plus de 96 000 caractères dont, en particulier, ceux de *l'Alphabet Phonétique International*. C'est une nécessité pour tous les travaux qui concernent la prononciation, mais aussi pour tous les cas où il est difficile de dégager des morphèmes stables qu'on pourrait écrire en orthographe standard : langage des très jeunes enfants (modèle international CHILDES), langage des étrangers en cours d'acquisition de la langue, notation de certains régionalismes, notation de certaines formes d'aphasie comme les jargons (Abou-Haidar 2002). Ces transcriptions, qui ne peuvent se faire que pour de petites quantités de corpus, sont souvent accompagnées de traductions juxtalinéaires. La représentation de la prosodie exige des modèles spécifiques, très développés dans les techniques récentes (Martin 1987). Les enregistrements vidéo demandent des notations spéciales, qu'on peut pousser plus ou moins loin (Van der Straten 1998, Mondada 2006).

En ce qui concerne les grands corpus de langue parlée, ils sont transcrits en orthographe standard, de façon à en rendre la lecture facilement accessible. A partir de ce choix, plusieurs options sont possibles : orthographe standard avec ou sans adaptations, avec ou sans ponctuation, avec ou sans indications de pauses, allongements, rythmes, accentuations, hésitations, toux, rires, gestuelle, etc. De grands débats ont eu lieu sur tous ces points, pour dégager les conditions optimales de transcription, adaptées aux objectifs de la recherche. Un exemple : les linguistes qui s'intéressent aux unités syntaxiques de la langue parlée se méfient généralement de la ponctuation, qui impose des délimitations propres à la langue écrite et qui s'avère souvent trompeuse quand on la met avant d'avoir suffisamment bien analysé les textes. Mais les textes non-ponctués indisposent les informaticiens, dont les analyses automatisées réclament des repères de ponctuation. Des négociations sont parfois menées entre les linguistes et les informaticiens (ICOR au laboratoire ICAR) afin d'établir des conventions de transcription qui tiennent compte de ces problèmes et des standards internationaux (GAT, TEI, Du Bois, Jefferson).

Les transcriptions qu'utilisent les linguistes conservent soigneusement toutes les particularités des productions orales : répétitions, hésitations, amorces de mots, retouches. Elles exigent que le transcripteur veille à ne pas projeter sur la transcription ses propres interprétations (ajouter ou ôter des *ne* de négation, par exemple, ou reconstruire une portion de texte selon les stéréotypes attendus). Ce souci du détail exige un entrainement et une formation spécifique des transcripteurs. La tâche, longue et couteuse, est pleine de pièges (Leech 1991). Selon les estimations courantes, un minimum de trente minutes de travail est nécessaire pour transcrire une minute d'enregistrement (les concepteurs du corpus néerlandais estiment que cela revient à un euro par mot graphique !). En raison même de leur fidélité, les

transcriptions de la langue parlée déplaisent aux profanes : ils y voient quantité de « fautes de français », de répétitions, de dissolutions de l'information. Montrer à un informateur profane une transcription de sa parole provoque souvent le rejet. Ce n'est pas un très bon moyen pour obtenir son autorisation de transcrire et publier le résultat de la recherche.

L'outillage informatique a transformé le travail de transcription, d'une part par les aides qu'il a apportées, d'autre part par les exigences nouvelles qu'il a introduites. Les aides à la transcription (Anvil, Clan, Elan, Ite, Praat, Transcriber…) facilitent les manipulations et permettent de réécouter facilement les portions d'enregistrement sous étude. La technique des *corpus synchronisés* permet de lire sur écran des portions de texte écrit en même temps qu'on écoute les mêmes portions dans leur déroulement sonore (*Speech Communication* 33, numéro spécial sur les annotations et les outils d'analyse des corpus). Les exigences nouvelles concernent les annotations informatisées : étiquetage morpho-syntaxique de tous les éléments du texte, arborescences, métadonnées (concernant les circonstances d'enregistrement, les situations et les locuteurs). Divers classements et codages permettent de faire les lemmatisations et les concordanciers nécessaires pour pouvoir formuler des requêtes sur l'ensemble du corpus. Une polémique s'est engagée, dans les années 2000, autour du degré de sophistication des annotations qui semblait nécessaire (Sinclair, Teubert). La standardisation se fait maintenant au plan européen (SpeechDat Exchange Format).

2.1.4 TRAITEMENT AUTOMATIQUE DE LA PAROLE

Contrairement à bien d'autres domaines de recherche autour de la parole, la transcription automatique de la parole, qui s'effectue sur un flux acoustique continu, nécessite une modélisation de l'ensemble des phénomènes observés dans le signal sonore. Il faut donc modéliser, au-delà des mots auxquels est associée une représentation phonologique dans le dictionnaire de prononciation, des phénomènes extra-lexicaux : respirations, hésitations, fragments de mots, etc.

Suivant les genres de document traités, les systèmes de reconnaissance automatique obtiennent des taux d'erreurs très variables[1]. Cependant, s'il y a un décalage entre les modèles (en gros les connaissances) du système et les corpus à transcrire, ces taux d'erreurs peuvent augmenter rapidement. Afin d'arriver aux meilleurs résultats possibles, les systèmes de transcription doivent être adaptés en fonction des corpus à transcrire.

Les recherches actuelles montrent que la transcription automatique est en train de devenir un instrument précieux pour aider la transcription et l'annotation de corpus. Par exemple dans (Barras *et al.* 2004) est montrée l'utilité de la transcription automatique pour la génération semi-manuelle de transcriptions acoustiques fines (c'est-à-dire comprenant non seulement tous les mots orthographiques mais également les « disfluences » et autres évènements extra-lexicaux). Les recherches en cours montrent également que la transcription automatique de la parole peut devenir un

[1] Les travaux du LIMSI (Barras 2004) présentent des résultats allant de 10 à 30 % d'erreur de mots avec des systèmes optimisés pour une tâche donnée.

instrument précis pour explorer, analyser des corpus, quantifier des phénomènes linguistiques. Plus généralement, on peut penser qu'à l'avenir il faudra de moins en moins opposer les visions des linguistes et des informaticiens. À cet égard, l'émergence de la linguistique des corpus oraux comme domaine de recherche doit reposer sur la formation de linguistes informaticiens et d'informaticiens linguistes.

2.1.5 *EXPLOITATIONS ET RESULTATS*

Les grands corpus actuels de langue parlée sont chers. Certains corpus, notamment dans l'ingénierie, sont exploités en association avec les industriels : dialogue homme-machine, reconnaissance et synthèse de la parole, communications téléphoniques, etc. (des organismes comme ELRA/ ELDA se sont spécialisés dans la diffusion des corpus et des ressources disponibles dans ce domaine).

Les grands corpus servent en premier lieu de documentation générale sur la langue nationale. Les grands *corpus de référence*, échantillonnés en tenant compte des régions et des données socio-économiques et culturelles, permettent de guider les politiques linguistiques à grande échelle. Par exemple, le corpus de référence du portugais parlé, qui comporte des enregistrements réalisés au Portugal, en Afrique, au Brésil et en Asie, permet d'évaluer les différences selon la géographie mondiale, et de fonder sur cet examen, certains usages de pratiques scolaires et même des décisions gouvernementales. Le *British National Corpus* a servi de base à la fabrication d'une grande grammaire, la *Longman Grammar of Spoken and Written English*, conçue sur des bases très nouvelles. Une grande activité éditoriale s'est développée en langue anglaise, en utilisant ces matériaux. C'est ainsi que l'éditeur Collins a utilisé les corpus anglais pour la publication de nombreux ouvrages didactiques servant à l'enseignement de l'anglais comme langue maternelle et comme langue étrangère. Une documentation sur la langue parlée est parfois le point de départ pour lancer des activités nouvelles : des corpus de langue parlée ont servi de base pour diffuser des langues peu (ou pas du tout) écrites, comme on l'a fait pour la langue maori, qui a servi de modèle pour développer des émissions de radio et de télévision (Kennedy 1998 : 72).

La comparaison entre langues parlées appartenant à un même groupe linguistique permet d'évaluer *in vivo* les ressemblances et différences à l'intérieur d'une grande aire linguistique.

Une exploitation importante est celle qu'offrent les corpus multilingues (appelés aussi corpus parallèles ou alignés), qui servent aux traducteurs, à l'enseignement des langues et à l'étude contrastive. Il en existe pour la langue écrite :

- anglais/français à l'université de Lancaster, à l'université d'Oslo, à Mannheim, à l'université de Gand en Belgique (Contragram, bank.ugent.be/contragram/newslet.html), à l'université de Montréal,
- français/ anglais/ néerlandais, à l'université de Courtrai,
- français/anglais/espagnol à l'université de Pennsylvanie.

Une étude récente, fondée sur des enregistrements et transcriptions de quatre langues romanes (italien, français, portugais, espagnol) permet de comparer la prosodie (intonations, accentuations, rythmes), en tenant compte de différentes situations et différents médias (C-ORAL-ROM, Cresti & Moneglia).

C'est ainsi que les grands corpus de langue parlée ont renouvelé quantité de problèmes linguistiques. Sur les données livrées par ces grandes collectes, de nouvelles disciplines se sont fondées, comme l'analyse conversationnelle et l'analyse des interactions, des négociations et des codes de politesse. Les recherches en pragmatique s'appuient massivement sur ces données. Certaines connaissances ont été nettement modifiées, comme par exemple les études portant sur la production et sur la perception du langage parlé et, par voie de conséquence, sur la fragilité de l'intuition linguistique (Blanche-Benveniste 1997). On a pu montrer quel est le degré d'organisation ordonnée et systématique dans les interactions. On s'en est servi pour remettre en cause certaines unités de base comme la *phrase*, et pour en introduire d'autres comme les unités de *macro-syntaxe*, utilisée maintenant par plusieurs équipes de linguistes (Blanche-Benveniste *et al.*, 1999, Scarano 2003, Nolke 2002). L'étude de l'intonation a été prise en charge très sérieusement dans la délimitation des unités de macro-syntaxe (Cresti & Moneglia 2005, Couper-Kuhlen & Selting, 1996). Dans les interactions, on a montré qu'intervenaient plusieurs niveaux d'organisation imbriqués (Turn-Constructional Units ou « Unités de Construction du Tour », Selting 1995, 1998, 2000, Auer *et al.* 1999, Ochs, Thompson & Schegloff, 1996). Dans différentes langues, on a pu montrer quel était le rôle des caractéristiques de productions orales que sont les particules discursives, les répétitions, les hésitations ou les « réparations », qui intéressent actuellement les neurosciences. Les perspectives sur l'histoire des langues en ont même été modifiées, dans la mesure où l'on peut maintenant étudier l'influence qu'exercent les différentes situations de parole sur le type de grammaire adopté (Biber 1987). On peut montrer par exemple, pour le français, que les récits d'explication et les argumentations révèlent des pratiques de syntaxe à haut degré d'enchâssement, alors qu'il y en a rarement dans les conversations, ou que les récits d'accidents contiennent des organisations chronologiques complexes. On sait que les thèmes réputés « sublimes » (discours sur la morale, la religion, la mort) déclenchent des caractéristiques de « langue de cérémonie », par exemple, en français, un grand nombre de liaisons, des emplois massifs du *ne* de négation et même parfois des emplois inattendus de passé simple. Les grands corpus permettent de suivre certains processus de grammaticalisation en cours. Ils montrent l'importance numérique des énoncés parenthétiques, des focalisations et des thématisations. Ils obligent à considérer que les locutions figées occupent une place très importante par rapport à la libre composition des énoncés, de sorte que le lien entre la grammaire et le vocabulaire apparait maintenant plus nettement qu'auparavant, beaucoup de tournures grammaticales n'étant utilisées par les locuteurs que pour une petite liste de mots du lexique. Il faut en conclure que, lorsqu'on parle, on ne choisit pas « un mot » mais un ensemble préconstruit (Sinclair, 1991).

Cela remet en cause, évidemment, les théories linguistiques qui visaient à isoler la syntaxe comme une composante du langage indépendante.

Ces grands corpus, lorsqu'ils existent, rendent un service primordial : ils servent de base de données pour toutes les comparaisons concernant le langage : pour évaluer le langage des enfants à divers stades d'acquisition, pour soutenir les diagnostics dans les pathologies de langage, pour évaluer le degré d'accomplissement dans l'acquisition des langues maternelle et étrangère, pour calculer les effets des langages

de groupes et des langages professionnels (Gadet), pour étudier les modes de coordination dans une équipe ou dans un groupe, pour comprendre les spécificités des types d'activités et des contributions qui y sont adéquates dans des contextes institutionnels différents ou pour connaitre l'effet des influences régionales. Un exemple : avant de juger qu'une tournure est caractéristique du parler des enfants de tel âge ou de telle origine, il est indispensable de recourir à une base de données de comparaison pour savoir si la tournure est spécifique ou non (les fautes les plus courantes sur les relatifs *dont* et *lequel*, premier degré, se retrouvent chez les adultes les plus scolarisés, et depuis assez longtemps, pour autant qu'on puisse en juger).

CORPUS DE LANGUES A TRADITION ORALE

Les problèmes rencontrés lors de la constitution, l'exploitation, la diffusion et la conservation de corpus oraux dans les sociétés dites « à tradition orale », ou « ethniques », ou « exotiques », recoupent partiellement ceux rencontrés lors de l'établissement des grands corpus de langues occidentales. Les précautions à prendre (telles que préconisées dans le guide) pour respecter les personnes sont alors à adapter au contexte dans lequel se déroule le travail de terrain.

Dans une société à tradition orale, l'autorisation après information (sur le modèle du « consentement éclairé » décrit dans le guide, là encore adapté à la situation) peut, dans certains cas, n'avoir de valeur que si elle est orale, et accordée par la personne qui en a le pouvoir (tout comme dans une situation d'enquête en milieu médical, l'autorisation n'aura de valeur que si elle est accordée par l'Ordre des médecins). Par ailleurs, l'information du locuteur n'est pas sans poser problème dans des sociétés où l'activité de recherche, les objectifs de la constitution du corpus et ses réseaux de diffusion (publications, internet) ne correspondent à rien de concret.

Le chercheur doit par ailleurs s'informer du droit en vigueur dans le pays dans lequel il va travailler. Par exemple, le droit français ne reconnait pas la propriété intellectuelle ni les droits d'auteur dans le cas de recueil de contes ou de mythes, considérés comme faisant partie du patrimoine et relevant du domaine public. Dans plusieurs pays d'Afrique en revanche (voir annexe), il a été créé des bureaux de droit d'auteur pour protéger ce type de productions et leurs auteurs. Par ailleurs, certaines communautés ne reconnaissent pas le droit national de l'État dans lequel elles vivent. C'est par exemple le cas dans certaines communautés amérindiennes en Guyane qui fonctionnent selon un droit collectif et non pas privé (Tiouka 2005, pour une réflexion sur l'intégration du droit coutumier dans le droit français et européen). Certaines autorisations n'auront de valeur pour ces communautés que si elles respectent le droit coutumier, et bien que le chercheur se sente protégé en respectant le droit national, il peut se retrouver en conflit avec les autorités coutumières et se voir refuser l'accès au terrain.

L'exploitation du corpus nécessite dans la plupart des cas l'intervention de plusieurs personnes : le transcripteur, qui peut être le locuteur de l'enregistrement, mais pas toujours ; le traducteur (id). Les droits de ces personnes sur le corpus sont là encore à définir selon plusieurs paramètres : le droit national s'il a un sens pour la communauté, ou bien le droit coutumier.

Il est deux points sur lesquels la constitution de corpus oraux dans certaines sociétés diffère de celle des grands corpus de langues nationales :

1. la taille du corpus

Il est difficilement envisageable d'arriver à recueillir des corpus sur certaines langues qui pourraient atteindre la taille des grands corpus de langues européennes, comprenant plusieurs millions de mots. Le problème de la représentativité des corpus se pose alors de façon différente.

2. le retour à la communauté

La pratique du terrain (anthropologie du début du 20^{e} siècle, linguistique missionnaire, etc.) jusqu'à il y a encore une cinquantaine d'années a laissé des

traces dans les communautés qui se sont senties pillées et exploitées sans avoir jamais eu accès aux résultats de la recherche. Celles-ci réclament maintenant que les recherches aient des retombées directes sous diverses formes, et ces revendications sont reprises par tous les organismes qui financent ou organisent des recherches sur les langues en danger (UNESCO, projet DOBES du Max-Planck Institute[2], etc.). Les revendications qui émanent des communautés n'ont rien à voir avec le dédommagement du travail individuel des différents locuteurs impliqués dans la constitution et l'exploitation du corpus, et demandent une implication du chercheur (participation à des programmes d'éducation, restitution des enregistrements et des matériaux collectés, constitutions d'archives accessibles aux communautés, etc.). La nécessité de rendre à la communauté les matériaux collectés devrait d'ailleurs motiver les chercheurs à constituer des bases de données et à archiver leurs corpus.

Cependant, il semblerait que ces corpus restent encore souvent des « outils personnels » uniquement destinés à servir de base à l'analyse linguistique du seul chercheur. Les raisons sont multiples : comme pour la constitution de n'importe quel corpus de langue orale, l'aspect technique et le temps nécessaire à la mise en forme (numérisation, dans certains cas synchronisation, etc.) rebutent le chercheur, et ce d'autant plus que ce travail n'est pas valorisé par les institutions scientifiques. Par ailleurs, sur certains terrains, la constitution du corpus est le résultat d'une relation de confiance qui s'est établie entre le chercheur en tant que personne (et non pas en tant que représentant d'une communauté scientifique) et la communauté ou certaines personnes de la communauté, dans des contextes difficiles. La décision de diffuser le corpus au sein de la communauté scientifique ou plus largement (accessibilité par Internet) interroge l'éthique du chercheur, et ne relève plus du cadre juridique. Dans tous les cas, il est souhaitable que cette diffusion se fasse après la restitution du corpus à la société concernée.

Le chercheur se trouve donc en porte-à-faux entre la volonté de préserver une relation privilégiée avec son terrain, et avec la croissante nécessité de mettre à la disposition de la communauté scientifique les ressources qui sont à la base de l'analyse et des résultats de la recherche.

Au fur et à mesure que se développent les grands corpus de langue parlée actuels, la standardisation progresse (depuis les consignes diffusées par EAGLES en 1993) et les champs de recherche deviennent de plus en plus intéressants. Dans cette perspective, rendre accessibles les corpus de français parlé existants ou ceux de toute autre langue et en créer de nouveaux est une tâche importante du « patrimoine immatériel ». Les problèmes juridiques de protection de la parole, qui ont longtemps été considérés, à tort, comme secondaires, sont actuellement des freins très importants : beaucoup de chercheurs refusent de faire circuler leurs corpus parce qu'ils ne sont pas sûrs d'avoir « les bonnes autorisations ». Beaucoup hésitent à en lancer de nouveaux, parce que la demande d'autorisations leur parait fondamentale mais difficile à accomplir. C'est pourquoi une réflexion collective sur cette question est maintenant indispensable.

[2] Voir dans la bibliographie les liens vers les sites de ces organisations.

2.2 CADRES POLITIQUES DE LA DIFFUSION DE LA RECHERCHE

LA DIFFUSION DES RESULTATS DE LA RECHERCHE FAIT PARTIE DES MISSIONS DES CHERCHEURS

> *« Les organismes publics doivent avoir le souci constant de faire bénéficier au mieux la collectivité nationale des fruits de leurs travaux... ».*
>
> *« La politique de la recherche et du développement technologique vise à l'accroissement des connaissances, à la valorisation des résultats de la recherche, à la diffusion de l'information scientifique et technique et à la promotion du français comme langue scientifique »*[3].

C'est en ces termes que le rapport annexé à la loi d'orientation du 15 juillet 1982 définit les contours de la valorisation. Il ne fait aucun doute que ces principes généraux trouvent à s'appliquer aux chercheurs dont les travaux aboutissent à la constitution de corpus oraux. Toutefois les conditions de la valorisation et de la diffusion dépendront aussi des possibles droits existants sur les contenus collectés et sur les résultats du traitement de ceux-ci par les chercheurs.

LA DYNAMIQUE DE L'ECHANGE ET LES OCCASIONS OFFERTES AUX TITULAIRES DE DROITS POUR FACILITER LA LIBERTE D'ACCES DANS LA SOCIETE DE L'INFORMATION

Sans doute peut-on parler aujourd'hui d'une nouvelle manière de voir le rapport de chacun dans l'échange de l'information. Cette dynamique de l'échange engendre, de fait, de nouveaux comportements. La liberté d'accès, la gratuité et le droit de réutilisation semblent aller de soi quand ils s'inscrivent dans la réciprocité.

Le 22 octobre 2003, à Berlin, la plupart des Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST) ont signé la *Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales*, dont l'objectif est de promouvoir Internet « comme instrument fonctionnel au service d'une base de connaissance globale de la pensée humaine ».

En signant cette déclaration, les responsables politiques chargés de la science, les institutions de recherche, les agences de financement, les bibliothèques, les archives et les musées se sont engagés à envisager un certain nombre de mesures. Ces mesures doivent permettre de « trouver des solutions aptes à soutenir le développement des cadres actuels, juridique et financier, en vue de faciliter un accès et un usage optimaux » d'Internet. Le texte reconnait aussi l'existence d'une possible contradiction entre les demandes de protection et de libre accès. Enfin, de cette déclaration, il ressort que le libre accès requiert l'engagement de chacun en tant que producteur de connaissances scientifiques ou détenteur du patrimoine culturel, ce

[3]Art 5 de la Loi n°82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France, aujourd'hui art. L 111-1 du code de la recherche. *JO* du 16-07-1982, p. 2273 et ss.

libre accès se faisant « dans le respect des droits des auteurs ou des titulaires ». Le libre accès doit donc être réglementé et modulé par les titulaires de droits. Les auteurs (ou l'institution) peuvent concéder un « droit gratuit, irrévocable et mondial d'accéder à l'œuvre » ou bien « une licence autorisant à copier, utiliser, distribuer, transmettre, montrer en public, réaliser et diffuser des œuvres dérivées, sur quelque support numérique que ce soit et dans quelque but responsable que ce soit, sous réserve de mentionner comme il se doit son auteur ». Peut être uniquement concédé « le droit d'en faire des copies imprimées en petit nombre pour un usage personnel ». La formalisation de ces autorisations peut se faire sous forme de licences de type **creatives commons* (autorisations d'utilisation données directement par les auteurs, sans contrepartie financière. Les auteurs peuvent en revanche, le cas échéant, poser des limites à cette utilisation en la réservant exclusivement à des usages à but éducatif).

Appliquées aux corpus oraux, ces licences peuvent être un moyen de mettre à la charge des futurs utilisateurs le respect des engagements souscrits par le chercheur créateur du corpus à l'égard de tous ceux qui ont contribué à son élaboration.

LES PROGRAMMES DE NUMERISATION PATRIMONIALE

Le contexte de la société de l'information a suscité de nombreuses initiatives publiques dans le dessein d'assurer la pérennisation de la mémoire culturelle. En 2001 à Lund, en Suède, un groupe de représentants nationaux des États membres de l'Union européenne, intéressés par les problèmes de numérisation, a élaboré un texte qui prône notamment : la mise en place de standards d'interopérabilité ; la diffusion de bonnes pratiques dont la gestion des *droits de propriété intellectuelle ; l'organisation de centres de compétences sur la numérisation dont les professionnels de l'information ont la responsabilité.

La question de la conservation des résultats de la recherche se pose aujourd'hui avec d'autant plus d'acuité que les résultats, mais aussi les matériaux mêmes qui ont servi à ces recherches sont sur des supports numériques. Comment assurer la « traçabilité » des différentes étapes du travail de recherche ? Que faut-il conserver ? Qui assurera cette conservation ? Dans quelles conditions ? Ces questions doivent aujourd'hui être posées et trouver des éléments de réponse pour chaque opération de recherche. Si des recommandations générales peuvent être données, cela n'exonère en rien les responsabilités de ceux qui initient une recherche dont l'un des objectifs, ou l'une des étapes consiste en l'élaboration d'un corpus oral.

2.3 CADRES JURIDIQUES

Le propos de ce guide qui s'adresse à des chercheurs n'est pas de traiter de toutes les techniques juridiques à appréhender (on renverra pour une présentation plus détaillée de certains des sujets abordés à des fiches spécifiques en annexe). Il s'agit de sensibiliser le lecteur et de l'inviter à se poser les questions nécessaires pour comprendre ses obligations mais aussi ses droits.

Quel peut être le statut juridique de chacun des corpus oraux constitués par les chercheurs ? Cette question peut *a priori* sembler théorique, mais nous ne pouvons pas l'occulter car c'est en fonction des réponses apportées qu'il sera possible de déterminer les conditions d'exploitation et de diffusion des corpus. Pour répondre à

cette question, il faut tout d'abord connaitre les conditions d'élaboration du corpus et de ses différentes composantes. Le corpus est-il constitué d'informations du *domaine public ? Est-il le produit d'une ou plusieurs créations intellectuelles susceptibles d'être protégées par le *droit d'auteur ? Les contenus du corpus sont-ils des *données personnelles ? Quels sont alors les droits des locuteurs ou des personnes concernées ?

Ces statuts juridiques déterminés et les droits qui en découlent une fois connus, il convient de s'enquérir des modalités de la gestion contractuelle de ces droits. Les titulaires des droits se sont-ils prononcés sur les conditions de mise à disposition et de réutilisation des corpus ?

Enfin, ce sont les questions de la responsabilité de tous ceux qui auront à intervenir dans la « vie du corpus » qui méritent attention : responsabilité des créateurs, responsabilités des hébergeurs, des diffuseurs, des archiveurs… (voir annexe).

Pour faciliter la démarche du chercheur, on donnera ici un aperçu sur quatre grandes questions qui reviennent de façon récurrente dans la constitution et la vie des corpus : qu'est-ce que le domaine public, c'est à dire « l'inappropriable » ? Quand est-il question de droit d'auteur à propos des corpus ? Comment assurer la protection des données personnelles au regard du traitement des informations constituant les corpus oraux ? Quelles sont les responsabilités des personnes en charge de la diffusion des corpus sur Internet ?

2.3.1 LE DOMAINE PUBLIC ET LE DROIT D'AUTEUR

QU'EST-CE QUE LE DOMAINE PUBLIC ?

Si l'expression « domaine public » est généralement connue de tous, l'acception juridique du terme peut être entendue dans des sens différents qu'il est important de préciser pour éviter des ambiguïtés ou des incompréhensions lors de la constitution des corpus oraux. Au sens juridique, le domaine public est un concept multiforme qui peut renvoyer autant à un lieu, qu'à un régime ou à des contenus.

Le domaine public peut, ainsi, être « l'endroit où la société civile s'efforce d'influer sur la manière dont les biens collectifs sont gérés et distribués ». C'est dans ce sens que l'UNESCO est à l'origine d'une véritable politique des contenus et développe une stratégie de promotion d'un domaine public fort, accessible en ligne et hors-ligne. Le domaine public recouvre non seulement les idées de liberté d'accès et de gratuité d'utilisation des données, mais aussi la possibilité pour chacun de les exploiter. Il se caractérise, en outre, par l'absence de monopole, puisque les informations qui tombent dans le domaine public deviennent *de facto* des « choses communes ».

En revanche, deux types d'informations peuvent être distingués : celles qui sont nées dans le domaine public et celles qui y sont « tombées ». Les idées, la langue, les textes de loi et tous les éléments qui fondent le patrimoine commun d'une communauté donnée, constituent, de par leur nature, le « fonds commun » du domaine public. Ce fonds commun reste pourtant difficile à délimiter. Les enregistrements linguistiques suscitent ainsi de nombreuses hésitations. Mis à part les droits de celui qui a enregistré, le contenu d'une langue, son expression phonique, font-ils ou non partie du domaine public ? La question peut aussi se poser à l'égard des traditions et des

coutumes. En outre, ce fonds commun est-il universel ou bien seulement commun à une petite communauté ? Aujourd'hui, il fait de plus en plus l'objet de revendications identitaires qui soulèvent de nouvelles interrogations.

Au-delà d'un certain délai, les œuvres protégées par le *droit de la propriété intellectuelle, notamment par le droit d'auteur ou les brevets, finissent par entrer dans le domaine public. Le droit d'auteur, par exemple, protège les œuvres soixante-dix ans après la mort de leur auteur. En droit français, à l'expiration de ce délai, d'autres types de protection peuvent subsister sur les œuvres de l'esprit : les droits patrimoniaux d'une part ; les attributs imprescriptibles du *droit moral d'autre part. Par conséquent, certains éléments du domaine public peuvent encore bénéficier de la protection du droit moral.

Ces distinctions font apparaitre deux types de situations apparemment opposés : soit les corpus sont constitués d'œuvres du domaine public ne pouvant faire l'objet d'une appropriation (de par leur nature ou du fait de l'expiration du délai de protection) et de ce fait sont libres de droit, soit les corpus sont soumis au droit d'auteur et donc soumis aux autorisations requises. En réalité, nous l'avons vu, il existe une possibilité intermédiaire où les corpus protégés par le droit d'auteur peuvent être mis en libre accès dans le cadre d'une licence accordée par les titulaires de droits autorisant l'utilisation et l'exploitation des résultats. Sans être dans le domaine public, ces corpus sont – de par la volonté de leurs créateurs – libres d'accès et d'utilisation. Néanmoins, si les créateurs peuvent renoncer à exercer leurs *droits patrimoniaux, il ne leur est pas possible de renoncer à leur droit moral, qui reste imprescriptible.

Le droit d'auteur et les corpus

Quelles sont les conditions pour qu'un corpus soit protégé ? Il y en a trois.

Il faut en premier lieu qu'il corresponde à l'exigence d'une *activité créatrice* : un travail de compilation d'informations n'est pas protégé en soi.

Pour être protégé, il est par ailleurs indispensable que le corpus ait une *forme définie*. Ce qui est protégé, ce n'est pas le contenu du corpus mais son enveloppe, son architecture.

Enfin, la forme du corpus doit répondre à la condition d'être *originale*. Que signifie l'originalité d'un corpus ? L'originalité de nombreuses créations de l'ère du numérique, comme les logiciels ou les bases de données, ne peut être appréciée que d'après des critères objectifs. Il semble qu'il en soit de même des corpus oraux, ceux-ci pouvant le plus souvent être assimilés à une base de données. C'est alors, le plus souvent, le fait que le corpus soit ou non copié et révèle un minimum d'activité créative qui servira de critère pour déterminer s'il est ou non original (et non pas uniquement la prise en compte de l'empreinte de la personnalité de son auteur).

« Il n'y a pas de place pour les droits des auteurs quand il n'y a pas d'auteur »

L'auteur est en principe la (ou les) personne(s) physique(s) sous le nom de laquelle (ou desquelles) l'œuvre est divulguée. Le travail scientifique suppose l'intervention de nombreux acteurs dont bon nombre sont susceptibles de revendiquer la qualité d'auteur sur les résultats de la recherche.

Certains corpus oraux, comme les autres produits de la recherche, peuvent rester l'œuvre d'un auteur unique, alors que d'autres peuvent être l'œuvre de plusieurs auteurs. Dans le cas de pluralité d'auteurs, le droit distingue les œuvres de collaboration des œuvres collectives. Pour les premières, chaque co-auteur dispose des mêmes prérogatives. D'autres œuvres – telles que les bases de données ou les dictionnaires – peuvent être qualifiées d'œuvre collective lorsqu'elles sont créées

> *« sur l'initiative d'une personne physique ou morale qui l'édite, la publie et la divulgue sous sa direction et sous son nom, et dans laquelle la contribution personnelle des divers auteurs se fond dans l'ensemble »*[4].

Dans ce dernier cas, c'est la personne physique ou morale qui a pris l'initiative de l'œuvre qui dispose des droits d'auteur. Par ailleurs, le contexte de la création ou le statut de l'auteur peuvent avoir des incidences sur la détermination du titulaire des droits d'auteur. L'œuvre a-t-elle été créée dans le cadre d'une mission de service par un employé ou un fonctionnaire ? Quels sont les droits respectifs de l'auteur et de son employeur ? Si la question est résolue le plus souvent par le contrat de travail, elle reste plus délicate quand le créateur est un fonctionnaire. En effet, depuis plusieurs années, deux logiques s'affrontent, celle de la reconnaissance d'un droit de la personne créatrice d'une part, et d'autre part celle de la reconnaissance uniquement d'un droit de l'État sur les créations de fonctionnaires. La transposition de la directive sur les droits des auteurs dans la société de l'information a incité les pouvoirs publics à proposer une voie médiane qui reconnait à la fois le droit des auteurs et les droits de l'employeur « État » quand la création de l'œuvre s'inscrit dans l'exécution de la mission de service public. Si ce texte est voté par le Parlement, les droits des auteurs pourraient naitre sur la tête du fonctionnaire.

En contrepartie, tous les droits d'exploitation de l'œuvre pour les besoins de sa mission seraient cédés à son employeur État (droit de communiquer ou de diffuser pour la mission). Toutefois, dans le cas d'une exploitation commerciale, l'auteur personne physique recouvrera ses droits avec l'obligation d'accorder un droit de préférence à son employeur et la possibilité d'être intéressé à l'exploitation commerciale. Ce texte n'est pas sans soulever des débats et des interrogations. Comment sera déterminé le périmètre de la mission de service des chercheurs qui interviennent dans l'établissement du corpus ? Comment distinguer l'exploitation pour la mission du service et l'exploitation commerciale quand – nous l'avons vu précédemment – le chercheur a pour mission de communiquer les résultats de sa recherche et de les valoriser par la publication ?

QUELS DROITS POUR LES AUTEURS SUR LES CORPUS ORAUX ASSIMILABLES A DES ŒUVRES ?

Il convient de distinguer les droits patrimoniaux des prérogatives du droit moral. On rappellera aussi que la loi pose quelques limites aux droits exclusifs des auteurs.

Les droits patrimoniaux se résument en un droit exclusif au profit de l'auteur (ou des titulaires) ou des ayants droit (bénéficiaire d'une cession, héritiers...) d'autoriser ou

[4] Art. L. 113-2 du CPI.

interdire la reproduction ou la communication au public de l'œuvre protégée. Si le corpus oral est une œuvre, toute reproduction (la numérisation est pour le droit une reproduction) et toute mise à disposition du public (sur un site Internet comme sur tout autre support) nécessitent l'autorisation expresse de l'auteur ou du titulaire de droit.

Quant aux prérogatives du droit moral, toujours attachées à la personne physique créatrice de l'œuvre protégée, elles sont au nombre de quatre : *le *droit de divulgation, le *droit de repentir et de retrait, le *droit à la paternité* et *le *droit au respect de l'œuvre*. Chacun de ces droits est applicable aux corpus oraux. L'auteur du corpus (au titre de son droit de divulgation) peut décider du moment ou des modalités de la mise à disposition du corpus au public, le dépôt aux archives ne valant pas nécessairement divulgation. Un corpus inédit ne peut donc être mis à la disposition du public sans l'autorisation de son auteur. Le chercheur auteur qui refuse de divulguer le corpus qu'il a créé est dans son droit (au titre du droit d'auteur), même si par ailleurs il peut être sanctionné administrativement pour ne pas avoir exécuté sa mission de service public qui est de communiquer les résultats de sa recherche. Le droit de repentir ou de retrait peut s'exercer aussi sur un corpus oral, ces regrets ne pouvant porter que sur le contenu intellectuel de l'œuvre et non pas sur les conditions matérielles de sa diffusion. Si le droit à la paternité est en soi facile à comprendre, on peut se demander ce que signifie le droit au respect de l'œuvre appliqué à un corpus oral. Ce droit correspond autant au respect de la forme de l'œuvre (pas de suppression, d'adjonction ou de modification…) qu'au droit au respect de l'esprit de l'œuvre (altération de la finalité du corpus).

Comme tout monopole, les droits exclusifs des auteurs souffrent des limites. On peut en premier lieu rappeler qu'ils sont limités dans le temps et qu'au-delà de cette limite, les œuvres tombent dans le domaine public (cf. *supra*). Ces limites peuvent aussi trouver leurs justifications dans le type d'usage qui est fait des œuvres. On parlera alors d'exceptions au droit d'auteur qui sont justifiées par les finalités ou le contexte ou encore l'intérêt général.

Enfin, *le droit à la copie privée* ou le droit de citation concernent directement les corpus oraux (voir fiche *Droit de citation*).

2.3.2 LE RESPECT DE LA VIE PRIVEE

LE RESPECT DE LA VIE PRIVEE DANS LA CONSTITUTION, L'EXPLOITATION, LA DIFFUSION ET LA CONSERVATION DES CORPUS

La création d'un corpus passe le plus souvent par la collecte de données. Celles-ci pouvant être des données personnelles, cette collecte doit être faite dans le respect de la loi *Informatique et libertés* : licéité et loyauté, information préalable, obtention du consentement des personnes concernées (voir fiche *Consentement*), respect des finalités annoncées[5]… Quand il s'agit de finalités de recherche, faut-il entendre de façon restrictive une recherche spécifique identifiée comme telle, ou peut-on entendre de façon plus large l'expression « finalités de recherche » ? Le problème se

[5] http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17_definitive-annotee.pdf

pose quand, une fois le corpus constitué et exploité scientifiquement par les chercheurs qui ont été à l'origine de sa création, on envisage une réutilisation et de nouvelles exploitations scientifiques. La recherche scientifique bénéficie aujourd'hui d'une exception au principe général avec l'application de ce que l'on appelle *l'extension de finalité*. Toutefois, toute nouvelle exploitation scientifique devra se faire en respectant les formalités préalables à tout traitement (nouvelle procédure de déclaration ou d'autorisation) et les principes posés par la loi (information, consentement et/ou autres garanties appropriées…).

Même si la diffusion des corpus et leurs nouvelles exploitations sont faites dans les conditions requises, se pose le problème de la conservation des données personnelles.

Si les données sont « anonymisées » de manière irréversible, elles sortent du champ de la loi et peuvent être conservées (voir fiche *Données personnelles et anonymisation*). Toutefois dans la recherche, le besoin de « traçabilité » nécessite souvent de sauvegarder les données personnelles.

Et pourtant, en principe, sur le fondement du **droit à l'oubli*, les données personnelles ne doivent pas être conservées au-delà de la durée initialement prévue, et quand la finalité initiale annoncée lors de la collecte de ces informations n'a plus de raison d'être, ces données doivent être détruites. Cela veut-il dire qu'il n'est pas possible de conserver certains corpus contenant des données personnelles si celles-ci n'ont pu être anonymisées ? Non, mais il ne peut s'agir que de cas exceptionnels où le maintien des données personnelles se justifie pour des raisons scientifiques. Dans ces cas, les corpus oraux pourraient bénéficier – en tant qu'archives publiques – d'une dérogation au droit à l'oubli permettant leur conservation au-delà de la durée prévue, en vue d'un traitement à des fins de recherche, historique ou scientifique. C'est alors la loi sur les archives qui fixera les conditions de leur mise à disposition en libre accès (délais plus ou moins longs – 60 à 150 ans[6] – suivant le degré de sensibilité des données contenues dans le corpus).

QUELLES SONT LES RESPONSABILITES DES PERSONNES CHARGEES DE LA DIFFUSION DES CORPUS SUR INTERNET ?

La diffusion des corpus oraux sur Internet peut être assimilée à « l'édition d'un service de communication au public en ligne ». Il est donc important d'apprécier les obligations et responsabilités des éditeurs d'un service de communication au public en ligne (voir fiche *Responsable du traitement*).

[6] Voir art. 213-2 du code du patrimoine (ancien article 7 loi de 1979).

3 LA DEMARCHE
CONSTITUTION, EXPLOITATION, CONSERVATION, DIFFUSION

3.1 EXPLICITER LA DEMARCHE

Les objectifs, notamment scientifiques, liés à la constitution, à l'exploitation, à la conservation et à la diffusion des corpus oraux sont très diversifiés, et le respect de ceux-ci, ainsi que leur hétérogénéité, impliquent que soit reconnue la diversité des *démarches* qui peuvent être adoptées par les chercheurs et par les responsables de la diffusion et de la conservation de ces corpus.

Le *Guide des bonnes pratiques* n'a pas vocation à contraindre cette démarche en prescrivant une méthodologie type, mais souhaite fournir toutes les informations nécessaires au repérage des points juridiques et éthiques « sensibles ». Seule l'identification précise et détaillée des éléments de la situation en jeu et notamment de la forme des données et de leurs supports, des pratiques de terrain, mais aussi des différentes étapes de leurs traitements, permet d'apporter à la fois des éléments de réponses juridiques correspondant à la situation, et une évaluation des « risques » éventuels. Enfin, une analyse réflexive sur la démarche liée à la constitution et aux traitements des corpus oraux est le premier élément de l'élaboration d'une éthique reconnue par l'ensemble d'une communauté scientifique.

3.2 ÉLEMENTS DE LA SITUATION EN JEU

Les enregistrements qui constituent les données primaires de l'enquête linguistique sont loin de former un objet uniforme. Ainsi, un conte enregistré sur une bande magnétique lors d'une cérémonie traditionnelle sur la place d'un village est un objet scientifique et patrimonial fort différent de l'enregistrement numérique d'un texte lu par un « informateur rémunéré » dans les locaux d'un laboratoire universitaire, des réponses à un questionnaire enregistrées sur minidisque par un chercheur au domicile de la personne interrogée ou bien encore d'une conversation spontanée non sollicitée par les chercheurs, se déroulant dans un café et filmée par une ou plusieurs caméras.

Il convient donc, dans un premier temps, d'identifier les éléments qui caractérisent les données récoltées en situation :

- le *type de données* qui constitue le corpus et leurs supports (d'enregistrement, mais aussi de stockage pour exploitation, et de conservation),
- les *différentes techniques* employées par les chercheurs pour récolter les données,
- la définition des *participants* et de leur rôle,
- la catégorisation des *lieux* de la collecte.

3.2.1 CORPUS ET TYPE DE DONNEES

Si la volonté de « capturer » la parole est fort ancienne, c'est récemment que les avancées technologiques et la recherche (notamment en linguistique) ont permis de concevoir les enregistrements comme de véritables « données ». Ainsi, l'Alphabet Phonétique International est un exemple de système « d'enregistrement »

alphabétique inventé par des linguistes afin de normaliser le codage de la transcription phonétique et/ou phonologique de la parole. L'histoire moderne des enregistrements audio et vidéo se déroule au fil des transformations des modes d'enregistrement comme à travers celui des supports d'inscription utilisés.

Les modes d'enregistrement

Le mode d'enregistrement analogique a été le premier à être utilisé pour l'enregistrement et la conservation du son. Il code les variations mesurées sous forme de signaux obéissant à la même loi de variation que celle qui régit leur propagation dans un milieu naturel. Depuis quelques décennies, c'est plutôt un mode d'enregistrement numérique qui est privilégié. Dans ce mode, les mesures ponctuelles de la pression de l'air sont régulièrement effectuées (échantillonnage). Ces mesures sont ensuite codées sous forme d'une valeur numérique exprimée dans une échelle de référence puis sont représentées sur le support de stockage sous la forme d'une suite organisée d'unités binaires.

Les supports d'enregistrement

- *Supports physiques*

Les premiers supports modernes permettant la conservation de la parole ont été les supports physiques. Ce terme est dû au fait que les variations de pression mesurées par un appareil (microphone) sont inscrites physiquement dans la matière du support. On compte parmi eux les anciens cylindres, les disques vinyles, etc. Ces supports conservent dans la matière qui les compose (vinyle, cire, etc.), sous la forme d'un sillon ondulé, une image analogique des variations de pression mesurées. Ces supports utilisés au siècle dernier sont pratiquement abandonnés. Ils posent aujourd'hui des problèmes d'accès et de conservation.

- *Supports magnétiques*

Les supports magnétiques sont apparus plus tardivement, dans la deuxième moitié du 20e siècle. Différents supports de stockage ont été et sont encore utilisés de nos jours (fil, bande, disque) dans différents conditionnements (bobine, cassette, cartouche, etc.). Le principe ici repose sur la rémanence des particules magnétiques réparties tout le long du support (i.e. la propriété qu'ont ces particules de conserver durablement leur aimantation). Cette aimantation des particules pourra, suivant les modes d'enregistrement, coder des informations sous forme binaire (comme dans les supports disque-dur, cassette DAT, disquette informatique, etc.) ou bien des informations sous forme analogique (comme dans les mini-cassettes audio, les cassettes VHS, etc.). Une partie de ces supports est destinée à être utilisée sur du matériel informatique, une autre sur du matériel audio/analogique. Ici encore, comme pour tout support, ceux-ci se dégradent inexorablement au cours du temps. Ces supports demeurant encore très populaires, l'accès aux outils qui en permettent la lecture et l'écriture reste aisé.

- *Supports optiques*

Les supports optiques sont les derniers apparus ; ils sont connus principalement dans leur forme de Compact-Disc (CD-audio, CD-ROM, etc.). La technologie repose sur les propriétés optiques des composants, à savoir par exemple pour les

CD-audio, la capacité des alvéoles qui les composent de réfléchir la lumière d'un faisceau laser. Ces supports sont principalement utilisés pour stocker des données numériques (exception faite de certains disques laser peu populaires, et des films argentiques peu utilisés pour l'enregistrement sonore). Une grande partie de ces supports est destinée à une utilisation sur des équipements informatiques, ce qui facilite l'accès, le transfert et le traitement des données. Les problèmes de conservation sont les mêmes que pour tout type de support, même s'ils ne sont pas sensibles aux mêmes agressions (lumière, chaleur, champs magnétique, humidité, etc.). Comme les supports magnétiques, ils ont l'avantage d'être récents et populaires, ce qui rend leur utilisation facile aujourd'hui.

Il existe d'autres types de supports mélangeant par exemple les techniques optiques et magnétiques. (voir fiche *Supports pour enregistrer et archiver le son*).

LES CRITERES DE CHOIX

La conservation des supports posant de toute façon des problèmes similaires quel que soit le type de support choisi, les critères de choix du bon support d'enregistrement puis de conservation reposeront plutôt sur la qualité du codage, la facilité d'accès et de traitement ainsi que sur la possibilité de reproduire son contenu sans perte d'information. On privilégiera donc les supports numériques par rapport aux supports analogiques, car ils peuvent être dupliqués à l'identique et à l'infini. On privilégiera aussi les supports informatiques en raison de la panoplie des outils que l'informatique offre pour la gestion, l'accès à du matériel de lecture, la diffusion et le traitement des données (cryptage, techniques d'anonymisation, etc.) tout en considérant que ces outils posent encore de nombreux problèmes de standardisation (par exemple en ce qui concerne le choix des logiciels, des formats, des codecs de compression). Enfin, pour la conservation, un support qui ne peut pas être effacé est aussi, peut être, une bonne garantie pour éviter les accidents malencontreux.

Le choix d'un format qui permet la reproduction à l'identique garantit une forme de pérennité aux données. Elle met en cause la notion même d'« original » qui se réfère alors moins au support qu'aux données elles-mêmes.

STANDARDISATION DES ANNOTATIONS

Les corpus oraux sont en général composés d'enregistrements audio ou vidéo et d'annotations de ces derniers.

- *Données primaires vs. données secondaires*

On distingue généralement entre données primaires et données secondaires :

- les *données primaires* sont constituées par les *enregistrements*, ayant un lien le plus proche possible avec l'évènement documenté. Elles comprennent aussi les autres objets recueillis dans le contexte de l'action, comme les documents lus ou écrits durant l'action enregistrée, les objets manipulés, les images consultées, etc. Elles comprennent aussi les traces informatiques laissées par l'activité.
- les *données secondaires* sont constituées par la série de descriptions, transcriptions, annotations qui viennent enrichir les données primaires et qui sont souvent fournies après coup et sur la base des données

primaires. Elles comprennent aussi les métadonnées, les conventions de transcription, les autorisations des participants, etc.

La distinction entre données primaires et données secondaires est utile notamment pour différencier des niveaux d'interprétation et souligner l'importance du retour aux données primaires et de leur disponibilité. Ainsi une analyse porte sur la bande audio ou vidéo et non exclusivement sur la transcription, même si celle-ci est un adjuvant important sans lequel l'analyse serait probablement impossible. C'est dans ce sens que sont développés les outils d'alignement entre la source sonore/visuelle et le texte de la transcription. Toutefois, cette distinction entre données primaires et données secondaires a ses limites : elle ne doit pas faire oublier le fait que tout enregistrement est le fruit de décisions à la fois techniques et théoriques – concernant par ex. le choix du moment à enregistrer et la délimitation du segment enregistré, le choix du cadrage et de l'optique pour la vidéo, du positionnement et de l'orientation du micro pour l'audio – qui reposent sur une connaissance préalable de l'activité enregistrée. Les « données » ne sont jamais « offertes » ni « (re)cueillies » mais elles sont activement produites par les chercheurs (Mondada, 2006).

- *Explicitation de la structure des données*

Pour l'écriture des annotations, on utilise des formalismes d'expression qui permettent à la fois de coder le contenu des commentaires ainsi que d'expliciter de quel type de commentaire il s'agit. Par exemple, dans les bases de données relationnelles, on va utiliser des tables comportant des champs avec des noms (i.e. *pos* pour « *part of speech* ») qui vont servir à stocker des valeurs (par ex. « verbe ») exprimées dans des types particuliers de structure (chaîne de caractères, nombre, etc.).

Un formalisme alternatif et très largement utilisé dans le domaine de l'annotation textuelle est celui apporté par la grande famille des langages de balisage de textes. Ce formalisme délimite les commentaires par des marques formelles (i.e. balises) indiquant de quel type de commentaire il s'agit. Il existe aujourd'hui un consensus assez vaste, toutes disciplines confondues, sur l'adoption du récent langage de balisage de texte XML comme formalisme de structuration et d'échange de documents (voir fiche *Codages et formats*).

- *Standardisation/Normalisation*

Alors que le choix d'un formalisme permettant d'exprimer l'ensemble des annotations ainsi que d'expliciter leur structure est indispensable, il n'est pas pour autant suffisant pour permettre l'échange ou la conservation d'un document. Pour échanger ou conserver un document, il faut que le langage utilisé pour coder sa structure ainsi que le contenu de ses commentaires soit commun entre les participants (dans le cadre d'un échange) ou qu'il puisse rester connu au cours du temps (dans le cadre d'une conservation à long terme). Dans le contexte d'un document utilisant un langage de balisage de textes, les noms des éléments de structure (balises, attributs...) doivent être connus et leur définition acceptée et partagée, ainsi que l'ensemble des contraintes (enchainement de balises, vocabulaires contrôlés, caractère optionnel ou obligatoire de certaines structures…).

Quand un grand nombre de personnes ou toute une communauté parviennent à s'entendre sur un langage commun, on parle de standardisation. C'est ce qui s'est passé par exemple avec l'Alphabet Phonétique International (API). Alors que la standardisation est nécessaire pour l'échange, la conservation à long terme réclame des garanties sur la transmission et sur l'accès à la documentation des langages communs mis en place. A ce titre, les organismes de normalisation doivent pouvoir apporter une certaine pérennité aux normes qu'ils mettent en place, ainsi qu'une indépendance vis-à-vis des intérêts privés. Ils doivent aussi être représentatifs de l'intérêt général. A ces conditions, il sera avantageux, partout où elles existent, d'utiliser des normes pour le codage et le formatage des données. On peut citer à titre d'exemple le codage des caractères ISO-10646, plus connu sous le nom d'Unicode, qui est un code-caractère qui se veut universel et prend en compte la plupart des écritures du monde, y compris l'Alphabet Phonétique International. Pour le codage de l'analyse linguistique, il sera intéressant de lire les recommandations de la Text Encoding Initiative (TEI) qui propose des analyses pour des structures de données telles que les dictionnaires, les poèmes ou la transcription de la parole. Il sera aussi très utile de suivre les progrès du groupe de travail de l'ISO sur la gestion des ressources linguistiques TC37 SC4 (voir fiche *Codages et formats*).

Ainsi, les principes qui doivent guider le choix d'une technologie plutôt que d'une autre pour l'annotation peuvent être résumés en quatre questions :

- Cette technologie permet-elle de *coder de manière explicite* toutes les annotations ?
- Cette technologie présente-t-elle un *caractère propriétaire* ou *une limite légale* qui empêcheraient de partager les annotations avec d'autres (formats propriétaires, techniques basées sur des brevets, etc.) ?
- Cette technologie est-elle *acceptée par la communauté* avec laquelle l'échange des données est envisagé ?
- Cette technologie a-t-elle fait l'objet d'une *normalisation* ?

3.2.2 TECHNIQUES D'ENQUETE

RECUEIL ET PRODUCTION DE DONNEES

Les enquêtes linguistiques n'ont pas toujours donné lieu à des enregistrements pour des raisons techniques (les premiers outils d'enregistrement de la parole ont à peine plus d'un siècle) mais aussi méthodologiques et théoriques. Ainsi, les questionnaires écrits, la prise de notes, le recours à l'intuition et/ou à l'observation du chercheur ont et sont encore des outils de description utilisés par les linguistes. La possibilité d'enregistrer la parole et l'évolution des techniques (miniaturisation des appareils, qualité du signal enregistré, numérisation et traitements informatiques des données sonores et vidéo), ont néanmoins permis aux enquêtes de terrain de développer des méthodologies qui restent toutefois très différentes ne serait-ce que par la diversité des domaines scientifiques concernés (dialectologie, sociolinguistique, analyse conversationnelle, psycholinguistique, linguistique de l'oral, traitement automatique de la parole, ethnolinguistique…). Cependant les recherches sur la méthodologie de l'enquête ont conduit les chercheurs à considérer les données enregistrées comme

étant le produit de la situation d'enquête par opposition à une conception de données préexistantes simplement (re)cueillies (Cameron *et al.*, 1991).

Enfin, les techniques d'enquête ont un rôle important dans la possibilité qu'elles offrent (ou qu'elles n'offrent pas) de contrôler les données fournies aux chercheurs par la personne interrogée. La suite de ce chapitre est consacrée à un inventaire des différentes techniques d'enquête utilisées lors de la constitution de corpus oraux.

Le questionnaire

Le questionnaire oral enregistré peut revêtir différentes formes ; il est le plus souvent composé de questions fermées ou semi-ouvertes et de listes de termes lexicaux ou de textes préparés par le chercheur. Seul le cas des textes préexistant au questionnaire peut poser éventuellement la question de la propriété intellectuelle (comme par exemple la lecture d'un texte protégé par le droit d'auteur, ou une production dont l'originalité du contenu serait protégée). Dans les autres cas il s'agit de capter, notamment, les variations, les régularités et les perceptions de ces régularités par le questionné, en référence à un système linguistique commun.

Le degré de sensibilité des informations collectées est le plus souvent prévisible, puisque c'est le chercheur qui élabore le questionnaire et qui peut donc évaluer les risques selon la nature des questions. Toutefois, des questions apparemment anodines peuvent aussi receler des enjeux, insoupçonnés par le chercheur, surgissant du contexte particulier de l'enquête. Il convient en outre de souligner que le questionnaire contient plus que toute autre technique la marque de l'acte de questionnement et de la *prise* du chercheur (Encrevé, 1983) et donc potentiellement un sentiment d'évaluation, même si celui-ci est souvent atténué par la possibilité explicitement offerte de ne pas répondre à tout ou partie des questions (pour une analyse des situations de questionnaire cf. Achard 1991). Enfin, soulignons ici un point qui concerne de nombreuses situations d'enquête, mais qui est particulièrement lié au questionnaire : celui-ci contient souvent une partie consacrée au recueil de données personnelles (âge, catégorie socio-professionnelle…) dans le but de dresser le profil sociologique de l'enquêté.

L'entretien

L'entretien est composé de questions ouvertes, l'objectif étant principalement de recueillir une quantité importante de données linguistiques. L'entretien suppose toujours un guidage de la part de l'enquêteur, qui peut être plus ou moins fort (de l'entretien directif au semi-directif, voire au non-directif ; du plus standardisé au moins standardisé), le rapprochant ainsi du questionnaire oral ou de l'interaction moins contrainte (Maynard *et al.*, 2002, Houtcoop-Streenstra, 2000). Bien que dans l'entretien le chercheur introduise souvent les catégories et les thèmes qu'il souhaite voir traités par les informateurs, la méthodologie des chercheurs peut aussi requérir, par souci de collecter les productions les plus naturelles possibles, que l'objet de la recherche ne soit pas précisé en détail avant l'entretien, et pose donc le problème du choix du moment et du contenu des informations fournies aux interviewés (cf. Mondada 2001).

Du point de vue juridique, les entretiens sont le plus souvent des sources de données et d'informations concernant la vie privée de l'interviewé ou de personnes mentionnées dans le cours de l'entretien et sont donc à protéger en tant que tels.

Le recueil de contes, chants…

Le recueil de contes, chants et productions orales de cultures traditionnelles est une pratique fréquente dans les domaines de la description des langues à tradition orale et de l'ethnolinguistique notamment. Outre l'importance de contextualiser ces chants, contes et récits (des significations implicites dans un contexte culturel peuvent échapper ou paraitre anodines dans un autre), deux éléments sont principalement à prendre en compte : la propriété intellectuelle de productions traditionnelles d'une communauté, et les conditions de recueil, souvent liées à des activités sociales dans un cadre public ou privé.

Les recits de vie

Les récits de vie sont couramment sollicités lors de recherches en anthropologie, en histoire, en ethnolinguistique, mais aussi en dialectologie, et dans de nombreux autres domaines (Guillaumou *et al.* 1997). Ces types d'enregistrement représentent nécessairement une source importante de données personnelles concernant l'auteur du récit et de tierces personnes, qui peuvent éventuellement être associées à un contexte social ou historique particulièrement sensible, notamment quand le récit personnel fait écho à un évènement vécu par une ou plusieurs communautés.

Ainsi, même dans le cas de recherche sur des phénomènes exclusivement linguistiques, les propos contenus dans des récits de vie et la question de l'impact de leur diffusion dans l'espace public ne peuvent échapper à la responsabilité du chercheur qui les sollicite et les exploite.

De plus, les conditions d'exploitation et de diffusion de ces récits peuvent se faire dans un contexte social très différent de celui, très particulier, qui a marqué le recueil et qui a souvent lieu dans un cadre précis et grâce à une relation privilégiée entre le chercheur et le témoin.

Enfin la question de la propriété intellectuelle d'un récit de vie et du droit moral inaliénable peut s'avérer particulièrement pertinente dans le cas de récits originaux.

L'enregistrement en laboratoire

Les enregistrements en laboratoire selon un protocole expérimental sont utilisés en sciences du langage notamment dans les domaines de la psycholinguistique, de la phonétique et du traitement automatique de la parole. Ainsi certains corpus intéressent directement la recherche appliquée et les entreprises concernées par l'ingénierie linguistique, et font donc parfois l'objet de financement partiel ou total sur des fonds privés.

De même que pour les questionnaires, sauf dans les cas particuliers de textes soumis aux droits d'auteur, les productions des participants selon un protocole expérimental élaboré par les chercheurs ne semblent pas devoir être concernées par le droit de la propriété intellectuelle (sauf les cas particuliers). La situation particulière de la personne enregistrée reste toutefois à rapprocher de tous les cas de recherches expérimentales sur personne humaine.

L'enregistrement d'activites provoquees

Il s'agit principalement d'enregistrements d'activités dans le contexte ordinaire des acteurs sociaux concernés, même si les consignes proviennent du chercheur (activités proposées à des enfants en milieu scolaire, tâches simulées en situation professionnelle, etc.). Cette situation combine à la fois les caractéristiques d'enregistrements selon un protocole expérimental (qui est de la responsabilité du chercheur) et les caractéristiques du contexte ordinaire en milieu écologique ; elle offre donc un double cadre contrôlable par le chercheur. Cette intervention explicite du chercheur (dont le rôle peut être clairement identifié par les participants) facilite les conditions d'obtention d'un consentement éclairé ; toutefois une attention particulière doit être apportée au milieu professionnel qui peut contraindre le consentement (confidentialité…).

L'enregistrement d'activites dans leur contexte ordinaire

Les recherches en sociolinguistique, analyse conversationnelle et analyse des usages des technologies (Computer Supported Cooperative Work ; Dialogue Homme Machine), s'intéressent au recueil de données en situation d'activité non orchestrée par le chercheur et non provoquée par ses consignes. Il s'agit ici d'activités telles qu'elles ont lieu de manière ordinaire, même en l'absence du chercheur. Ces activités peuvent être fort variées : réunions, activités professionnelles, demandes de renseignements, interactions téléphoniques, etc. Les techniques de collecte sont également très différentes. Elles vont de l'observation participante à l'enregistrement autorisé, en passant par l'utilisation « de personnes ressources » choisies au sein du groupe de pairs observés et en particulier chargées de porter le dispositif d'enregistrement (micro, éventuellement caméra).

L'objectif partagé par ces techniques est la recherche de données en situation naturelle et suppose donc une méthodologie s'efforçant de minimiser les effets produits par les dispositifs d'enregistrement (Heath, 1997 ; Jordan & Henderson, 1995). Il y a donc de fortes probabilités pour que les données contiennent des informations sensibles au regard de la protection de la vie privée. Les modalités du recueil du consentement doivent en tenir compte et s'y adapter.

La reprise d'enregistrements

Certains corpus constitués d'enregistrements produits par des acteurs différents des enquêteurs pour des finalités autres que scientifiques ou autres que les finalités évoquées lors du recueil de consentement peuvent donner lieu à une reprise dans un but de recherches linguistiques ou à une mise à disposition du public à des fins patrimoniales, mémoriales ou politiques (c'est ainsi que par exemple la Brigade des Pompiers de New York a mis à disposition en aout 2005 les enregistrements des communications radio durant l'attentat du 11 septembre 2001). Ces corpus sont donc caractérisés par l'absence de consentement pour la nouvelle finalité et par le fait que les propos archivés n'ont pas été produits en connaissance de cette finalité, mais dans un autre cadre et avec d'autres objectifs. Ainsi, lors d'interviews ou de séminaires – enregistrés par exemple dans le but d'une diffusion des contenus transcrits – l'autorisation de diffusion peut concerner les propos transcrits et validés, et non une reprise ultérieure des enregistrements.

LA REPRISE D'ENREGISTREMENTS MEDIATIQUES

La reprise d'enregistrements médiatiques est un cas particulier de la catégorie précédente, qui offre la particularité de concerner des données produites dans un cadre de diffusion publique.

Là encore, si le contenu des enregistrements est protégé par le droit d'auteur (par exemple dans le cas d'une production originale), le recueil du consentement est un préalable à toute exploitation. Une exception existe toutefois pour un laps de temps déterminé, lorsqu'il s'agit de discours destinés au public et prononcés en public, tels que spécifiés dans les lignes suivantes :

> *Code de Propriété Intellectuelle, art 122.5 :*
>
> *La diffusion, même intégrale, par la voie de presse ou de télédiffusion, à titre d'information d'actualité, des discours destinés au public prononcés dans les assemblées politiques, administratives, judiciaires ou académiques, ainsi que dans les réunions publiques d'ordre politique et les cérémonies officielles*[7].

Rappelons que l'enregistrement personnel d'une émission correspond à une licence légale pour cette copie strictement réservée à l'usage privé. La représentation de celle-ci ne peut se faire que dans le cadre du « cercle de famille ». De même pour une cassette ou un dévédé du commerce, le droit de copie (avec ses limites) ne donne aucun droit d'exploitation.

Enfin précisons que le caractère public du contexte de diffusion médiatique ne signifie pas une restriction de la protection des données personnelles.

La diversité des techniques utilisées pour la collecte de données, défini autant de *situations* qui mettent en évidence des *participants* dont le rôle est le premier élément de catégorisation.

3.2.3 ROLE DES PARTICIPANTS

Les participants à l'enquête et aux activités enregistrées sont catégorisables de différentes manières, qui toutes éclairent de façon spécifique ce qu'ils font et ce qu'ils disent (Sacks, 1972). Ainsi les participants à une situation d'enregistrement peuvent-ils être à la fois considérés comme des enquêtés (si l'on rapporte la situation au fait qu'elle est un objet d'enquête) et comme des acteurs sociaux – dont la caractérisation précise dépend du contexte, de l'activité, des formes d'engagement et de participation, impliquant à la fois l'histoire sociale des personnes et l'accomplissement local de leur rôle, mais aussi de leur identité durant la rencontre. Selon la manière dont les chercheurs eux-mêmes traitent ces multiples catégories, différentes conséquences peuvent apparaitre à la fois pour l'objet de l'enquête et pour l'évaluation du caractère plus ou moins sensible de l'activité.

CATEGORIES DE PARTICIPANTS

La terminologie très variée utilisée dans la littérature pour définir les catégories de participants à une enquête révèle des implications éthiques et théoriques diverses

[7] Art. 122.5 du code de la propriété intellectuelle.

(Cameron *et al.*, 1991). Voici une liste non exhaustive des termes utilisés dans différents contextes de recherche pour caractériser les participants du point de vue de leur engagement dans l'enquête :

– informateurs,
– locuteurs,
– sujets,
– « cobayes »,
– natifs,
– acteurs sociaux,
– participants,
– collaborateurs,
– partenaires,
– enquêtés,
– témoins.

Ces choix terminologiques sont le plus souvent le produit de considérations théoriques et politiques qui révèlent le type de relations préexistantes, construites, ou développées entre l'enquêté et l'enquêteur.

Si nous ne pouvons développer ici les enjeux de ces considérations théoriques, il est néanmoins important de repérer les marques d'une *relation* particulière qui fonde différentes réalisations de la *paire* enquêté/enquêteur, impliquant différents droits et obligations selon les caractéristiques de cette relation (Sacks, 1972).

Deux éléments définissent notamment cette relation : la proximité/distance des participants et les rôles en action et en situation.

- *Proximité/distance*

La question de l'accessibilité des situations enquêtées pour le chercheur s'est posée depuis toujours et a motivé différentes formes de *fieldwork* (travail de terrain), allant de l'immersion dans une communauté totalement étrangère au chercheur à l'exploitation de ses liens d'appartenance à sa communauté.

Ces problèmes ont été traités en termes de paradoxe de l'observateur - selon lequel le phénomène enquêté se dissout dès qu'il est observé (tel le vernaculaire pour Labov, 1972) - aussi bien qu'en termes de violence symbolique entre l'enquêté et l'enquêteur (Bourdieu, 1993). Ils ont aussi été traités en termes de *réflexivité* - par des chercheurs intégrant leur présence et celle du dispositif d'enquête dans l'analyse de l'objet enquêté (en anthropologie notamment, Clifford & Marcus, 1986 ; Mondada, 1998).

Les enquêtes chez les « proches » du chercheur, lorsque celui-ci exploite ses propres réseaux pour un travail d'enquête, facilitent les prises de contact et l'accès au terrain, tout en posant souvent des problèmes d'indistinction entre les relations dictées par l'enquête et les relations personnelles. Ces questions ne se posent pas dans le cas des enquêtes chez les « lointains » (communauté observée, panel échantillonné, témoins non sélectionnés par le chercheur,…) où les difficultés d'accès peuvent être supérieures mais où une fois gagnée la confiance et établie une relation, l'enquêteur a un statut souvent plus clair et mieux reconnu en tant que tel (Beaud & Weber, 1977).

La recherche en sciences sociales et humaines a en outre souvent utilisé des « populations captives », dans le sens où l'enquêteur dispose d'un accès facilité par des institutions (l'école, l'hôpital...) et où ces populations ont des possibilités limitées de refuser de collaborer (enfants, élèves…). Dès lors, une attention particulière doit être consacrée à l'approche de populations telles que :

– les personnes défavorisées,
– les personnes handicapées,
– les enfants,
– les élèves et étudiants,
– les employés d'entreprises ou d'institutions contactés par le biais de leur hiérarchie,
– etc.

L'usage est alors de doubler les autorisations pour les personnes par l'autorisation d'un responsable légal (enfants relayés par les adultes).

Ce cas particulier montre que l'autorisation signée ne peut pas toujours être considérée comme un acte suffisant et qu'il convient de protéger certains enquêtés au-delà de ce qu'ils ont signé (responsabilité de l'enquêteur).

- *Rôles en situation*

Lors de la situation d'enquête, et selon les techniques utilisées, la relation enquêteur–enquêté peut prendre des formes très différentes et impliquer des engagements plus ou moins directs.

- *Rôles de l'enquêteur*

– *observateur extérieur,*
– *observateur participant,*
– *observateur engagé* (défendant la communauté),
– *membre de la communauté* participant à une recherche-action (projet émanant de la communauté ou tenant compte de ses problèmes et objectifs, et y intervenant par une démarche spécifique sous la forme d'une « recherche-action »),
– *observateur déguisé* (*cross-dressing* dans la tradition ethnographique) s'intégrant dans la communauté par le biais de relations, d'un travail ou d'une fonction, mais ne déclarant pas son identité d'enquêteur,
– « *magicien d'Oz* » : enquêteur qui se dissimule derrière un dispositif technologique qui est censé répondre à l'informateur.

- *Rôles des enquêtés*

– *enquêté*/informateur/locuteur focalisé
– « *périphériques* » : les techniciens, les passants, les spectateurs...
– *associés* aux participants ratifiés à l'enquête (ex. clients appelant un centre d'appels, ou bien époux de la femme interviewée)
– le « *compère* » : informateur privilégié qui porte l'outil enregistreur et qui permet à l'enquêteur de pénétrer un groupe dont le compère fait partie ou auquel il a accès.

Ces rôles rendent compte notamment des variations possibles entre participation et observation, dans la tension entre les deux reflétée par le terme d'« observation participante » (Becker, 1960 ; Platt, 1983 ; Spradley, 1980). Selon ces rôles (Adler, 1987), l'engagement par rapport à l'enquête et aux enregistrements sera très différent, ainsi que les modalités de contact pour l'obtention d'un consentement éclairé.

3.2.4 LIEUX

L'information sur le lieu de la collecte conditionne des éléments de réponses juridiques particuliers de par ses propres caractéristiques et le rôle qu'il tient dans la situation d'enquête.

Ainsi on peut tout d'abord différencier les *lieux publics,* au sein desquels l'activité scientifique d'enregistrement audio-vidéo ne requiert pas d'autorisation autre que celle de la personne enregistrée, et les *lieux privés,* soumis à l'autorisation préalable du propriétaire/responsable qui est distincte du recueil du consentement de l'enquêté.

Le lieu peut également être défini selon la relation que les participants établissent. S'agit-il d'un lieu où la présence de la personne enregistrée est du fait de l'enquêteur (laboratoire, salle d'enregistrement…) ou est-ce celui-ci qui se déplace sur le terrain et investit donc l'espace propre de l'enquêté ?

Enfin, le lieu d'enregistrement peut être intégré aux données (caractéristiques audios ou visuelles présentes dans les données) ou ne relever que d'une information éventuellement présente parmi les métadonnées.

3.3 PRATIQUES DE TERRAIN

Ce chapitre a pour but de montrer l'omniprésence des enjeux éthiques et juridiques dans les étapes qui constituent la démarche de terrain ayant pour fin la constitution de corpus de données orales, interactives et multimodales. Nous insisterons notamment sur les phases *préparatoires* de l'enquête, préalables à l'enregistrement des données, où il s'agit notamment d'établir une relation avec les personnes concernées : ces modes d'approche sont étroitement liés non seulement aux méthodologies d'enquête (cf. *supra* 3.3.2) mais aussi aux possibilités et limitations techniques du dispositif d'enregistrement choisi, dont dépendent les contraintes spécifiques pour les autorisations à effectuer un enregistrement. Une fois terminée l'enquête et analysées les données, il s'agit d'organiser le *retour sur le terrain* pour différentes formes de « rendu » des résultats et des expériences – retour qu'il vaut mieux anticiper et qui configure le type d'engagement pris envers les personnes concernées.

3.3.1 MODES D'APPROCHE

Les enquêtes dont la finalité est le recueil de données enregistrées dépendent nécessairement de la qualité de la relation avec les personnes ressources – qu'on les appelle des informateurs ou des partenaires (cf. *supra* 3.2.3). La mobilisation de ces personnes varie selon la méthode d'enquête choisie : nous insisterons ci-dessous sur la temporalité des différentes approches auprès des personnes directement

concernées ou de leur hiérarchie, et sur la question de savoir comment organiser le retour, le contre-don, éventuellement la rémunération de ces personnes.

Typologie des relations et modes d'approche

On peut considérer que la façon dont les personnes sont approchées sur le terrain – la façon dont une relation personnelle et sociale est établie – est un acte ayant immédiatement des implications éthiques et juridiques. L'établissement de la relation avec les informateurs a d'une part des effets sur la qualité de leur collaboration et donc, en définitive, sur la qualité des données ainsi constituées ; d'autre part, elle a des effets sur les relations de confiance, d'acceptation, voire d'intérêt ou de curiosité scientifique que les informateurs nourriront envers les enquêteurs.

On peut esquisser une typologie des relations établies avec les informateurs en l'articulant au moment où ils sont approchés dans le processus de l'enquête :

- Quand l'enquête procède par *convocation nominale* des informateurs en laboratoire, les modalités de leur engagement sont généralement explicitées *préalablement*, au moment où les personnes acceptent de collaborer à l'enregistrement, effectué dans des lieux et à des moments convenus à l'avance. Les personnes sont alors soit sélectionnées et contactées par le chercheur (ou par une institution travaillant pour lui), soit elles répondent à un « appel à volontaires ». L'appel ou l'annonce de recrutement est le premier acte de communication qui manifeste (ou suscite des attentes quant à) la forme du contact, voire du contrat qui s'établit avec le chercheur.
- Quand l'enquête procède sous la forme d'un *fieldwork* (travail de terrain) impliquant une présence plus ou moins longue de l'enquêteur sur le terrain et des formes d'*observation participante* – classiquement discutées au sein des méthodes ethnographiques empruntées par les linguistes comme par d'autres chercheurs en SHS (Depperman 2000, Duranti, 1997, Hammersley, Atkinson 1995 & Moerman 1988) – la relation aux informateurs s'établit dans la *durée* de cette présence et est souvent associée à la construction de relations personnelles impliquant entre autres une confiance réciproque. Sur certains terrains, le chercheur n'est pas le premier à intervenir et d'autres l'ont peut-être précédé. Selon le comportement de ses prédécesseurs, l'accueil sera plus ou moins facile de la part de la communauté et, en particulier, les exigences en matière de retour (cf. 3.3.4) seront plus ou moins grandes.
- Quand l'enquête procède par *des entretiens, des « micro-trottoirs », des enregistrements d'activités* réalisés de manière *aléatoire* dans des espaces publics, sans viser des témoins particuliers mais des passants choisis simplement à cause de leur présence sur les lieux au moment de l'enregistrement, une rencontre préalable avec les informateurs est par définition impossible. C'est donc *juste avant, pendant* ou *juste après* la réalisation de l'enregistrement qu'ont lieu l'explication des finalités et la demande d'autorisation.

– Dans certains cas, il est possible d'envisager un contact *postérieur* à l'enregistrement : tel est le cas d'enregistrements réalisés à l'insu d'une partie des participants dont l'entrée sur la scène enregistrée n'était pas prévisible (c'est le cas des conversations téléphoniques par exemple, où une partie collabore à l'enquête et l'autre n'est pas toujours au courant de l'enregistrement ; elle est contactée ensuite pour donner son accord).

La forme du contact, de l'engagement, de la crédibilité, de la confiance varie énormément selon que la relation d'enquêteur à enquêté est établie au préalable ou durant le travail sur le terrain, de manière durable ou au moment même de l'enregistrement, ou encore après celui-ci.

Les personnes contactees

Dans la présentation que nous venons de faire, nous avons considéré, pour des raisons de simplicité, que le contact s'établissait avec la ou les personnes directement concernées par l'enregistrement ; or il s'agit souvent de personnes appartenant à un groupe ou à une institution – ce qui implique des prises de contacts multiples. Il s'agit ainsi de distinguer :

– Le cas où l'informateur agit *en son propre nom,* de manière individuelle.
– Le cas où l'informateur *est contacté* dans le cadre de ses activités professionnelles ou institutionnelles, et intervient donc en tant qu'appartenant à une organisation. La hiérarchie des personnes visées par l'enquête est aussi contactée au préalable : tel peut être le cas de la direction d'une entreprise, ou du chef d'une tribu, ou des parents d'élèves. Il convient de remarquer que la relation entre la personne et sa hiérarchie ne va souvent pas de soi et invite à différencier ce qui sera promis, expliqué, montré, etc. aux personnes et à leur hiérarchie.

Remunerations

Lors de l'approche des personnes concernées par l'enquête, des promesses peuvent être faites, de véritables contrats peuvent être proposés, des contreparties, rémunérations, remboursements peuvent être proposés. Ces engagements peuvent être à la fois éthiques et juridiques, sociaux, matériels voire financiers. De toute façon, la question se pose d'une forme de « dédommagement » des informateurs – qui est très différente si on la catégorise comme « contre-don », « rémunération », « dédommagement », « service rendu »…

Plusieurs cas de figure sont envisageables :

– *durant, voire avant l'enquête :* rémunérations financières promises dès l'établissement du premier contrat, contre-dons en nature, contre-dons symboliques, prestations pour la communauté concernée ;
– *après l'enquête :* reconnaissance de l'informateur sous des formes allant du remerciement ou de la citation à la mention comme co-auteur ou comme collaborateur, voire comme partenaire de la recherche, restitution des résultats, restitution des données/corpus sous forme d'archives, diffusion de savoir-faire, retours bénéfiques attendus pour la

communauté au sens large et sur le long terme (sur le modèle des bénéfices attendus d'une recherche médicale).

Pour une discussion de ces formes de « rendu » nous renvoyons (*infra*, 3.3.4) à la discussion du « retour » sur le terrain. La question reste de savoir ce qu'on peut/doit promettre aux informateurs lors de l'établissement de la relation, en tenant compte du fait que :

- Cette relation se modifie dans le *temps* (notamment si l'enquête de terrain implique une durée).
- Cette relation peut plus ou moins reconnaitre l'« informateur » comme un *partenaire* du projet de recherche (et non seulement comme un « objet »), dans des projets participatifs où le « natif » apporte plus que ses propres performances (par exemple en collaborant aux transcriptions, aux traductions, aux gloses des données).
- La *rémunération financière* peut être moins problématique pour des informateurs recrutés (parfois par des organismes spécialisés) dans le cadre d'un contrat formel ; elle peut être plus problématique sur le terrain, où elle implique une mise en concurrence non seulement entre les informateurs possibles, mais aussi entre les chercheurs pouvant y avoir accès (tel est le problème par exemple pour des linguistes d'universités moins dotées de moyens face à des chercheurs venant d'universités mieux dotées – et pouvant de ce fait être privilégiés par les informateurs ou générer chez eux des demandes difficiles à satisfaire). Les pratiques des anthropologues et des linguistes diffèrent sur ce point. Dans le cas d'une observation participante, il peut être délicat pour un anthropologue de rémunérer les personnes qui lui délivrent les informations, au risque d'entrainer une surenchère du cout de l'information. En revanche, la rémunération du locuteur et/ou traducteur qui passe plusieurs heures par jour avec le linguiste est un juste dédommagement pour un véritable travail, et n'entrave pas forcément la relation de confiance qui a pu s'instaurer entre les deux personnes.
- La rémunération financière n'est qu'un cas parmi d'autres de « *retour* » (ou de dédommagement, de salaire…), qui pour les enquêtes de terrain se fait toujours de manière plus ou moins implicite, au fil de la vie quotidienne et de la négociation des relations mutuelles.

3.3.2 DISPOSITIF D'ENREGISTREMENT

Le choix du dispositif d'enregistrement des corpus a des effets sur la manière dont les personnes concernées vont être traitées, dont leur consentement va être obtenu, dont l'acceptation ou l'acceptabilité de l'enregistrement vont se négocier.

Nous allons ici discuter quelques aspects qui peuvent se révéler pertinents, allant du choix des contextes dans lesquels effectuer l'enregistrement aux modalités de l'enregistrement.

Contexte de l'enregistrement

Par définition, il n'est pas possible de *tout* enregistrer et les chercheurs sont obligés de faire des choix. Ceux-ci dépendent de l'objet de recherche visé, des contraintes techniques (par exemple, difficulté à enregistrer en vidéo la nuit ou en audio dans des lieux très bruyants), et aussi du respect des personnes enregistrées.

Interviennent notamment :

- le choix du *moment* à enregistrer : il s'agit de trouver un équilibre entre les moments intéressants pour l'enquêteur et le respect de la vie privée de l'enquêté ;
- le choix des *activités* à enregistrer : celles-ci peuvent être davantage publiques et sociales ou bien intimes et privées ;
- le choix du *lieu* où enregistrer : là aussi il y a une tension entre des lieux publics détachés de la vie privée ordinaire des personnes et des lieux intimes : le *laboratoire* est un lieu totalement détaché de l'espace de vie des informateurs – et c'est d'ailleurs ce qui fait que les chercheurs voulant travailler sur les pratiques sociales situées l'évitent ; le *domicile* des personnes est leur lieu de vie, lui-même articulé en lieux plus « publics » ou plus « intimes » (un repas pris à la salle à manger, à la cuisine ou au lit n'a pas la même teneur, ainsi qu'un entretien effectué au salon ou autour de la table de la cuisine) ; les *espaces de travail* sont eux aussi, quoique de manière différente, structurés par des questions de confidentialité qu'il s'agit de respecter ; leur non-respect peut risquer d'impliquer pour les données recueillies un devoir de confidentialité qui signifie l'impossibilité de leur exploitation (voir 3.4) ; les *espaces religieux*, sacrés et/ou soumis à des tabous doivent également être respectés. De manière générale, une bonne connaissance du lieu et de son organisation géographique et sociale est nécessaire avant d'envisager tout enregistrement (image ou son).

L'équilibre à trouver se situe donc entre contextualité et naturalité des données enregistrées et voyeurisme – le choix des moments à enregistrer pouvant avoir des conséquences importantes sur la suite de l'enquête (sur les autorisations pour exploiter les données et sur le droit de rétractation *post hoc* des sujets).

Modalites d'enregistrement

Les modalités d'enregistrement interviennent souvent dans le choix des contextes à enregistrer (cf. *supra*), des activités visées ainsi que dans les modalités d'acceptation ou de résistance des personnes concernées. Différentes dimensions techniques peuvent intervenir concernant l'acceptabilité de l'enregistrement par les personnes enregistrées :

- Le fait que l'enregistrement soit réalisé en *audio* ou en *vidéo* : pour certaines activités, les personnes concernées peuvent préférer l'audio à la vidéo – jugée plus invasive -, quittes à passer de l'audio à la vidéo dans un deuxième temps, une fois constatés les modalités et les effets de l'enregistrement sur l'activité.

– Le fait que l'enregistrement soit réalisé par *l'enquêteur présent*, par des *techniciens* ou par un *dispositif pré-installé* et fonctionnant en l'absence du chercheur a des effets sur son acceptation : même si la caméra ou le micro sont souvent traités comme des « prothèses » ou des prolongements du chercheur (par exemple, quand les participants s'adressent directement à eux), l'absence du chercheur peut être préférée par certains participants.
– Le fait que l'enregistrement soit réalisé par le *chercheur* ou par *les participants eux-mêmes* : d'une part, la délégation de l'enregistrement aux participants peut être vue comme une forme de contrôle de leur part sur ce qui est enregistré ; d'autre part, cette délégation peut être refusée comme une forme trop poussée de collaboration détournant le participant de son activité.
– Le fait que l'enregistrement soit réalisé par un *dispositif voyant ou discret*, voire caché : il existe de nombreux débats sur le fait de recourir à un micro caché et sur les conséquences de ce choix sur les relations possibles avec les participants (Mitchell, 1991, Mondada, à paraitre, Welland & Pugsley, 2002) ; par ailleurs, même lorsque les participants sont au courant de l'enregistrement, le fait de recourir à un dispositif voyant peut aussi bien être perçu comme un gage de transparence que comme une gêne. Souvent, la miniaturisation des dispositifs permet de les installer d'une manière qui, sans du tout les dissimuler, en fait rapidement des éléments intégrés dans le décor.
– Le fait que l'enregistrement dépende de *moyens techniques nécessitant une intervention à brève échéance* (relative par exemple à la durée de la batterie ou à la durée de la cassette) implique des perturbations de l'activité par le chercheur (ou par les participants qui effectueraient le remplacement de la cassette) qu'évitent d'autres dispositifs dotés d'une plus grande autonomie (enregistrant par exemple directement sur des disques durs). Cela peut avoir des répercutions sur le comportement des témoins en raison du dérangement occasionné, en particulier pour certaines activités (comme opérer un patient, effectuer une consultation en thérapie, discuter d'un contrat délicat, être engagé dans un processus de création).
– Le fait que l'enregistrement offre ou non des *angles morts* aux participants qui voudraient lui échapper un instant : par exemple, le cadre et le champ délimités par une seule caméra permettent d'inférer des zones qu'elle ne couvre pas alors que la puissance imaginée d'un micro ou le fait de recourir à plusieurs caméras sur la même scène peuvent donner l'impression d'un dispositif de surveillance auquel on ne peut se soustraire.
– Le fait de pouvoir arrêter ou imposer des *coupures à l'enregistrement* peut intervenir comme une matérialisation de la possibilité de rétractation ; le fait que l'effacement ou la coupure de l'enregistrement puissent être effectués par les participants quand ils le désirent ou bien doivent être

effectués plus tard, ou par des tiers, peut donner l'impression d'une plus ou moins grande latitude pour intervenir sur les données et suppose des relations de confiance différentes. Cette question – comme bien d'autres – est, là aussi, liée aux contraintes techniques de l'enregistrement et à la sophistication du dispositif. On pourra en tenir compte dans le choix de supports permettant ou non un effacement immédiat des données ou bien permettant ou non un visionnement sur place de ce qui a été enregistré.

Ces considérations (Mondada 2006) montrent bien l'imbrication des questions techniques et des questions juridiques, le respect à la fois personnel, éthique et juridique des participants étant matérialisé dans les choix techniques mis en œuvre.

3.3.3 DEMANDE D'AUTORISATION ET CONSENTEMENT ECLAIRE

La définition du « consentement éclairé » et sa traduction dans des formes de relation sociale (le contact avec les informateurs) et des formes matérialisées (les documents échangés et signés) sont sensibles au contexte et aux objets de l'enquête, ainsi qu'aux conditions socio-culturelles du groupe dans lequel cette enquête se déroule. Nous esquissons ici quelques pistes de réflexion, en partant de la définition même du « consentement éclairé », en reprenant la question du moment auquel ces questions se posent, ainsi que la question des personnes que l'on informe et à qui on demande l'autorisation, des formes que prend cette information, des objets à propos desquels on choisit d'informer, et des formes du consentement lui-même.

DEFINITION DU « CONSENTEMENT ECLAIRE »

On parle souvent de formulaires d'autorisation à soumettre aux informateurs ; il est cependant important de faire dépendre cette autorisation de l'information préalable donnée aux personnes concernées : sans *information*, la *demande d'autorisation* n'a pas d'objet ni de sens. C'est pourquoi on parle de *consentement éclairé (informed consent*), dans le sens où l'acceptation de l'enregistrement est étroitement dépendante de la compréhension des finalités pour lesquelles il est effectué. Sur certains terrains, la difficulté de faire comprendre les finalités de la recherche ne doit cependant pas inciter le chercheur à passer outre la demande de consentement, et celle-ci doit alors être formulée en accord avec le type de société dans laquelle se déroule le terrain (par exemple, comment concevoir un consentement individuel signé dans une société à tradition orale dans laquelle le droit privé n'a aucun sens ?).

MOMENT DE L'INFORMATION ET DE LA DEMANDE

La demande d'autorisation dépend du mode d'approche des personnes enregistrées. Elle peut différer selon le moment où elle a lieu :

- information et demande *préparée à l'avance* durant une permanence sur le terrain et dépendant de la relation d'interconnaissance et de confiance avec l'enquêteur,
- information et demande faite *juste avant* l'enregistrement,
- information et demande faite *juste après* l'enregistrement,
- information et demande orale effectuée *avant* et demande écrite effectuée *après* l'enregistrement (avec possibilité de rétractation).

L'information est plus abondante lorsqu'elle bénéficie de la présence prolongée de l'enquêteur sur le terrain ; elle est plus restreinte lorsque la demande d'autorisation se fait rapidement avant ou après l'enregistrement, sans autre forme de contact entre les enquêteurs et les enquêtés.

Le moment où se situent l'information et la demande d'autorisation peut être choisi en relation avec ses effets envisagés sur la structuration de l'activité enregistrée : souvent le moment de l'information et de la demande d'autorisation est choisi de manière à ne pas perturber l'activité du point de vue des participants (par ex. une demande d'autorisation à un client au moment de la vente peut provoquer un risque de perturbation de la vente pour le vendeur et donc être refusée à l'enquêteur qui désirerait documenter cette activité), ou du point de vue des enquêteurs (par ex. une demande d'autorisation en ouverture de conversation modifie l'organisation du déroulement séquentiel de cette ouverture).

Si l'information et la demande interviennent *après* l'enregistrement, l'information peut apparaitre comme un « dévoilement », une « révélation » qui *a posteriori* qualifie l'enregistrement de « dissimulation » : cela peut faire intervenir des recatégorisations des participants et des activités (celui qui s'était présenté comme un touriste perdu dans la ville demandant son chemin devient un enquêteur travaillant sur les descriptions spatiales dans les demandes d'itinéraire) (Mondada à paraitre). En outre, cette technique n'est pas envisageable pour de nombreux terrains de recherche. Ainsi ces cas de dissimulation sont particulièrement mal venus dans certaines communautés et font alors beaucoup de tort à la communauté scientifique dans son ensemble et aux chercheurs qui suivront.

Statut du demandeur

Même si le chercheur est celui qui informe et demande habituellement l'autorisation d'enregistrer, différents cas de figure sont envisageables :

- Le cas le plus classique est celui de *l'enquêteur* se chargeant de l'information et de la demande d'autorisation.
- Souvent toutefois le chercheur envoie sur le terrain des *étudiants* ou *des collaborateurs* qui sont autant de porte-paroles du projet.
- Dans certains cas, il est envisageable que les participants deviennent eux-mêmes *les porte-paroles du projet* : cela est le cas lorsque le chercheur demande à un participant d'informer d'autres participants (par ex. l'hôte qui invite chez lui des amis à un repas qui sera enregistré ; le commerçant qui demande à ses clients d'accepter de se laisser enregistrer ; l'enseignant qui demande l'autorisation à ses élèves ou étudiants, etc.). Cette délégation fait partie des collaborations sur le terrain entre enquêtés et enquêteurs ; elle peut toutefois être la source de malentendus et de difficultés.

De même, l'autorisation peut concerner les signataires eux-mêmes ou des personnes qui dépendent d'eux (subalternes, enfants, étudiants, etc.). Dans ce dernier cas, il est important de tenir compte du fait que *autorisation* ne se confond pas toujours avec *acceptation*. Dans les sociétés où le droit individuel n'existe pas, l'avis et l'autorisation

du groupe dans son ensemble ou de certains de ses responsables (politiques ou religieux) sont souvent indispensables.

QU'EST-CE QU'*INFORMER* ?

Au cœur du consentement éclairé, il y a l'exigence d'informer les participants enregistrés. Toutefois, dès que l'on interroge cette exigence, les questions surgissent. Qu'est-ce qu'« informer » ? Informer « à propos de quoi » ? A quelles conditions peut-on dire que cette information produit le statut « éclairé » de son destinataire ?

La notion même d'« information » peut laisser penser à un simple transfert de messages et de contenus ; elle tend à gommer les processus, les contextes et les contingences qui caractérisent cette activité communicationnelle par laquelle un enquêteur explique l'objet de son enquête à ses partenaires sur le terrain. Dès que l'on réfléchit en termes de type d'activité, l'« information » aux enquêtés pose une série de problèmes à résoudre :

- *l'adéquation au destinataire* : l'explication du projet de recherche, pour être comprise et partagée, demande à être ajustée aux compétences, au niveau de langue et de compréhension du destinataire, cet ajustement concerne aussi le contexte et les modalités de l'enquête, prenant en compte l'adéquation entre ce que les partenaires voient faire sur le terrain et les explications qu'on en donne ;
- l'explicitation des *finalités de l'enquête* doit se faire sans *nuire* à celle-ci : cela pose la question de l'équilibre à trouver entre la transparence de l'enquête et les transformations qu'elle peut induire sur les conduites des participants ;
- l'explication du projet de recherche peut se faire à des *niveaux de généralité différents* (de « c'est une enquête sur les façons de parler des gens » à « c'est une enquête sur la fréquence et les contextes de la liaison non obligatoire en français »).

L'information aux enquêtés comprend non seulement des explications du projet scientifique mais aussi des informations précises concernant par exemple :

- les *responsables* de l'enquête et leur affiliation institutionnelle, ainsi que les financeurs,
- une *adresse* de contact,
- les *personnes qui auront accès aux données* et qui travailleront sur elles,
- *la façon* dont les enquêtés ont été choisis et la population dont ils font partie,
- la façon dont les données seront *anonymisées*,
- le fait que les données seront transcrites selon des *conventions particulières* (possibilité de donner un exemple),
- la façon dont les données seront *archivées* une fois l'enquête terminée (conservation ou destruction à la fin de l'enquête, conservation auprès de quel garant, modalités de réutilisation éventuelle, transmission à d'autres chercheurs),
- les *modalités d'accès* aux informations relatives au projet et concernant tout particulièrement les données/analyses faisant référence à la

personne (possibilité d'accès aux fichiers et informations concernant tout particulièrement la personne),
- les droits de la personne, notamment le droit de *rétractation*,
- les *risques* éventuels ainsi que les retombées positives, morales ou matérielles, de l'étude.

Les modalités d'information peuvent, elles aussi, varier selon la culture des destinataires, en particulier :

- l'information peut se faire de manière orale : individuellement dans des *conversations familières*, collectivement dans des *réunions d'information*…
- elle peut se faire de manière écrite (par une brochure, un dépliant…) ou par courriel.

Dans le contexte d'une culture écrite, il est recommandé de laisser un texte ; de même, l'indication d'un site Internet où suivre l'évolution du projet (éventuellement avec des modes d'accès particuliers) peut être utile.

L'OBJET DE LA DEMANDE D'AUTORISATION

Ce n'est qu'après cette phase d'information que la demande d'autorisation de collecter les données peut intervenir. La question qui se pose est de savoir comment circonscrire l'objet de cette autorisation.

L'autorisation concerne en effet les dimensions suivantes, qui peuvent interagir et se superposer :

- les *actions* effectuées par les chercheurs dans le cadre du projet : l'enregistrement, la préparation du corpus (transcription, traduction, annotation, etc.), les conditions d'archivage (lieu de dépôt, durée prévue de la conservation, institutions garantes...), l'analyse dans le cadre des objectifs annoncés, les usages des données de manière intégrale ou non, la diffusion des résultats de l'analyse, la conservation/destruction des données une fois terminée l'enquête ;
- les *formats* et les conditions de l'enregistrement : audio/vidéo, avec plusieurs caméras/micros, à des moments connus ou non des enquêtés, bien circonscrits ou couvrant de longues durées, tout choix technique intervenant dans la façon dont la personne figurera dans les données peut être explicité voire négocié ;
- les conditions *de diffusion* des données et des résultats : sous forme intégrale ou partielle (courts extraits dont la longueur maximale peut être prévue), sous forme uniquement textuelle (transcriptions) ou audio-visuelle (dans des documents Powerpoint par exemple) ;
- les contextes de diffusion des données et des résultats : des contextes de recherche (*workshops* [ateliers], colloques, congrès), des contextes d'enseignement universitaire, des contextes de formation et de vulgarisation plus larges, des contextes liés au terrain (par exemple il faut demander explicitement l'autorisation de réutiliser les données dans le contexte d'une formation dans la même institution où elles ont été recueillies – où elles peuvent se révéler très sensibles) ;

– des contextes larges de diffusion : sous la forme d'un cédé ou sur un site internet.

L'explicitation de ces contextes se superpose avec celle des activités dans lesquelles les données seront utilisées ; l'enjeu dans les deux cas est celui des personnes qui auront accès aux données dans le cadre de ces activités. On peut différencier les contextes de diffusion soumis à un certain contrôle de la part du chercheur (par des conventions, par exemple) et les contextes de diffusion incontrôlables par définition (site Internet par exemple).

Il est envisageable de laisser la possibilité à l'enquêté d'ajouter des contraintes qui lui seraient personnelles ; toutefois cette éventualité pose le double problème de sa légalité ainsi que celui de son interprétabilité. Un des problèmes majeurs qui se posent dans la demande d'autorisation – comme d'ailleurs pour l'information – concerne l'évolution toujours possible des finalités de l'enquête, qui peuvent ne pas être totalement fixées à son début et surtout se transformer au fil du travail sur le terrain et sur les corpus. Pour cela, il est important de formuler les finalités de manière suffisamment générale pour intégrer d'éventuelles évolutions des finalités pouvant émerger au cours du travail de recherche. Par contre, tout changement de finalité devra faire l'objet d'une nouvelle demande (cf. *infra*).

Les formes de l'autorisation

La demande d'autorisation peut prendre différentes formes, qui dépendent elles aussi du contexte socio-culturel dans lequel se déroule l'enquête : ainsi par exemple exiger la signature de l'enquêté n'a de sens que dans les cultures de l'écrit, de la *littéracie* où cette procédure a un sens, n'effraie pas et n'est pas liée à d'autres pratiques avec lesquelles elle pourrait être confondue (comme la signature de chèques).

On peut donc différencier les formes de la demande selon le support sur lequel elles sont consignées :

– demande écrite et signée,
– demande orale,
– il est possible et utile de prévoir que l'autorisation orale soit elle-même enregistrée, sous forme audio ou vidéo, ce qui permet d'en assurer la traçabilité ; c'est la solution à favoriser lors du travail dans des sociétés à tradition orale, en respectant, en fonction des besoins, le degré de formalité requis par les pratiques langagières de la communauté concernée et le choix de la langue (par ex. enregistrement individuel avec le locuteur pour une autorisation ponctuelle, ou autorisation enregistrée lors d'une réunion plus formelle avec les autorités).

Dans le cas de la demande écrite, celle-ci peut se présenter sous différentes formes – dans un texte préformé (formulaire) :

– Un texte *compact* qui synthétise les différents aspects de la demande d'autorisation et qui demande un accord (ou un refus) global.
– Un texte présentant des cases à cocher et donc des *choix* : cette forme a l'avantage sur la première de matérialiser des choix véritables pour l'enquêté et donc de lui laisser la possibilité de refus partiels (par ex. il

peut accepter l'enregistrement audio mais refuser l'enregistrement vidéo) voire d'ajouts de contraintes (par ex. il peut demander l'anonymisation de la vidéo en plus de l'audio). La question qui se pose alors est celle de la formulation des alternatives, de manière à ce qu'elles ne soient pas redondantes et qu'elles ne soient ni trop compliquées ni trop longues à traiter pour l'enquêté.

Un problème peut se poser lors des demandes collectives, lorsque des groupes sont concernés (par exemple dans le cadre d'enregistrements de réunions) : si de trop nombreuses alternatives sont laissées au choix des participants, il est possible que les réponses mènent à des résultats contradictoires où n'émerge aucun dénominateur commun ; dans ce sens, les demandes à des groupes présentent des problèmes et des contraintes qui ne sont pas les mêmes que pour les individus.

Pour aller plus loin :
voir Exemples de demande d'autorisation en annexe.

3.3.4 APRES L'ENQUETE : RETOURS, DEBRIEFINGS

On insiste souvent sur la préparation du terrain, mais il est également important de prévoir le départ et le retour sur le terrain. Cela présente une importance à la fois scientifique et éthico-juridique : le retour sur le terrain peut se révéler nécessaire à tout moment pour une vérification, un complément d'enquête, une reprise de contact avec les informateurs. Si le départ du terrain s'est mal passé, le retour sera impossible. Par ailleurs, la présence sur le terrain produit non seulement des relations de confiance, mais aussi des attentes qui engagent dans la durée : quitter le terrain en disparaissant tout simplement, après avoir pratiqué une immersion qui souvent établit des relations étroites avec les participants et leur demande de l'aide et des prestations, peut produire de grosses déceptions. Une fois « prélevé » du savoir, des réponses, des corpus sur le terrain, il s'agit donc de savoir comment « rendre » quelque chose aux personnes sans lesquelles l'enquête aurait été impossible (voir aussi les questions de rémunération traitées *supra*). Il est par exemple désormais impossible de travailler sur certains terrains (dans le cas des langues en danger) sans envisager une restitution au locuteur et à la communauté, voire un engagement du chercheur, sous quelque forme que ce soit (implication dans des projets éducationnels, de littéracie, etc.)[8].

Il convient en outre de signaler que les « feedbacks », les « debriefings », les retours d'expérience peuvent se faire déjà pendant le travail sur le terrain, sous la forme de comptesrendus de résultats partiels par exemple. La distinction entre le « pendant » et l'« après » du terrain peut ainsi être relativisée.

---

[8] Rapport de l'UNESCO (2001), *Language vitality and Endangerment* : « Any research in endangered language communities must be reciprocal and collaborative. Reciprocity here entails researchers not only offering their services as a quid pro quo for what they receive from the speech community, but being more actively involved with the community in designing, implementing, and evaluating their research projects ».

Plusieurs types de pratiques sont envisageables pour assurer un « retour » auprès des populations enquêtées. Nous en énumérons quelques-unes, allant de la présentation de résultats la plus proche du contexte académique à la formulation de savoirs et savoir-faires la plus proche du terrain. C'est sans doute dans l'évaluation de la distance entre le « retour » et l'académie ou le terrain que se situent les choix de « politique du terrain » :

– Présentation des résultats à la fin du projet : la formulation des résultats peut être plus ou moins vulgarisée, plus ou moins proche des préoccupations des enquêtés, la présentation des résultats peut comporter notamment des exemples de *transcription* et d'analyse de transcriptions : les participants réagissent de manière très différente (parfois surpris, parfois choqués) à la représentation de leur voix.
– Démarche *d'empowerment* (restitution) : elle consiste à ne pas simplement penser le « retour » en termes d'« information » mais aussi en termes d'apport en savoirs et savoir-faires à la communauté des enquêtés (Cameron *et al.*, 1991) : on peut ainsi songer non seulement à présenter des analyses mais à permettre aux participants de continuer à *collecter* des données et d'analyser leurs propres données pour leurs propres fins, on peut formuler les *retombées de l'analyse* dans les termes de l'agenda, des thèmes, des préoccupations des acteurs, on peut répondre, dans la mesure des compétences du chercheur, aux *demandes d'expertises* souvent exprimées par les communautés (par ex., ateliers de réflexion sur le passage à l'écrit, ou sur la traduction de documents officiels, implication dans des programmes d'éducation bilingue), on peut mettre au service de la communauté les *savoirs produits* par l'enquête en les matérialisant dans d'autres formes que les écrits universitaires traditionnels (par ex.. sous forme d'expositions, ou d'autres produits culturels dérivés), on peut offrir une *formation* basée sur les résultats/les méthodes de l'enquête ; de manière plus générale, on peut songer à transmettre des outils d'analyse, à transférer des compétences qui pourraient être utiles sur le terrain.
– La question du « retour » des données elles-mêmes sous forme de corpus ou d'archives peut se révéler délicate : elle peut s'imposer dans certains cas (ainsi, pour les langues en danger, il devrait[9] être constitué des archives patrimoniales léguées à la communauté) mais aussi devoir être évitée pour protéger les informateurs (ainsi dans le cas d'enquêtes dans des entreprises ou des institutions, les données collectées pourraient intéresser certains niveaux de la hiérarchie mais nuire à des subalternes). Le retour des archives, s'il est pertinent, pose donc souvent des questions : d'accès limité des personnes pouvant consulter ces archives, en tenant compte des risques et des avantages que produit la mise à disposition sur le terrain, de modes et de technologies d'accès

[9] C'est même un devoir selon les recommandations de l'UNESCO en la matière (voir fiche *Unesco*).

aux archives : si les archives sont formatées pour que la population concernée puisse y avoir accès, les technologies doivent être adaptées aux usages et aux possibilités de ces populations (il ne sert à rien de faire un DVD si personne n'a de lecteur de DVD, ou de faire un site Internet si personne n'a d'accès à l'informatique, se pose ici la question de la gestion de l'asymétrie entre « l'académie » et le « terrain »), la garantie *d'accès aux publications* pose des questions analogues à celle de l'accès aux données, quoique de manière souvent moins difficile.

3.4 ANONYMISATION

La possibilité ou la garantie (que nous relativiserons plus bas) de rendre les données recueillies anonymes est importante pour la protection de la vie privée des personnes concernées par l'enquête et pour la légalité des corpus recueillis par les chercheurs. L'anonymisation des données n'est toutefois ni un processus simple ni une garantie non-problématique, car elle fait surgir de nombreux problèmes à la fois techniques, scientifiques et sociologiques.

L'anonymisation des données est une garantie importante en matière de légalité des données et de leur usage ; dans certains cas, si elle garantit véritablement la non-identification des personnes concernées, et si par ailleurs les données ne sont pas protégées par le droit d'auteur, elle peut permettre d'utiliser des données même en l'absence de demande d'autorisation préalable. Il convient toutefois d'être prudent sur ce point – en considérant toutes les limitations et les difficultés auxquelles on se heurte dans l'anonymisation (cf. *infra*).

3.4.1 DEFINITION

Bien qu'on parle souvent d'anonymisation, la question légale qui se pose est celle de *l'impossibilité d'identifier des personnes* : l'enjeu est que, sur la base des données recueillies et de leurs modes de représentation (transcription par exemple), on ne puisse pas identifier les personnes concernées. Les procédures d'identification sont bouleversées par les technologies actuelles qui offrent des facilités de stockage et de diffusion des données, mais aussi de puissants outils de traitement des informations (tri, recoupement, requêtes croisées…).

Ces considérations concernent :

- tout ce qui permet d'identifier *directement* une personne : par référence au locuteur ou à un tiers et à sa sphère privée, sur la base des manifestations du locuteur, comme sa voix ou son apparence physique ;
- tout ce qui peut lui porter préjudice ;
- tout ce qui peut indirectement permettre, par recoupement d'informations, de remonter au locuteur concerné.

Les opérations qui suppriment ces références ou ces manifestations sont appelées des procédures d'« anonymisation » des données.

3.4.2 *DONNEES CONCERNEES*

L'anonymisation ne concerne pas uniquement les enregistrements ou les transcriptions, mais un ensemble de données qui sont contenues dans les corpus et qui se différencient selon divers supports et formats – dont dépendront les techniques d'anonymisation :

- les données primaires vidéo,
- les données primaires audio,
- les données primaires textuelles : documents, officiels ou non recueillis sur le terrain,
- les données secondaires : transcription, notes de terrain, métadonnées, analyses, descriptions ethnographiques,
- les données secondaires visuelles : copies d'écran (*screen shots*), voire représentations de la voix (oscillogrammes, spectrogrammes…).

On remarquera que certaines données personnelles échappent à l'anonymisation : tel est le cas des hommes et des femmes publics, dans des interventions à caractère public (par exemple des hommes politiques à la télévision), où ils interviennent en connaissance de cause en ce qui concerne la diffusion de leur image et où leurs propos sont eux-mêmes considérés comme un discours public. Dans ce cas, les propos, s'ils sont considérés comme « originaux », seront soumis aux contraintes de diffusion régissant le droit d'auteur avec une tolérance pour un laps de temps déterminé par « l'actualité ». Dès lors que ces interventions ne sont plus considérées comme liées à l'actualité, elles échappent à cette qualification[10].

3.4.3 *QUAND ANONYMISER ?*

On peut distinguer différents moments auxquels peut intervenir l'anonymisation. Selon les finalités de l'étude et les contextes de l'enquête, on peut considérer que l'anonymisation doit se faire le plus *tôt* ou le plus *tard* possible. La première solution augmente les garanties de confidentialité pour la personne, la seconde maximise les possibilités d'analyse pour le chercheur. Les temporalités peuvent varier selon les types de données aussi :

- on évite l'anonymisation sur les données primaires originales de référence car elle pourrait endommager les données elles-mêmes ; par contre les données ainsi non anonymisées doivent être conservées dans un lieu sûr,
- les données peuvent/doivent/ne doivent pas (selon les politiques adoptées) être anonymisées lors de leur dépôt pour conservation ; le rôle de garant des institutions assurant la conservation est ici concerné,
- on peut travailler (dans un groupe de recherche bien délimité et qui garantit la non circulation externe des données) sur des données non anonymisées et garantir en revanche une anonymisation de tout extrait figurant dans un écrit ou une présentation orale,

[10] Cf. l'art. 122.5 du code de la propriété intellectuelle.

– on effectue toujours l'anonymisation sur les copies destinées à circuler entre chercheurs extérieurs au projet et parfois entre chercheurs internes au projet (c'est le cas notamment pour de grands consortiums de recherche ou des projets articulant des réseaux d'équipes importants).

3.4.4 COMMENT ANONYMISER ?

Les modes d'anonymisation touchent à la fois les supports et les formats des données et mettent ainsi en jeu des possibilités et des contraintes technologiques ; ils concernent aussi des formes et des manifestations symboliques de l'identité des personnes et mettent ainsi en jeu des questions d'analyse.

FORMES OU ELEMENTS CONCERNES PAR L'ANONYMISATION

Comme nous allons le voir, il est difficile – voire impossible – de constituer une liste finie des formes concernées par l'anonymisation. On peut toutefois souligner les formes principales :

– formes nominatives (nom, prénom, surnom ou petit nom, sigle d'entreprise…),
– données personnelles (adresse, numéro de téléphone, numéro de passeport, numéro de compte, âge, lieu de naissance…),
– profession, statut, titres,
– activités sociales,
– parenté, réseaux,
– référence à des lieux (toponymes, institutions, services…),
– référence à des caractéristiques de la personne (physiques, culturelles, médicales…) uniques ou rares dans son milieu identifié,
– caractéristiques physiques : voix, visage, caractéristiques corporelles,
– etc.

L'« etc. » clôturant cette liste souligne le fait que tout élément, selon les contextes d'enregistrement et de réception de cet enregistrement, peut devenir un porteur d'informations sur l'identité des personnes. L'identification des formes concernées par l'anonymisation suppose donc une compétence sociologique et culturelle qui rende le chercheur capable d'imaginer les usages, les connaissances et les associations qui pourraient permettre l'identification d'une personne sur la base d'une forme donnée.

FORMES DE REMPLACEMENT

Une fois identifiées les formes pouvant porter à l'identification des personnes, il s'agit de les transformer pour effectuer les opérations d'anonymisation.

On fera remarquer que la forme la plus radicale d'anonymisation est la *suppression* pure et simple des données – bien que l'on cherche souvent d'autres moyens d'assurer l'anonymisation qui puissent mieux les préserver. On notera cependant que la suppression peut être partielle (on peut envisager de détruire des extraits qui seraient porteurs de trop d'éléments problématiques et confidentiels pour qu'ils soient utilisables en l'état).

La forme d'anonymisation généralement adoptée procède par *remplacement* d'éléments confidentiels par des formes neutres. Ces formes varient selon les supports techniques concernés : nous distinguerons ici entre le texte, l'audio et la vidéo.

- *Texte*

Les textes concernés sont d'abord la transcription et toutes ses mentions dans des articles, exempliers, cours, conférences… D'autres textes devant être anonymisés sont les données primaires textuelles (documents recueillis sur le terrain). Celles-ci peuvent se présenter d'ailleurs sous une forme textuelle ou sous la forme d'image (tel est le cas d'une lettre, d'un document administratif, d'un manuscrit qui est conservé sous forme photocopiée ou numérisée).

Le principe de la substitution consiste à rendre visible la portion de texte qui a été remplacée, et ainsi à donner des informations générales sur elle (concernant au moins sa durée).

– *Remplacement par un « blanc »* : c'est la solution la moins informative et surtout la moins visible.
– *Remplacement par un hyperonyme* ou une abréviation, tel que NN ou NVILLE ou NHOPITAL pour nom, nom de ville, nom d'hôpital, etc. Cette solution peut rester informative (on précise le type de référence de la forme anonymisée). Elle est utile dans les cas où la substitution par pseudonyme (cf. *infra* ici-même) est impossible, difficile ou non vraisemblable. Cette solution implique le développement de conventions spécifiques pour la notation de ces hyperonymes, qui ne sont pas de même nature que le texte qu'ils remplacent (c'est pourquoi l'emploi des majuscules est parfois proposé, quand il n'entre pas en contradiction avec d'autres emplois de majuscules prévus dans les conventions de transcription).
– *Remplacement par un pseudonyme* : c'est la solution la plus souvent utilisée, du moins pour les noms de personnes car elle permet une bonne intégration de la forme de remplacement dans le fil du discours, n'attire pas l'attention sur elle, est vraisemblable et garde un certain nombre d'indications contenues dans la forme initiale. Cela n'est toutefois possible que si le choix des pseudonymes est réfléchi et répond aux problèmes suivants : le pseudonyme est choisi dans le même champ paradigmatique que la forme qu'il remplace (par exemple « Ahmed » sera remplacé par « Moustapha » plutôt que par « Albert », le pseudonyme tentant de conserver des traits d'ethnicité), dans certains cas, notamment si l'interaction enregistrée le rend pertinent, on veillera à conserver les connotations possibles du nom (par ex. s'il est à la base de plaisanteries ou de jeux de mots) et le nombre de syllabes et certaines caractéristiques phonétiques et prosodiques (si elles sont exploitées dans l'interaction) ; le pseudonyme est choisi de manière à éviter de pouvoir reconstituer le nom initial (dans ce sens, le choix d'un pseudonyme commençant par les mêmes lettres que l'original est à éviter, même s'il présente des avantages pour sa mémorisation) ; le pseudonyme est choisi de manière à éviter de ridiculiser la personne (dans ce sens, sont à

éviter les pseudonymes qui renverraient à des caractéristiques de la personne – par ex. « Monsieur Gros ») ; les noms des rues, les numéros de téléphone, etc. peuvent être remplacés de la même manière que les noms de personne.

On remarquera qu'il est plus facile de choisir un pseudonyme pour les personnes que pour les noms de villes (on peut imaginer un nom de petite ville ou de quartier ou encore de rue mais beaucoup moins un nom de grande ville ou de capitale) ; il est parfois envisageable mais pas toujours possible de penser à des pseudonymes pour des noms de services institutionnels (cela n'a pas de sens de remplacer « département de chirurgie » par « département de dermatologie » dans le cas d'un hôpital). Dans le cas où le choix d'un pseudonyme est difficile ou invraisemblable, on recourra à la solution de l'hyperonyme.

- *Audio*

– *remplacement par du silence* ; cette solution a comme désavantage le fait que le remplacement peut être confondu avec une pause,
– *remplacement par un bip ou un autre bruit* qui ne se confond avec aucun signal pouvant intervenir dans l'enregistrement,
– *remplacement par le signal original filtré et déformé* ; cette technique est surtout utilisée dans les médias pour rendre la voix non identifiable ; quand elle est pratiquée par des non spécialistes, elle peut poser des problèmes quant à son irréversibilité (possibilité de rétablir le signal original).

- *Image*

L'image concernée est surtout celle, dynamique, des enregistrements vidéo. Mais on peut penser aussi aux images fixes, par exemple à des photographies sur des documents et à des captures d'écran dans les transcriptions. De même, on peut songer à l'anonymisation d'une représentation visuelle du flux sonore (dans un spectrogramme par exemple) lorsqu'elle pourrait rendre reconnaissable la prononciation d'un nom ou d'un numéro.

– pour ces données, la *suppression* est envisageable sous forme de coupures lors du montage ; dans ce cas, il est conseillé de marquer la durée du segment coupé sur la bande et de ne pas donner l'impression d'une continuité ;
– *remplacement par un brouillage du signal* : par floutage, par pixélisation ou par contourage de l'image ou par application d'autres types de filtres (ce traitement peut concerner *toute l'image ou un détail uniquement*) ; dans ce dernier cas, elle est d'une technique plus complexe à réaliser quand ce détail est en mouvement ;
– placement d'un bandeau noir sur les yeux de la personne.

3.4.5 LES LIMITES DE L'ANONYMISATION

Même si l'anonymisation est une opération fondamentale pour assurer la circulation légale des données, il convient d'être prudent par rapport aux promesses et garanties faites aux enquêtés concernant l'anonymisation des données.

Les limitations sont essentiellement de deux ordres très différents, le premier concernant les contextes qui augmentent ou diminuent la reconnaissabilité des personnes, le second concernant les contraintes que l'anonymisation fait peser sur les objets mêmes de la recherche.

Contextes de production et de circulation

L'anonymisation est relativisée par différents facteurs intervenant soit lors de la production des données – et selon les spécificités de ce qui se passe durant l'enregistrement –, soit lors de la réception de ces données :

- L'anonymisation opère d'abord sur une série de formes censées contenir les indications principales permettant l'identification de la personne ; néanmoins n'importe quelle référence ou forme peut, selon les contextes, conduire à l'identification de la personne, et souvent d'une manière qui passe au premier abord inaperçue pour l'enquêteur. Ainsi, par exemple, la mention d'un détail rare dans l'interaction (une pathologie rare de la personne, un attribut extraordinaire, une caractéristique unique et connue dans la région de la personne…) peut se révéler significative pour certains (dans certains cas sans que l'enquêteur ne s'en aperçoive).
- Le caractère reconnaissable de ces détails dépend de manière cruciale du contexte de réception et plus spécifiquement du public qui consultera ou prendra connaissance des corpus. Ainsi les membres d'un département d'anesthésie reconnaitront facilement un de leurs collègues sur la base d'expressions typiques, d'expertises spécifiques ou de façons propres de parler ou d'agir ; en revanche les mêmes détails passeront inaperçus chez les professionnels d'un autre hôpital ou a fortiori chez des étudiants en linguistique. Mais, là encore, la reconnaissabilité ne dépend pas simplement de l'éloignement géographique ou social du contexte dans lequel ont été enregistrées les données : les personnes sont mobiles dans l'espace et dans les milieux sociaux et il n'est pas impossible que le fils d'un patient puisse reconnaitre son père dans un cours universitaire portant sur des consultations thérapeutiques. La valeur identifiante d'un détail dépend donc du contexte de réception des données.
- Selon les cas, la référence à une institution ou à un organisme peut rendre nécessaire ou non l'anonymisation : par exemple la référence à une grande enseigne doit être anonymisée s'il s'agit du lieu de travail d'un employé, mais n'a pas besoin de l'être si elle intervient comme élément du paysage dans une indication d'itinéraire, et doit à nouveau être anonymisée si elle est citée dans des propos diffamatoires.
- D'autres aspects sont liés au *recoupement* d'informations venant de plusieurs sources (cela peut concerner par exemple la relation entre données anonymisées et métadonnées).

Pratiques d'analyse

Les limitations de l'anonymisation peuvent venir d'un autre type de considérations, davantage liées aux pratiques d'analyse des chercheurs.

Le problème fondamental est posé par la contradiction éventuelle entre anonymisation et disponibilité des détails pour l'analyse (sur le principe de disponibilité Mondada, 2003). En effet, les enregistrements et les transcriptions visent à produire la disponibilité des détails observables pour qu'ils puissent être exploités par l'analyse ; l'anonymisation au contraire peut rendre indisponibles certains de ces détails en les effaçant ou en les transformant.

Cela peut être le cas par exemple de l'anonymisation par bipage d'un nom qui est prononcé en chevauchement avec un autre tour de parole et qui rend impossible l'analyse de ce chevauchement.

Cela peut être le cas de l'anonymisation de numéros de téléphone lors d'appels d'urgence qui rend indisponible la manière dont l'appelant donne son numéro de téléphone dans une situation de stress et d'émotion et qui peut donc affecter de manière cruciale cette information.

Cela peut être le cas de l'anonymisation des visages sur une bande vidéo qui rend impossible une analyse des regards.

De manière analogue, le filtrage de la voix (tel que pratiqué par les médias) n'est pas envisageable pour la plupart des études linguistiques qui se basent sur les qualités intrinsèques du signal sonore.

C'est pourquoi les chercheurs affirment souvent la nécessité et revendiquent le droit de travailler – en garantissant la sécurité et l'inaccessibilité des données – sur des données non anonymisées, de les conserver sous cette forme et de faire intervenir l'anonymisation le plus tardivement possible et d'une manière qui tienne compte de ce qui est pertinent pour l'analyse.

3.5 TRANSCRIPTION

La transcription est une pratique qui, loin de se limiter à un exercice technique de reproduction, intègre de nombreux enjeux théoriques et interprétatifs (déjà Ochs, 1979). Dans le passage de l'oral à l'écrit graphico-visuel, de nombreuses opérations de catégorisation sont effectuées, soit quant aux formes linguistiques, segmentées visuellement en unités (Blanche-Benveniste & Jeanjean, 1987 ; Mondada, 2000), soit quant à l'identité des locuteurs eux-mêmes (Mondada, 2003). Du point de vue de la protection de l'image et de l'identité des personnes enquêtées et enregistrées, il convient d'apprécier ces effets pour éviter la surinterprétation, la stéréotypisation (Jefferson 1996) et la stigmatisation des locuteurs et de leurs façons de parler. Nous nous limiterons ici à ces enjeux de la transcription ; dans la section suivante, nous prendrons en compte un tout autre aspect, celui des questions de standardisation des transcriptions et de leurs conventions.

3.5.1 LES DESCRIPTIONS ETHNOGRAPHIQUES

La transcription est souvent accompagnée d'une brève description ethnographique qui esquisse le contexte dans lequel elle a été recueillie ainsi que le type d'activité et l'identité des participants. Cette description, qui intègre des éléments issus des métadonnées du corpus, peut avoir plusieurs effets sur la lecture (ou sur la réception d'un exposé oral) :

– Elle peut contenir des informations, permettant l'identification des personnes, qui entrent en contradiction avec les principes de l'anonymisation.
– Elle peut contenir des indications qui forcent la lecture ou l'interprétation des données. En restituant l'appartenance à telle catégorie ou à telle autre dimension pertinente de l'enquêté, ces indications peuvent donner une image particulière de l'activité et des locuteurs.
– En particulier, elle peut contenir des allusions, permettre des inférences qui renforcent certains stéréotypes (voire qui les utilisent pour provoquer des effets comiques pour conquérir le public – cela n'étant pas rare dans les exposés oraux).

Ces remarques ne concernent pas uniquement la description des données mais aussi les noms des corpus, qui peuvent parfois intégrer des éléments confidentiels. Dans ce sens, même si cela a souvent une fonction mémorielle, il convient d'éviter d'intégrer le nom des acteurs concernés dans le nom du corpus.

3.5.2 L'IDENTIFICATION DES LOCUTEURS

La transcription doit intégrer les résultats de l'anonymisation. Là où l'annotation prévoit un codage des tours de parole, des parties de transcription peuvent être attribuées à des locuteurs distincts et identifiés de diverses manières. L'usage des pseudonymes est assez répandu, mais d'autres possibilités sont envisageables, qui ont néanmoins des effets variables sur l'interprétation du texte qui les suit. Tout choix effectué en la matière pose le problème de la manière dont est traité le locuteur. Par exemple :

– A, B, C… : solution qui est la moins connotée mais qui en adoptant l'ordre alphabétique ordonne les locuteurs en premier, deuxième, troisième....
– E1, E2, E3... (pour des élèves) : choix qui homogénéise les personnes au sein d'une même classe, désignée par une catégorie unique. La même chose vaut pour L1, L2, L3 où L renvoie au Locuteur : si le linguiste peut considérer que tous les locuteurs sont égaux et que les acteurs sociaux l'intéressent avant tout en tant qu'êtres parlants, du point de vue de l'activité en cours, ceux-ci participent d'abord sous d'autres catégories, que ce soit enquêteur/enquêté, père/fils, médecin/patient, etc.
– H, F (pour homme et femme) : là encore, le choix privilégie la catégorie du sexe/genre sur toute autre catégorie, en postulant ainsi la pertinence généralisée de cette catégorie pour la compréhension des activités en cours.

Ces remarques invitent à se demander quels effets interprétatifs produisent les choix des identifiants. Il convient de ce point de vue de se demander quels sont les identifiants pertinents pour les participants – surtout dans des démarches analytiques qui se préoccupent de la perspective des participants (comme l'analyse conversationnelle). C'est pourquoi les solutions alternatives peuvent être les suivantes :

– EVA, MAR, ROB, AND… : indication des 3 premières lettres des pseudonymes, que ce soit des prénoms ou des noms propres – selon la tonalité de la conversation,

– APP/OPE pour appelant/opérateur ou DOC/PAT pour docteur/patient, ou encore INTE/IEUR pour interviewé/intervieweur lorsque l'activité institutionnelle est régie par des paires catégorielles de ce type. Sur ces questions, on peut renvoyer aux réflexions de H. Sacks sur les catégorisations des personnes et sur la pertinence des catégories selon l'activité et le contexte en cours (une personne qui est médecin dans un contexte peut très bien être père de famille dans un autre ; la manière de l'identifier dépend donc de l'activité en cours) (Sacks, 1972, 1992).

La notion de vie privée et d'intimité n'ayant pas la même valeur dans toutes les sociétés, il conviendra que le chercheur se renseigne sur les souhaits des locuteurs concernant l'anonymisation des données. Dans certaines communautés, le fait de ne pas mentionner les noms des personnes est considéré comme un manque de respect pour l'auteur du récit ou les personnes qui y participent, alors que dans d'autres, les mentionner est une atteinte à la vie privée. Sur ce point, il semble y avoir par exemple de grandes différences entre certains terrains en Afrique (où les locuteurs souhaitent être cités) et des terrains comme ceux de l'Amazonie, en particulier en Guyane française.

3.5.3 ENJEUX

Lorsqu'on transcrit, on prend sans cesse des décisions quant à la manière de représenter les locuteurs et leurs manières de parler. Ainsi, l'analyse – et parfois le jugement – se glissent immédiatement dans la pratique de la transcription. Nous soulignerons quelques enjeux des choix effectués dans la transcription elle-même.

ENJEUX (ORTHO)GRAPHIQUES

Depuis plus de vingt ans, de nombreuses discussions ont eu lieu sur l'emploi de l'orthographe standard, de l'orthographe adaptée et de l'API dans les transcriptions (voir 2.1.3). Les transcriptions phonétiques (API ou autres) ne sont lisibles que par les spécialistes et seulement pour des textes courts. Ainsi, pour lire de grands corpus, de nombreux linguistes européens ont choisi l'orthographe standard, mais proposent aussi de pouvoir superposer d'autres notations, lorsqu'il s'agit d'observer plus en détail certains phénomènes.

A l'inverse, dans certaines disciplines comme la phonétique, une transcription orthographique peut dans certains cas être non pertinente (par exemple pour la transcription de logatomes, de pseudo-mots, etc.).

Toutefois, la représentation écrite de la langue surprend souvent les locuteurs, et peut même leur déplaire considérablement. Il arrive qu'ils refusent l'image de leur langue transmise par la transcription, qu'ils désavouent le chercheur et qu'ils refusent son travail.

LA REPRESENTATION DU PARLER EXOLINGUE

Le choix de transcrire en API certains passages ou uniquement ceux de certains locuteurs plutôt que d'autres permet certes une plus grande précision dans la représentation des détails de leur parler mais risque aussi de provoquer des effets d'asymétrie non maitrisés.

Ainsi le recours à l'API et à l'orthographe adaptée peut produire des effets de stigmatisation et d'asymétrie à l'encontre de locuteurs « non-natifs » – lorsque ces derniers sont représentés de manière différente par rapport aux locuteurs « natifs » (ceux-ci par des notations standard, les « non-natifs » par des orthographes spéciales qui en mettent en relief non seulement la différence mais aussi l'« anormalité », l'« anormativité »).

De manière comparable, la notation explicite, par convention, de la variété de langue du locuteur (différenciation grâce à des polices, styles, alphabets spécifiques aux différentes langues utilisées dans une conversation bilingue, ou spécifique à l'interlangue de l'apprenant dans une conversation exolingue) opère une précatégorisation de cette variété : or cette variété se trouve être souvent un élément négocié par les participants et changeant au fil de la conversation (où par moment certaines formes sont marquées comme « étrangères » ou « étranges » et où à d'autres moments leur différence n'est pas du tout prise en considération).

Les mêmes questions se posent pour la traduction de la transcription :

- le fait de traduire les paroles de certains locuteurs plutôt que d'autres peut être considéré comme un jugement de valeur ;
- la façon dont on traduit, plus ou moins littéralement, peut amener à produire une version appauvrie de la parole du locuteur, et à en effacer ou au contraire à en souligner la différence ;
- différents formats existent pour la traduction (fournie en note, à la suite de l'original, ligne par ligne ; ou bien de manière à proposer un équivalent à la forme originale, de manière à respecter un lien quasi littéral à l'original, de manière à en fournir une glose grammaticale) qui produisent chacun une image différente de la culture et de la langue de l'autre (Traverso, 2003).

Précisons qu'il s'agit ici de traduction dans le cadre spécifique des corpus oraux. Cette traduction est indispensable pour le travail sur des langues autres que le français, mais reste souvent un outil pour le chercheur, et dans ce cas il ne doit pas chercher à être le reflet de la parole du locuteur. Elle doit s'accompagner de renseignements métalinguistiques qui permettent de mieux retranscrire les nuances nécessaires à une analyse approfondie de la langue. Ainsi, si une publication bilingue du corpus est prévue, un véritable travail de traduction devra alors être envisagé, dans une optique totalement différente de celle du recueil des données en vue de l'analyse de la langue.

Enjeux du multimodal et du detail de la transcription

Le fait de ne noter que les activités verbales et d'ignorer d'autres indications communicationnelles – comme c'est actuellement le cas dans la plupart des transcriptions – peut produire une image aberrante de certains comportements des locuteurs. Cela peut être le cas notamment de locuteurs aphasiques ou d'enfants s'exprimant par d'autres moyens que les moyens linguistiques standards : ne pas tenir compte de la totalité des ressources mobilisées par ces locuteurs signifie en donner une image réduite, qui pathologise ou anormalise leur comportement.

De même, différents degrés de granularité de la transcription (Jefferson, 1985) peuvent nuire à la représentation de conduites non-standards (par ex. la vocalisation prononcée par un patient aphasique peut être significative et demander une transcription adéquate ; mais elle peut aussi être réduite à un simple « bruit » sans aucun sens dans une transcription superficielle).

Le caractère plus ou moins approfondi ou détaillé de la transcription ne répond donc pas uniquement à des exigences scientifiques ; elle répond aussi à des exigences éthiques et juridiques, qui permettent de nuancer et de complexifier l'image que l'on donne des locuteurs – en s'éloignant d'autant plus du risque de le caricaturer et de le stigmatiser à travers des comportements stéréotypés.

4 LES CORPUS ORAUX, OBJETS DE PATRIMOINE ? UNE SOLUTION POUR LA PRESERVATION ET L'ACCES AUX CORPUS ORAUX ?

4.1 RAPPEL DE LA SITUATION

LES CORPUS ORAUX, PRODUITS PAR DES CHERCHEURS, AU SEIN DES INSTITUTIONS

L'enregistrement de corpus oraux s'inscrit dans une histoire déjà longue d'un siècle, à laquelle la possibilité de fixer la voix a conféré une dimension nouvelle et singulière. Dès 1896, érudits, chercheurs (anthropologues, ethnomusicologues, linguistes) fixent leurs collectes sur des cylindres, puis des disques. Les chercheurs étant conscients de créer des collections nouvelles à transmettre aux générations futures, les productions enregistrées lors des « missions ethnographiques » trouvent naturellement place dans des instituts sous l'égide de l'État. Les Archives de la Parole, conservatoire des langues et dialectes de France, naissent au sein de l'Université de Paris en 1911, la phonothèque du Musée de l'Homme en 1932, la Phonothèque Nationale en 1938, et elle sera en 1977 intégrée au sein du Département de l'Audiovisuel de la BnF. Les grandes collectes ethnographiques menées par le Musée National des Arts et Traditions Populaires[11], également Centre d'ethnologie de la France, concernent par exemple la Bretagne en 1939 et l'importante enquête pluridisciplinaire sur l'Aubrac qui, entre 1964 et 1968, produisit notamment près de quatre mille phonogrammes et une douzaine de films. Ce sont les linguistes puis les ethnologues qui se soucient de façon prioritaire de l'avenir de leurs enregistrements, y compris de leur utilisation par d'autres chercheurs. Dans les années 70, certains sociologues comme Daniel Berteaux[12] introduisent le « récit de vie » dans leurs méthodes. Cette piste ouvre la voie à des recherches pluridisciplinaires dont les « Ethnotextes » ont constitué une voie, expérimentée par Jean-Claude Bouvier et Philippe Joutard.

MAIS LA FRANCE EST UN PAYS DE TRADITION ECRITE ET L'ORAL NE BENEFICIE PAS DE VALEUR CULTURELLE, ENCORE MOINS DE STATUT PATRIMONIAL.

L'Université n'a donc pas développé de méthodologie critique spécifique et adaptée à sa problématique. L'absence de vocabulaire normalisé pour définir les différentes formes de corpus oraux est révélatrice de l'absence de statut scientifique et patrimonial des corpus oraux. Chaque discipline utilise sa terminologie en lui conférant un sens précis. Claude Martel[13] rappelle la variété des définitions connues

[11] Le MNATP est devenu en juin 2005 le MCEM, Musée national des Civilisations de l'Europe et de la Méditerranée.

[12] Daniel Berteaux, « L'approche biographique. Sa validité méthodologique, ses potentialités » *Cahiers internationaux de sociologie*, 1980.

[13] Voir l'article de Claude Martel « la recherche et les sources orales, les mots pour le dire » in : *Bulletin de liaison des adhérents de l'AFAS* 10, 1998.

pour des termes comme récits de vie, témoignages, entretiens selon le domaine disciplinaire de celui qui les utilise.

Les historiens ont éprouvé pendant fort longtemps des réticences à considérer le témoignage oral comme une source fiable et digne de considération. Philippe Joutard, un des promoteurs de l'histoire orale, rappelle l'isolement de la France face aux autres pays européens comme par exemple la Grande-Bretagne, l'Italie, l'Espagne, l'Argentine qui connaissent, au sein même de l'Université, un développement dynamique et foisonnant de cette discipline. De nombreuses revues attestent de cette vitalité (voir bibliographie).

L'excellente enquête[14] menée entre 2001 et 2003 à la demande du Ministère de la Recherche par Françoise Cribier et Elise Feller, a prouvé que, dans les trente dernières années, les chercheurs français, dans toutes les disciplines des sciences humaines et sociales à l'exception de l'histoire, ont énormément enregistré. Mais leurs enregistrements, sans reconnaissance officielle ni lieu pour les accueillir, sont restés dans les laboratoires. Surtout ils n'ont été ni décrits ni documentés, et les autorisations des témoins, lorsqu'elles existent, sont limitées, dans le meilleur des cas, à l'usage des chercheurs qui les ont réalisés.

Les fonds sont souvent conservés en mains privées car, la plupart du temps, les collectes orales réalisées lors des campagnes d'enregistrement officielles embarrassent les pouvoirs publics. A cet égard, la grande entreprise coordonnée par la DGRST au début des années 1960 autour de Plozévet, village bigouden, est tout à fait exemplaire. L'enquête très importante, menée par le Musée de l'Homme, qui a duré pendant près de cinq années, a mobilisé des historiens, des géographes, des sociologues, des économistes, des ethnologues. Nombre d'entre eux étaient équipés de magnétophones. Mais cette enquête, au lieu de fournir un travail pluridisciplinaire, n'a produit qu'un ensemble de monographies et personne ne s'est soucié des enregistrements réalisés, à l'exception de ceux produits par l'ethnologue Donatien Laurent. Il est l'un des rares chercheurs qui, non seulement a documenté l'ensemble de sa collecte, mais l'a déposée au Centre de recherche et de culture celte et bretonne de l'Université de Brest. Aujourd'hui ses enregistrements sont numérisés et consultables dans le cadre universitaire. Les autres enregistrements ont été perdus, ou, par manque de crédits, les bandes ont été réenregistrées.

L'ERE DU NUMERIQUE ET DU TRAVAIL EN RESEAU : LES ANNEES 80

Aussi, ces collections sans statut scientifique posent, pour certaines encore, du point de vue de leur préservation et de leur consultation, des questions juridiques toujours non résolues.

---

[14] Cribier F. & Feller E. (2003) *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile »* rapport présenté à Madame la Ministre déléguée à la Recherche et aux nouvelles technologies. dactylogr. 2 vol.
http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf Une autre enquête succincte a été réalisée par Dubar C. à la demande du CNRS (voir bibliographie).

Enregistrés en analogique, les documents sonores ne peuvent être consultés qu'en temps réel. Leur indexation ne suffit pas toujours à en prendre rapidement connaissance. Ce travail rebute la plupart des chercheurs.

Dans les années quatre-vingt, les techniques de numérisation[15] marquent un nouvel intérêt pour l'oral, donnée sensible et contenu souvent unique. En effet les enregistrements produits numériquement, indexés par le chercheur lui-même au moment de sa réalisation, permettent de « feuilleter » rapidement le son comme on peut le faire avec de l'écrit.

Mais, si les techniques numériques ont, comme pour l'écrit et l'image, révolutionné l'accès aux corpus oraux, elles ont introduit, par le caractère parfait des copies réalisées, une autre révolution intellectuelle beaucoup plus importante, notamment pour les usages ultérieurs. *En gommant la notion d'original, elles ont oblitéré les repères* qui jusqu'alors jalonnaient le domaine des collections. Versés par leur producteur au sein d'une institution patrimoniale, les corpus oraux deviennent objets de collection mais il devient alors impossible de distinguer entre le premier enregistrement réputé « original » et les copies successives d'un corpus oral.

Le support ne permettant plus d'identifier les différents éléments, qui décidera de sélectionner et de figer l'instant T de la version qui, en entrant dans une institution, témoignera de la recherche de son producteur ? Quel type de *métadonnées* seront simultanément intégrées aux collections ?

COLLECTIONS ORALES SANS STATUT

Les corpus oraux ne figurent pas dans le Code de la Propriété au titre des œuvres protégées, sauf si elles ont une forme identifiée et, comme telle, protégeable : *les témoignages, les interviews, les entretiens, les émissions radiophoniques*.

D'une façon générale, les collections orales, mais également la dimension sonore en général, ne sont pas prises en compte dans cette grande entreprise culturelle lancée en 1964 par André Malraux : *l'Inventaire général des monuments et richesses artistiques de la France.* Aucun des dispositifs qui fondent un patrimoine[16] ne leur est attribuable. Pas de classement ni d'inscription et, par voie de conséquence, aucune commission spécialisée « du patrimoine » ne s'en préoccupe. Seule, l'UNESCO a pris des initiatives dans ce sens (voir fiche *UNESCO*). Plus modestement, la Mission du Patrimoine ethnologique créée dans les années 80 au sein du Ministère de la culture et de la communication va placer les corpus oraux au rang d'objets. Cette préoccupation a très vite disparu des programmes.

---

[15] « Musique et son : les enjeux de l'ère numérique. Création musicale, recherche, archivage, transmission », *Culture et Recherche* 91-92, 2002.

[16] Sur le terme très galvaudé de « patrimoine » on lira Jean-Pierre Babelon et André Chastel, *La notion de patrimoine*, Liana Levi, 1994 et l'analyse historique très complète que lui a consacrée André Desvallées, « Emergence et cheminement du mot Patrimoine » dans *Musées et collections publiques de France* 208 : 6-29,1995.

4.1.1 LES COLLECTIONS DE CORPUS ORAUX

PRATIQUES ET USAGES DES INSTITUTIONS PATRIMONIALES

Les corpus oraux produits de façon unique par des producteurs individuels ou institutionnels ne constituant pas une catégorie particulière au regard du patrimoine et du Code de la Propriété intellectuelle, le législateur n'a pas prévu de dispositif particulier pour les collecter et organiser leur préservation.

L'Université s'est désintéressée de cet ensemble riche et foisonnant qui ressortissait de domaines disciplinaires trop diversifiés. *Il n'existe donc pas de dépôt légal des corpus oraux*.

Les corpus oraux ne peuvent être protégés dans une institution patrimoniale qu'au travers d'une *initiative volontaire* (don ou dépôt) de celui qui les a collectés ou par *décision de l'institution* soucieuse de constituer des collections orales sur des thématiques qui lui sont propres. Les institutions patrimoniales peuvent donc être, à la fois ou successivement, productrices de corpus oraux et conservatrices de documents oraux produits par d'autres. Les institutions responsables de ce type de collections engagent des recherches sur la conservation des documents sonores.[17] Elles mettent également en œuvre des critères sélectifs de constitution des fonds.

- De façon générale, c'est *le principe de cohérence des fonds* qui préside à la constitution des collections au sein des institutions patrimoniales (archives, bibliothèques patrimoniales, musées). Un enregistrement isolé ne signifiera que pour lui-même. L'enregistrement unique de la voix d'un écrivain dans le musée qui lui est consacré demeure anecdotique.
- Cela signifie que la constitution d'un fonds cohérent est le résultat d'une *politique de tri et de sélection exigeante* selon les axes prioritaires définis par l'institution (fonds parlé pour la BnF, fonds sur la déportation pour les Archives Nationales) mais suffisamment larges et complets pour qu'ils constituent pour demain des sources de référence significatives. Dans les musées de société, héritiers des écomusées définis dans les années 1970 à l'initiative de Georges-Henri Rivière, la collecte d'enquêtes orales vise à combler l'absence d'objets ou leur difficulté à témoigner de la dimension humaine à l'intérieur d'une collectivité. A Fécamp, l'enregistrement des ouvrières des anciennes pêcheries révèle une forme d'organisation sociale de la cité dans la première moitié du 20e siècle dont aucun objet ni aucun écrit ne peut rendre compte[18]. Il en est de même au Musée de la manufacture des tabacs à Morlaix, à l'Ecomusée de la communauté urbaine du Creusot-Montceau-les-Mines (Saône-et-Loire).

---

[17] Calas, M.-F. Fontaine, J-M. (1996) *La conservation des documents sonores*, Paris : CNRS-Editions.

[18] Cette série d'entretiens réalisés en collaboration entre le Musée et le service d'archives municipales a donné lieu à un disque avec livret *Femmes de marins, compagnes de pêche*, Fécamp, Musée des Terre-Neuvas, 2003.

- La collecte n'est pas toujours considérée comme un objet de collection ou comme une œuvre. A la BnF, aux Archives nationales, le traitement documentaire n'est pas déterminé par le support de la collection. Rien de semblable dans les musées. A l'exception du Musée National des Civilisations de l'Europe et de la Méditerranée (ancien Musée national des Arts et Traditions populaires), du Musée Dauphinois qui, très tôt, a intégré au même titre que les objets, les enquêtes de Charles Joïsten sur l'inventaire du musée, la plupart des musées comme le Musée-conservatoire de Salagon par exemple, portent les corpus oraux sur des inventaires de type bibliothèque. De même, l'Ecomusée de Saint-Quentin-en-Yvelines, a choisi d'inscrire les entretiens qu'il mène avec les acteurs politiques et les habitants sur un registre à part qui répertorie les collections d'études. A la fin des années 90, on a assisté à un intérêt très fort, voire excessif, pour la quête identitaire et le devoir de mémoire. Ces archives orales ne bénéficient pas encore d'une reconnaissance bien établie.
- Les collectes orales ne sont pas réductibles à l'enregistrement des voix. Elles ne prennent sens que dans la mise à disposition des données temporelles, techniques, scientifiques de leur production. L'ensemble de ces éléments de contextualisation (métadonnées), spécifiques du corpus enregistré, constitue avec lui un tout indissociable, sans lequel l'enregistrement serait privé de temporalité et de sens. Et on pourrait alors lui faire signifier tout et son contraire.
- Comme tout objet patrimonial, le document oral, bien que daté, identifié, n'est pas, comme nombre de chercheurs l'ont cru très longtemps, réductible au seul usage de son producteur. Les enquêtes orales dépassent souvent le projet dans lequel elles ont été menées. Elles peuvent être utilisées dans le cadre d'autres disciplines

  > *« Une nouvelle lecture conduit à porter un autre regard sur ce qui a été dit, parce que le temps a passé, et que les questions qu'on se pose se sont déplacées »*[19].

- Elles doivent pouvoir être analysées, au cours des temps, par différents chercheurs à travers leur grille d'analyse personnelle. Mais le Plan de numérisation des documents sonores mis en place fin 1999 par le Ministère de la Culture et de la Communication a révélé le déficit d'informations relatif à ces collections orales. Certains fonds considérés comme historiques ne pouvaient témoigner convenablement de leur intérêt en l'absence de documents indispensables de contextualisation. En outre, aucune des collections ayant répondu à l'appel à numérisation ne détenait les droits d'exploitation permettant d'organiser la consultation du public, notamment via internet.

[19] Françoise Cribier & Elise Feller, *op. cit.*

- De quel type de protection les corpus oraux bénéficient-ils dans les institutions publiques et privées ? Le versement d'une collecte au sein d'une institution n'a pas de **valeur probatoire*. La date de versement peut-elle indiquer une preuve éventuelle d'antériorité par rapport à un enregistrement qui se révèlerait être une contrefaçon du premier ? A l'exception des dépôts qui, par nature, sont toujours révocables, les collections entrent (sous la forme de supports ou de données numériques) de façon définitive et imprescriptible dans les fonds de l'institution. Cette cession, nous venons de le rappeler, n'emporte pas, sauf accord spécifique, cession des droits d'exploitation. Les institutions s'engagent a priori à assurer la pérennité physique et à organiser la consultation des corpus oraux dans le respect des droits de ceux qui ont participé à la création, mais il est indispensable que les cessionnaires cèdent au moins les autorisations de consultation. Depuis les années 80, la consultation à distance des collections a, en quelque sorte, « réveillé » l'intérêt pour les corpus oraux, et laissé entrevoir des possibilités d'accès autrefois inimaginables L'accessibilité aux corpus analogiques pose le problème en termes de conservation et d'identification des sources préalables à la numérisation. Il se chiffre également en moyens financiers et en personnel.

4.1.2 LA BIBLIOTHEQUE NATIONALE DE FRANCE

CONSERVATOIRE DE L'ORALITE

Héritier des Archives de la Parole fondées en 1911 par Ferdinand Brunot, du Musée de la Parole et du Geste qui leur succède en 1928, puis de la Phonothèque nationale créée en 1938, le département de l'Audiovisuel de la Bibliothèque nationale de France inscrit son action dans la continuité de ces institutions. Aujourd'hui, c'est donc plus d'un siècle d'une mémoire de l'oralité qui est ainsi conservée et mise à la disposition du public.

Mais, parallèlement, le département de l'Audiovisuel mène une politique active de développement de ses collections, notamment dans le domaine de l'oralité. En effet, outre la collecte du dépôt légal (voir fiche *Bibliothèque nationale de France*), la Bibliothèque nationale de France a vocation et mission d'enrichir ses collections par acquisitions, donations, dons, legs, dations, etc. C'est donc le cas du département de l'Audiovisuel qui, de manière complémentaire au dépôt légal des documents sonores, vidéographiques, multimédia et informatiques dont il a la charge, a défini les grands axes d'une politique d'enrichissement de ses collections en matière d'enregistrements sonores inédits. On trouvera à la fin de cette présentation quelques-uns des fonds entrés récemment au département de l'Audiovisuel, représentatifs de la place de l'oral dans ses collections.

LES GRANDES LIGNES D'ENRICHISSEMENT DES COLLECTIONS DU DEPARTEMENT DE L'AUDIOVISUEL

Le département de l'Audiovisuel définit comme documents « inédits », des documents « source » à l'état « unique », non diffusés en nombre, et qui ne sont pas

déterminés par une forme éditoriale précise. Cela posé, face à l'extension indéfinie du champ et à la multiplicité des contenus (linguistique, ethnologie, histoire orale…), à la multiplicité des sources possibles (institutionnelles, chercheurs indépendants…), à la nécessaire complémentarité avec d'autres institutions en même temps que face aux vides à combler en matière de conservation, de diffusion et de valorisation, le département de l'Audiovisuel a déterminé un certain nombre de principes forts à même de guider sa politique d'enrichissement en la matière.

Le critere documentaire et patrimonial

La politique du département repose tout d'abord sur un principe de sélection. Le critère fondamental qui amène à accepter ou à refuser un don d'inédits est avant tout l'intérêt documentaire et/ou patrimonial du fonds proposé. Ce critère peut être assimilé à celui de « mémoire nationale ». En d'autres termes, quels sont les enregistrements inédits que l'on peut considérer comme relevant d'une mémoire, d'un patrimoine national ? Ce critère ne limite pas le champ de la politique documentaire au « terrain » français, mais donne priorité aux fonds ayant – soit en termes de source (le collecteur, l'institution…), soit en termes de contenu – un rapport avec la France. Le don du fonds de Deben Bhattacharya, ethnomusicologue indien ayant enregistré à travers le monde, mais ayant vécu à Paris de 1954 à 2001, ou celui des collectes pygmées de Simha Arom (Lacito-CNRS), en sont l'illustration.

En étroite articulation avec ce critère d'intérêt documentaire et/ou patrimonial, et étroitement délimité par lui, le département de l'Audiovisuel accorde une attention privilégiée à des documents ou à des fonds pour lesquels n'existe a priori aucun lieu de conservation et/ou de consultation déterminé. C'est le cas, par exemple, de certaines archives personnelles ou de fonds en déshérence dans certains laboratoires, faute de structures appropriées.

L'acceptabilite du fonds et le principe documentaire

Ce principe de sélection et les critères documentaire et patrimonial établis, des conditions d'acceptabilité sont posées quant à la réception d'un fonds. Il s'agit tout d'abord de conditions documentaires. Ainsi, pour être reçues ou acquises, les sources inédites doivent être documentées et/ou exploitables d'un point de vue documentaire. On pourra envisager, soit que le traitement documentaire soit fourni en même temps que l'archive sous forme de métadonnées ; soit, éventuellement, que toutes les informations soient fournies à la BnF sous une forme ou une autre pour lui en permettre le traitement documentaire.

L'acceptabilite du fonds et le principe juridique

Les conditions juridiques forment une autre composante des conditions d'acceptabilité. La personne – physique ou morale – qui réalise le don doit notamment s'assurer :

– Qu'il est le propriétaire des supports physiques sur lesquels ont été réalisés les enregistrements, et que ces enregistrements sont susceptibles d'être donnés à la Bibliothèque ;

– Qu'il est titulaire ou qu'il peut garantir, les droits d'auteur sur les œuvres réalisées et les droits voisins du producteur de phonogrammes et éventuellement des interprètes musicaux.

Pour la BnF, recevoir les supports nécessite également de disposer des droits d'auteur et droits voisins requis pour leur reproduction et leur communication aux lecteurs, les documents sonores devant faire l'objet d'actes de reproduction et de représentation pour être conservés et consultés. Or, la personne – physique ou morale – qui réalise le don n'a pas toujours la capacité juridique de délivrer ces autorisations de reproduction et de communication.

Doivent pouvoir être cédés à la BnF :

– le droit de reproduction du document, c'est-à-dire la possibilité de transférer son contenu sur un support idoine (numérique) pour des raisons de conservation du signal ;
– le droit de représentation. Ce droit se comprend comme étant, au minimum, la possibilité d'une consultation par le public de chercheurs en salle P (au niveau « Recherche » de la Bibliothèque). On pourra admettre le principe d'une autorisation de communication au cas par cas. De même, pour certains documents, on acceptera qu'un délai de réserve de communication puisse être exigé pour des raisons autres que celles tenant au droit d'auteur (confidentialité de données relatives à la vie privée…).

Quelques exemples parmi les derniers fonds inedits reçus en don par le departement de l'Audiovisuel

(classés par ordre d'arrivée dans les collections) :

– fonds des atlas linguistiques régionaux (1979 et suivantes) ;
– fonds du Centre de Recherche Historique, EHESS/CNRS (1979) : histoire orale, récits de vie, années 1970-1980 ;
– fonds Félix Quilici (1981) : musiques corses de tradition orale, 1959-1963 ;
– fonds Geneviève Massignon (1985) : collectes ethno-linguistiques, Acadie, Ouest de la France, Corse…, 1946-1963 ;
– fonds Nicole Revel (1995) : épopées Palawan, Philippines, années 1980 ;
– fonds Gilles Deleuze (1997) : cours, Université Paris VIII, 1979-1984 ;
– fonds Deben Bhattacharya (2003) : collectes ethnomusicologiques, Asie, Europe…, 1954-2000 ;
– programme « Archivage », LACITO/CNRS (2005) : langues rares, transcriptions, annotations, http://lacito.vjf.cnrs.fr/archivage/

4.1.3 LES ARCHIVES DE FRANCE

Dans le Livre II du Code du Patrimoine, les archives sont définies à l'article L 211-1 comme suit :

« Les archives sont l'ensemble des documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé, dans l'exercice de leur activité. La conservation de ces documents est organisée dans l'intérêt public tant pour les besoins de la gestion et de la justification des droits des personnes physiques ou morales, publiques ou privées, que pour la documentation historique de la recherche. ». Les archives constituent deux catégories : les archives publiques qui procèdent de l'activité de l'État, des collectivités locales et des entreprises publiques et les archives privées (voir fiche *Archives : législation*).

C'est le mode de production et non le type de support ou le sujet qui définit l'appartenance à l'une ou l'autre catégorie. L'enregistrement d'une séance du Conseil général est un document d'archive public alors que l'enregistrement d'un personnage politique à la radio est un document d'archive privée.

La consultation des fonds sonores varie selon qu'il s'agit d'archives publiques ou privées. Si les premières sont clairement réglementées, c'est la volonté du déposant qui fixe les règles en matière d'archives privées.

QUELQUES EXEMPLES DE CORPUS ORAUX DANS DES FONDS D'ARCHIVES

- *Les Archives nationales*

Placées sous la responsabilité de la direction des Archives de France, elles regroupent cinq centres.

- Le Centre Historique des Archives Nationales (CHAN) à Paris. C'est au sein de la section XX^e^ qu'a été créée dans les années 80 une cellule d'archives orales. Cette cellule reçoit des versements, par exemple ceux réalisés par la Fondation pour la mémoire des déportés, mais elle produit des témoignages en complémentarité des archives écrites « en disant ce qui ne s'écrit pas, en redimensionnant l'évènementiel à l'échelle humaine, et en venant le cas échéant, par la narration de détails occultés, combler les lacunes historiques existantes » [20]. Il en va de même pour les enregistrements vidéo des archives judiciaires (procès de Klaus Barbie, de Paul Touvier ou du « sang contaminé ») et les Archives de la Présidence de la République : discours et conférences de presse des présidents de la République Georges Pompidou et Valéry Giscard d'Estaing.
- Les sources créées par les conservateurs correspondent à deux approches : le récit autobiographique sert l'écriture de l'histoire des élites, et les corpus thématiques peuvent permettre de croiser

[20] Agnès Callu, « Aux Archives nationales, une politique raisonnée en faveur des témoignages oraux » *Colonnes : archives d'architecture du XX^e^ siècle*, 20, décembre 2002.

plusieurs récits sur un même fait (par exemple la fonction d'instituteur dans les années 50).

- Le Centre des Archives contemporaines (CAC) à Fontainebleau. C'est là, par exemple, que sont versées les 400 heures d'enregistrement réalisées dans le cadre du programme lancé par le Comité d'histoire de la Sécurité sociale par Dominique Aron-Schnapper (voir point : Statut des collections d'archives...).
- Le Centre des Archives du Monde du Travail (CAMT) à Roubaix qui collecte tout type d'archive sur son domaine, dont des enregistrements.
- Des deux autres centres, celui d'Esperran ne conserve que des microfilms et celui des archives d'outre-mer conserve surtout un fonds imprimé clos.

- *Les Services d'archives départementales*

Décentralisés bien avant d'autres, ces services collectent souvent des copies d'émissions de radio, des films d'amateurs, des documentaires, conduisent des programmes d'enquêtes orales seuls ou avec des concours associatifs et universitaires. Leur situation est très diversifiée et l'importance des fonds oraux tient aux thématiques couvertes et surtout à la motivation et à l'intérêt du directeur.

Les services d'archives municipales, dans le courant de la *patrimonialisation* de la mémoire, ont souvent confié la réalisation d'archives orales à des emplois-jeunes recrutés sur des postes de « gardiens de la mémoire » (exemples : Martigues, Lille).

4.1.4 PLACE DES CORPUS ORAUX DANS LES MUSEES

Est considéré comme musée, dans son acception la plus large, toute collection permanente composée de biens dont la conservation et la présentation revêtent un intérêt public, et organisée en vue de la connaissance, de l'éducation et du plaisir du public.

Les collections sont constituées de tout type d'objet et d'œuvre dont la matérialité est tangible.

Les enregistrements oraux, par définition, constituent, pour le musée, de l'immatériel. Pourtant l'ICOM, l'association internationale des collections de musées, ONG qui au sein de l'UNESCO, préside au développement de toutes les formes de musées, a lancé le débat en 2004 sur la dimension immatérielle du patrimoine intangible. Le malaise ressenti par les musées occidentaux en général, face à l'intégration de la dimension sonore, audiovisuelle, paysagère au sein des musées, révèle parfaitement cette forme de contradiction, pour un musée, entre objets et oralité.

Par contre, les musées d'histoire, les écomusées, les musées de société utilisent parfois depuis de très longues années (exemple, le Musée dauphinois de Grenoble) l'enregistrement de la mémoire orale comme un des éléments essentiels du projet culturel et scientifique autour duquel le musée va s'organiser. Les collections sonores sont inscrites sur le Registre d'inventaire du Musée comme les autres collections, mais le cas est loin d'être généralisé et nombre d'enregistrements sonores et vidéo sont au mieux inscrits sur le Registre des collections d'étude ou documentaires.

Si les corpus oraux étaient, comme au Musée dauphinois, reconnus comme des œuvres inscrites sur le Registre d'inventaire dont les modalités de rédaction sont définies par les textes législatifs, ils seraient inaliénables et imprescriptibles.

4.1.5 *LES « CORPUS ORAUX » A L'INA*

Au travers de la consultation du dépôt légal de la radio télévision, l'Ina, de fait, donne accès à une grande variété de corpus oraux constitués par divers témoignages, paroles, allocutions, discours enregistrés dans une perspective de diffusion.

Les chercheurs, usagers du Centre de consultation de l'Inathèque, constituent pour leurs besoins spécifiques des corpus à partir des sources de la radio télévision, qui s'inscrivent dans différentes logiques disciplinaires d'exploitation de corpus oraux : linguistique, sociologie, histoire…

L'étude de ces corpus peut porter sur les procédés discursifs dans tel ou tel genre d'émission (l'interview télévisée, les commentaires radiophoniques…), sur différents types d'analyse du discours (politique, journalistique…), sur la création de répertoires lexicographiques, sur des analyses sociolinguistiques (la parole du danseur, paroles d'ouvrières) etc.

Certaines collections d'émissions archivées à l'Ina constituent d'emblée des « corpus fermés » de productions orales.

Pour n'en citer que quelques unes : « *Les archives du vingtième siècle* » produites par Jean-José Marchand, recueils d'entretiens avec des personnalités du monde littéraire et artistique, « *Les conteurs* », une collection réalisée par André Voisin, produite par le service de la recherche de l'ORTF, recueil d'histoires personnelles, régionales (*Ceux de La Hague, Au cœur de l'Aubrac*…).

Par ailleurs l'Ina est engagé depuis sa création dans la production de collections d'enregistrements patrimoniaux et de recueil de témoignages.

Ces entretiens, de durée variable (jusqu'à 15 heures d'entretien), sont accessibles par le biais d'une interface de consultation interactive, «@propos», facilitant la navigation dans le programme.

- Ainsi, la collection « Musique Mémoires » est fondée sur une campagne d'archivage visant à recueillir le témoignage de compositeurs, interprètes, chefs d'orchestres et personnalités dont les créations et l'action ont marqué la vie musicale des soixante dernières années. Ces entretiens, menés par Bruno Serrou, explorent le parcours propre à chacun des artistes : origine, formation, influences, rencontres, exercice du métier... Entretiens déjà réalisés : François Bayle, Claude Helffer, Betsy Jolas, Claude Ballif, Pierre Boulez, Marius Constant, Antoine Duhamel, Luis de Pablo, Yvonne Loriod, Michel Fano, Ivo Malec.
- « Histoires d'historiens » offre une collection d'autoportraits d'historiens contemporains ; l'histoire de leur vie ainsi racontée permet une meilleure intelligence de leur œuvre. Entretiens déjà réalisés : Maurice Agulhon, Pierre Chaunu, Emmanuel Le Roy

Ladurie, Claude Nicolet, Pierre Nora, Robert Paxton, Madeleine Rebérioux, René Rémond, Zeev Sternhell, Jean Tulard.

- « Télé notre histoire » est une collection de longs entretiens offrant une véritable mémoire de la télévision racontée par ceux dont l'itinéraire personnel et la pratique professionnelle éclairent l'histoire de ce média : auteurs, artistes, producteurs, programmateurs, ingénieurs, techniciens, décideurs, pionniers ou praticiens plus récent. Entretiens déjà réalisés : Igor Barrère, Marcel Bluwal, Yves Jaigu, Jacques Krier, Claude Santelli, Pierre Tchernia…
- D'autres entretiens qui ne s'inscrivent pas dans une logique de collection offrent néanmoins les témoignages d'acteurs essentiels de la vie culturelle, scientifique et artistique contemporaine. Entretiens déjà réalisés : Françoise Gilot, K.S. Karol, Claude Lévi-Strauss.
- « Mémoires de la Shoah » en cours de production est une collection de 110 entretiens de 3 heures environ de témoins de la Shoah : déportés, orphelins, « justes ».

Toutes ces collections seront, à terme, accessibles au centre de consultation de l'Inathèque de France.

4.2 LES INITIATIVES PRIVEES

L'enregistrement de témoignages oraux connait depuis 1972 (date de la création de la Commission permanente d'histoire de l'Éducation) un développement notable au sein de programmes mis en place par les *Comités d'Histoire orale* créés par les institutions publiques soucieuses de valoriser la mémoire de leurs institutions.

Aujourd'hui on dénombre 67 comités et services[21] intégrés à une institution (Comité d'histoire du Ministère de la Culture et de la Communication, Comité d'histoire de la BnF).

L'*AHICF***,** Association pour l'histoire des chemins de fer en France, occupe une place à part. Elle se met au service des institutions dont elle se propose de faire l'histoire. L'AHICF, créée en 1987, a deux missions : recherche et sauvegarde du Patrimoine. Elle favorise la préservation des sources, mais n'a pas vocation à l'assurer elle-même. Il existe des services à la carte (historiens) pour aider à la création de la mémoire dans le domaine industriel.

D'une façon générale, ces comités considèrent les enregistrements réalisés comme des archives privées couvertes par le droit d'auteur La clause de dévolution des corpus oraux produits, au bénéfice d'un service d'archives en cas de dissolution des associations, est une règle assez répandue.

---

[21] Guide des Comités d'histoire et des services historiques. Paris, Comité pour l'Histoire économique et financière de la France

On peut citer parmi les partenaires actifs d'un réseau « archives orales » les Pôles associés de la BnF comme la FAMDT, DASTUM, la MMSH Maison Méditerranéenne des Sciences de l'Homme à Aix-en-Provence (cf. BnF, Pôles associés). Ces centres ne disposent que très rarement des droits complets des corpus qu'ils conservent.

4.3 L'ACCES AUX COLLECTIONS

Il n'existe pas de catalogue collectif des corpus oraux. Plusieurs initiatives ont permis d'identifier des structures institutionnelles et associatives qui produisent ou collectent des corpus oraux à des fins de préservation et de consultation. Encore font-elles davantage du signalement global que du détail des contenus, la plupart de ces corpus étant très peu décrits par leurs producteurs. La publication[22] qui vient de paraitre cette année résulte du dépouillement d'une vaste enquête sur les sources orales en sciences sociales conservées en France. Elle marquera peut-être, si le catalogue informatisé permet une actualisation par le réseau des producteurs, le début de la constitution d'une source collective pour l'oralité.

Les conditions de consultation sont définies par le contrat. Or, il n'existe pas de contrat-type.

Dans les institutions, les enregistrements oraux, dans leur majorité, sont traités à travers le Code de la Propriété intellectuelle. D'une façon générale, le témoin a un droit de regard sur l'utilisation de sa voix (loi du 17 juillet 1970). Nul ne peut fixer, conserver, divulguer sans son accord les propos et l'image d'une personne privée se trouvant dans un lieu privé. Le Code civil article 9 et le Code pénal article 226-1 obligent à obtenir le consentement écrit de la personne. Le témoin, s'il fait preuve d'originalité dans ses propos, peut être considéré comme auteur, et bénéficier à ce titre d'un droit moral et des droits moraux afférents. L'utilisation de son enregistrement peut passer par l'obligation d'une rémunération définie dans le cadre d'un contrat. Le collecteur devra obtenir l'autorisation de consultation la plus large.

> *L'accessibilité pose des questions de droit et de déontologie (respect de la vie privée, droit à sa voix pour un témoin, histoires de vie, témoignages délicats, propos qui risquent de devenir diffamatoires...). Or, pour des raisons qui tiennent à la nature du contenu (récits de vie et témoignages mettant en cause d'autres personnes, entretiens en milieu psychiatrique), ces corpus oraux ne peuvent être donnés en consultation sur place, encore moins être diffusés sur Internet.*

Chaque cas est donc particulier et la reconnaissance des droits des uns et des autres, relève d'une analyse fine et périlleuse au cours de laquelle les questions suivantes devront avoir reçu une réponse : qui détient des droits ? Le détenteur accepte-t-il de les céder, dans quelles conditions et pour quel usage ? Pour quelle durée ? De façon immédiate ? Différée ?

---

[22] Callu, A., Lemoine, H. (2004) *Patrimoine sonore et audiovisuel français : entre archive et témoignage : guide de recherche en sciences sociales*, Paris, Belin, 7 vol., 1 CD-Rom, 1 DVD-Rom.

Le collecteur-chercheur pour qui l'enregistrement des corpus constitue un moment dans une recherche approfondie devrait pouvoir être protégé en tant qu'auteur. Il est dans la plupart des cas appelé *collecteur*. Pour reconnaitre un droit d'auteur à l'intervieweur, il faudrait pouvoir mettre en évidence la forme originale de son propos.

Les institutions, en conséquence, elles ne peuvent souvent que donner en consultation dans leurs propres locaux, et les travaux de numérisation dont ils peuvent prendre l'initiative sont souvent faits sans autorisation des véritables détenteurs de droits (Plan national de numérisation).

Il subsiste bien des difficultés.

La question du collecteur salarié agissant dans le cadre de ses missions publiques, censé faire l'abandon de ses droits au bénéfice de l'État souligne un problème sur les droits des salariés « auteurs » qui bute dans la fonction publique sur des questions financières non résolues.

Que dire des droits que pourraient revendiquer des étudiants bien peu aguerris à la technique de l'interview et qui sont payés pour poser les questions dans l'ordre d'un questionnaire préétabli ?

Le statut des collections d'archives orales n'est pas indifférent

Le cas de la première grande enquête sur l'histoire de la sécurité sociale, conduite entre 1973 et 1975 par Dominique Schnapper à la demande du Comité d'histoire de cette institution créée en 1973, a permis l'enregistrement de 200 témoins qui ont donné lieu à 400 heures d'interviews et de témoignages. Il s'agissait, par définition, d'archives privées. Or, avant que ne débute la campagne, il a été décidé que l'ensemble de l'enquête serait classée comme une archive publique et, comme telle, consultable au bout de soixante ans. Cette décision a eu des conséquences importantes. Philippe Joutard à plusieurs reprises a évoqué cet exemple, dans lequel il voit une des raisons possibles du manque de dynamisme du développement de l'histoire orale en France.

De même, Florence Descamps partage cette analyse en stigmatisant ces archives orales novatrices qui ont été, dès le début « gelées ».

Les chercheurs, peu enclins à voir institutionnalisés leurs corpus, les ont gardés par-devers eux, peu encouragés par les organismes comme le CNRS et l'Université (exception : la convention signée entre le CNRS et la BN en 1979 pour la sauvegarde des Atlas linguistiques) qui, jusqu'à une date récente, n'ont jamais pris d'initiatives constructives pour préserver des corpus oraux qui échappaient à toute définition académique, alors que l'histoire orale connait en Grande-Bretagne où elle est née, tout comme dans les pays latins autres que la France, un grand foisonnement.

4.3.1 *QUEL RESEAU POUR DEMAIN ?*

UN RESEAU DE GESTION, DE PROTECTION DES COLLECTIONS DE CORPUS ORAUX ORGANISE PAR LES UNIVERSITES ET LES INSTITUTIONS DE RECHERCHE OU DES INSTITUTIONS PATRIMONIALES ?

En dehors des institutions patrimoniales, les universités, les organismes de recherche, à l'instar de ce qui existe dans de nombreux pays européens, pourraient avoir la capacité et la volonté de créer un grand réseau des sciences humaines et sociales, à travers lequel les corpus mis à la disposition des autres chercheurs pourraient être protégés et rendus accessibles à d'autres.

Le Rapport[23] rédigé par Françoise Cribier avec la collaboration d'Elise Feller a étudié la situation et les réseaux existant pour la sauvegarde et l'accès aux données qualitatives des sciences sociales dans six pays européens. Deux initiatives sont présentées comme d'éventuels modèles pour les chercheurs français : Qualidata (Grande-Bretagne) et SIDOS (Suisse).

Qualidata[24] en Grande-Bretagne a été créé en 1994. Il est implanté à Colchester dans le département de sociologie de l'Université d'Essex. Cette initiative s'est inscrite dans un contexte universitaire largement sensibilisé à la préservation des données orales notamment par l'enquête menée par Paul Thomson à l'initiative de l'ESRC (Conseil de la Recherche Économique et Sociale du Royaume Uni). Elle pourrait servir d'exemple. Le service est très sélectif pour les fonds produits après 1995 (parmi les critères : thèmes bien identifiés, corpus documentés, documents sonores numérisés et en excellent état et dont les caractéristiques sont juridiquement établies).

Le service retient des critères utiles dans la perspective d'une analyse secondaire à venir. Il est intéressant de noter l'investissement de la structure dans la formation des chercheurs futurs producteurs de données.

Elle peut ainsi constituer un moyen de mieux maitriser la recherche dans certains secteurs en évitant les redondances.

---

[23] *op.cit.*

[24] Qualidata, UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK. www.qualidata.ac.uk.Voir aussi l'Annexe 3 (Cribier, 2005).

Le SIDOS, Service suisse d'information et d'archivage des données pour les sciences sociales, créé en 1992 par l'Académie suisse des sciences humaines et sociales, constitue lui aussi une sorte d'agence de gestion des données qualitatives ou quantitatives produites par les chercheurs[25].

Le SIDOS considère le producteur de données comme un auteur et tout travail d'archivage comme un travail d'édition des données et de la documentation.

L'archivage est orienté vers l'échange de données entre chercheurs. Il constitue un instrument d'enrichissement de l'activité scientifique, à la condition que ces données soient ensuite convenablement diffusées.

La mise en place de tels réseaux aurait un intérêt incontestable pour la recherche. Nous ne sommes pas persuadés que le statut patrimonial et la pérennité de ces collections orales seraient mieux garantis.

QUELLES SOURCES ORALES POUR DEMAIN ?

Depuis le début de l'enregistrement numérique, la question de la pérennité à long terme fait encore problème, notamment de par l'obsolescence rapide des standards et de la compatibilité des systèmes. Mais la cohérence future des collections est bousculée par les modalités d'archivage des données. Verser ses fonds représente pour le chercheur un véritable *travail d'édition* des corpus et de leur documentation, afin de toujours rendre accessibles des documents compréhensifs et cohérents. Ce travail devrait toujours être réalisé par le chercheur. Quand en prendra-t-il le temps ? Quelle image de ses travaux souhaitera-t-il verser ? Quelle forme conserver ? Quel intérêt pour le chercheur de demain ? Il n'y a pas de réponse unique.

Le chercheur qui souhaite utiliser des corpus oraux créés par d'autres a besoin d'une médiation, c'est-à-dire d'une documentation qui décrit les variables mais aussi la collecte des données et le contexte du projet.

Dans ce dernier cas, le chercheur producteur n'est pas le mieux à même de décrire ses données, dont l'usage sera fait par des personnes non familières de son domaine. Il appartient au professionnel du traitement documentaire, bibliothécaire, documentaliste, archiviste de décrire *grâce à des outils normalisés et compréhensibles par tous les corpus destinés à des tiers.*

La description trop précise, *témoignages* « *ultérieurs* », « *rétrospectifs* » « *récits de vie a posteriori* », fondée sur la notion de temporalité, certes utile pour les besoins d'analyse du chercheur, n'est pas opérante pour la gestion de ces collections au sein d'une institution de conservation. Ces critères font certes partie de la description objective

[25] Voir enquête réalisée par F. Cribier & E. Feller , *op. cit.* Annexe 3 : 14-20.

du document oral, mais il n'appartient pas à l'institution de les *classer dans des catégories trop étroites qui procèdent déjà de l'analyse et limitent la liberté des futurs usagers en contraignant leur point de vue.*

En résumé, le producteur de corpus est certainement le seul à pouvoir documenter ses corpus oraux. Leur utilisation par des tiers ne pourra se faire que si le signalement est rédigé par des professionnels de la documentation.

4.3.2 VERS LA RECONNAISSANCE D'UN STATUT DU PATRIMOINE ORAL

L'avenir des sources orales n'est pas une question exclusivement juridique. Cette dimension peut être résolue par des solutions contractuelles pragmatiques. Ce *Guide* n'a d'autre ambition que de le montrer.

Mais le véritable enjeu de la question des sources orales est d'ordre culturel et politique. Leur reconnaissance nécessite à la fois l'élaboration de critères de tri exigeants, sans lesquels aucun patrimoine digne de ce nom ne peut exister, et dans le même temps une prise de conscience de la société, qui consiste à conférer, à ces documents produits scientifiquement, *un statut d'objet du patrimoine.*

Leur intégration au sein du dispositif qui régit les objets du patrimoine sera alors chose naturelle.

La France, il faut le noter, accuse à l'égard du patrimoine immatériel un retard singulier.

5

6 ANNEXES

Fiches juridiques

L'Œuvre Orale
Les œuvres protégées
Données personnelles et anonymisation
Le droit de citation
Le Consentement
Exemples d'autorisations
Bases de données, objet d'un droit « sui generis »
Responsable du traitement
Le Patrimoine immatériel et l'UNESCO

Fiches techniques

Prise de son et enregistrement sur le terrain
Supports pour enregistrer et archiver le son
Supports pour enregistrer et archiver la vidéo
Codages et formats

Institutions

Bibliothèque nationale de France
Les Archives : législation
Musées de France : législation
Inathèque de France

Travaux

Programme « ARCHIVAGE » du LACITO
CLAPI
PFC
DELIC
ESLO
Inventaire des corpus

L'ŒUVRE ORALE

L'élaboration des corpus oraux peut se faire à partir de différents types de productions orales. Un certain nombre de ces productions peuvent être des œuvres de l'esprit protégées par le droit d'auteur. Nous présenterons ici quelques-uns des exemples les plus courants en analysant leur statut juridique et les conséquences qui en découlent.

L'INTERVIEW, RECITS DE VIE

Le fait de répondre à des questions ou de livrer un témoignage, voire d'enregistrer une personne en situation peut constituer un élément important dans la réalisation d'un certain nombre de corpus oraux. Du fait de cette importance, il est nécessaire de cerner le cadre juridique dans lequel s'inscrivent la réalisation et l'exploitation des interviews. Des jurisprudences récentes concernant des personnes publiques permettent aujourd'hui de faire le point et d'inciter à une certaine prudence dans la prise en compte des droits des interviewés et des obligations du responsable du corpus. Les conséquences du non-respect de ces obligations peuvent être un obstacle à l'utilisation, voire à la diffusion ou la publication du corpus. Il faut cependant garder à l'esprit que tout enregistrement ne constitue pas nécessairement une interview. Lorsqu'il n'y a pas communication d'une pensée, l'enregistrement ne sera pas considéré comme une interview. Ainsi, lorsqu'il s'agit de faire réciter une liste de nombres, lire un texte imposé ou répondre à des questions qui n'impliquent aucun élément personnel (ex. le temps qu'il fait), le régime de l'interview ne trouve pas à s'appliquer.

Premier principe :

Toute interview ne peut se faire qu'avec l'ACCORD de la personne interrogée. La jurisprudence reste constante sur ce point. Par exemple, un journaliste s'est vu condamner car il avait enregistré une personne clandestinement en dissimulant le magnétophone sous sa serviette[26]. Dans le cas où il n'est pas souhaitable pour des raisons liées à l'intérêt scientifique, l'accord peut être demandé après l'interview. Par exemple, pour des enregistrements en situation, pour ne pas influencer le sujet, on peut lui demander son accord lorsqu'il sort du lieu choisi. En principe, il est préférable pour des raisons de garantie des droits des personnes que cet accord soit formalisé dans un écrit. Mais ce n'est pas toujours possible ni souhaitable compte tenu du contexte. Dans tous les cas, il faut d'une façon ou d'une autre garder une trace de cet accord (un enregistrement oral, un écrit…). Voir fiche *Consentement*.

Deuxième principe :

Une fois l'enregistrement terminé, si nous sommes en présence d'une œuvre de l'esprit, ce qui sera bien souvent le cas, il faut déterminer quels sont les droits de l'interviewé. Est-il auteur et peut-il bénéficier à ce titre des droits correspondants (voir fiche *Œuvres protégées*). Partage-t-il ces droits avec l'intervieweur ?

C'est en fonction de l'apport de l'un et de l'autre que l'on pourra déterminer si l'on est ou non en présence d'une œuvre et qui en est le (ou les) titulaire(s) : banalités (ex. : questions courantes, propos impersonnels) ou originalité (travail créatif de clarification et structuration des questions ou des réponses)

[26] Cour d'appel de Versailles première chambre, 29 novembre 2001.

des questions posées et des propos enregistrés. La jurisprudence a récemment admis que lors des longues séries d'interviews filmées du président Mitterrand, ses propos constituaient en eux-mêmes une œuvre de l'esprit protégée par le droit d'auteur[27]. Le cas échéant, l'intervieweur et l'interviewé peuvent être co-auteurs du résultat de l'interview (voir fiche *Œuvres protégées*)[28]. Quand le corpus oral est aussi une œuvre audiovisuelle, la loi énumère les différentes catégories de personnes pouvant revendiquer, aussi, la qualité de co-auteurs (voir fiche *Œuvres protégées*)[29]. Quand cela est possible, et si la personnalité de l'interviewé ou le contexte le justifient, un contrat peut être une bonne solution pour ménager les droits de chacun.

Il reste un dernier aspect à ne pas négliger, celui de la relecture. En effet, toute personne ayant un droit d'accès et de rectification sur ses propos (voir fiche *Consentement*), mieux vaut offrir à la personne interrogée la possibilité de corriger ses dires.

LES DISCOURS POLITIQUES

Les actualités télévisées utilisent abondamment les discours politiques, d'où la naissance d'une confusion. Les actes officiels, tels les textes de lois, les décrets ou encore les rapports gouvernementaux sont dépourvus de toute protection et appartiennent au domaine public. Les discours et autres allocutions sont d'une nature bien différente ; ils ne fixent pas de règle pour les citoyens. On ne peut donc pas les définir comme des actes officiels, même s'ils sont prononcés dans le cadre de fonctions officielles ; ce sont des créations intellectuelles qui appartiennent à leurs auteurs. Ainsi le texte d'une loi votée au Parlement relève du domaine public, mais le discours de présentation bénéficie de la protection du droit d'auteur[30]. L'auteur jouit donc de droits moraux et patrimoniaux.

Il existe deux exceptions au droit de reproduction :

- L'article L122-5 3° CPI autorise une reproduction, même intégrale, pour la presse ou la télévision à des fins d'informations d'actualité[31]. Cela signifie que ce type de reproduction est réservé aux organes de presse ou de télévision (pas aux chercheurs). Il faut en plus qu'il y ait un rapport avec l'actualité immédiate. L'autorisation est donc limitée dans le temps. La notion d'actualité s'apprécie en fonction de la périodicité du média (un hebdomadaire aura plus de temps qu'un journal télévisé quotidien). L'œuvre devra avoir un rapport direct avec l'actualité immédiate. Lors d'une prise d'otages, la photo du preneur d'otage fait partie de l'actualité, pas celle de gens passant à proximité du lieu. Dès que le temps de

[27] Note P. Sirinelli à propos de TGI Paris, 16 Septembre 2003, C.Sosnowski c/ France 2 *et al. Propriétés Intellectuelles*, 9, Octobre 2003, : 380-382. On notera que le Président Mitterrand avait aménagé par contrat les droits qu'il revendiquait sur le résultat de l'interview.

[28] CA Paris, 4e chambre, 5 décembre 1997, SA Les Belles Lettres et autres c/ Éditions Albin Michel et autres, *Recueil Dalloz* 1999, 65.

[29] Voir TGI Paris 16 septembre 2003, précité.

[30] TGI Paris 3e chambre, 25 Octobre 1995, François Mitterrand c/ ID Éditions et autres in *Revue Internationale du Droit d'Auteur*, 167, Juillet 1995 : 294-298.

[31] Art. cité en note 3.

l'actualité passe, toute utilisation d'une œuvre de l'esprit ne peut se faire sans autorisation; et versement de droits le cas échéant.

- Le droit de citation permet de s'exonérer du droit de reproduction à condition de citer ses sources, que l'extrait soit court et qu'il n'ait qu'un rôle d'illustration. Si l'on soustrait les citations, l'œuvre doit conserver son caractère original. (voir fiche *Citation*).

LES EMISSIONS DE RADIO ET DE TELEVISION

Ce sont des œuvres qui sont destinées au public, ce qui ne veut pas dire qu'elles soient libres de droits. Une chaine, qu'elle soit radiophonique ou de télévision, reçoit la qualification juridique d'entreprise de communication audiovisuelle selon l'article L216-1 al. 2 du CPI.

L'entreprise dispose de droits voisins sur les programmes (voir fiche *Œuvres protégées)*. Toute utilisation dans un cadre professionnel devra donc se faire uniquement après l'obtention d'un accord de l'entreprise. Un corpus oral ne pourra utiliser des émissions de radio ou de télévision que si les droits ont été acquis préalablement. Le plus souvent, c'est le producteur qui doit donner son autorisation. Il faut donc s'adresser à la station de radio, à la chaine de télévision ou encore à l'INA.

Le corpus ne va pas, la plupart du temps, utiliser l'intégralité d'une œuvre radiophonique ou audiovisuelle ; il se contentera d'extraits. Le droit de citation existe à condition d'indiquer clairement la source et que l'œuvre à laquelle on l'intègre ait un caractère « critique, polémique, pédagogique, scientifique ou d'information... » selon l'article L213 3° CPI. L'indication de la source doit apparaitre explicitement. Un extrait d'émission de télévision peut se faire si le logo de la chaine est visible[32]. Toutes les règles énoncées dans la fiche *Citation* restent valables.

[32] TGI Paris 31 mars 1999.

LES ŒUVRES PROTEGEES

LES ŒUVRES PROTEGEES PAR LE DROIT D'AUTEUR ET LES DROITS VOISINS

Un corpus oral s'obtient au terme de nombreuses étapes. Une ou plusieurs personnes réaliseront une collecte, puis d'autres (ou les mêmes) effectueront la transcription ; et ainsi de suite jusqu'à l'obtention du corpus. La question qui se pose alors est de savoir qui sont le ou les auteurs ? Quels droits cela leur confère-t-il ?

La réponse s'obtient en s'intéressant d'abord aux notions d'auteur et d'œuvre. Il faut ensuite décrire les droits liés à la qualité d'auteur. Enfin, nous ne devons pas oublier les auxiliaires de la création que sont les producteurs et les artistes-interprètes.

1. L'AUTEUR ET SON ŒUVRE

Aucune définition légale ne vient déterminer ce qu'est un auteur si ce n'est son lien avec sa création : « la qualité d'auteur appartient, sauf cas contraire, à celui ou à ceux sous le nom de qui l'œuvre est divulguée »[33]. C'est donc l'œuvre qui est l'élément déterminant du droit d'auteur.

L'ŒUVRE

La qualification d'œuvre ne prend en compte ni le genre, ni la forme d'expression, ni le mérite, ni la destination.

Sans liste limitative, peuvent être œuvres de l'esprit des œuvres écrites, orales, audiovisuelles, produits de la recherche littéraire, artistique, musicale, quel que soit le mode d'exploitation (vidéo, cinéma...), les destinataires ou les sens auxquels elles s'adressent (l'ouïe, la vue, l'odorat...).

Toutefois, pour qu'une œuvre soit protégée par le droit d'auteur, il faut :

- qu'elle soit concrétisée dans une forme ;
- qu'elle soit originale.

Le droit d'auteur ne protège que la forme de l'œuvre et non les idées contenues dans celle-ci. Un cours d'université donné oralement, une conférence sont protégés. En revanche, les hypothèses scientifiques traitées différemment ne sont pas protégées. À propos des logiciels, seule la forme du programme, c'est-à-dire l'enchainement des instructions, peut être protégée.

L'originalité : l'œuvre ne doit pas avoir été copiée ou être le résultat d'un plagiat. Elle doit présenter une certaine créativité marquant soit l'empreinte de la personnalité de son auteur, soit un apport intellectuel, un effort personnalisé allant au-delà de la simple logique automatique.

Une œuvre est protégée dès sa création, aucune formalité n'est nécessaire pour jouir des droits qui lui sont attachés. Cette protection dure toute la vie du créateur et se prolonge après sa mort, sans limite pour le droit moral durant soixante-dix ans pour les droits pécuniaires.

[33] Art. 113-1 du CPI.

UNICITE OU PLURALITE D'AUTEURS

Suivant les conditions dans lesquelles l'œuvre a été créée, il peut y avoir un ou plusieurs auteurs :

- **Œuvre dérivée ou seconde :** lorsqu'une œuvre existante est incorporée, sans l'intervention de son auteur, à une autre œuvre. Une nouvelle œuvre est ainsi créée, donnant des droits à son auteur, malgré son originalité relative. Mais celui dont l'œuvre a été utilisée conserve sur celle-ci toutes ses prérogatives. Ainsi, une traduction ou une adaptation d'un texte ouvrent des droits à celui qui la réalise ainsi qu'à l'auteur du texte traduit ou adapté.
- **Œuvre collective** : quand une personne physique ou morale dirige ou coordonne plusieurs contributions qui se retrouvent fondues pour former une œuvre unique[34]. Il n'est alors pas possible de distinguer les apports de chacun. L'auteur sera donc celui qui en aura eu l'initiative ou qui aura joué le rôle de coordination. Il est primordial d'étudier le processus qui a conduit à sa création.
- **Œuvre de collaboration :** œuvre créée par plusieurs auteurs. La contribution de chaque auteur est clairement identifiable[35]. Les œuvres cinématographiques ou audiovisuelles comportent le plus souvent une pluralité d'auteurs. Sont présumés coauteurs, sauf preuve contraire, le réalisateur, le scénariste, le dialoguiste, le compositeur, l'adaptateur (ainsi que l'auteur de l'œuvre adaptée). Pourra également être considérée comme coauteur toute autre personne qui fera la preuve d'un apport original (création spéciale ou indépendance vis-à-vis du réalisateur).

Un corpus oral peut être qualifié, selon les cas, d'œuvre dérivée, d'œuvre collective ou d'œuvre de collaboration.

La frontière entre ces notions n'est pas toujours évidente à définir. Il faut pourtant bien prendre garde à ne pas vouloir les faire entrer trop hâtivement dans une catégorie afin d'éviter des procédures devant les juridictions civiles. Voyons maintenant quels sont les droits que la création ouvre à son ou ses auteurs.

2. LES DROITS DE L'AUTEUR

Le droit d'auteur se décompose en deux prérogatives bien distinctes :

LE DROIT MORAL

Le droit moral est *perpétuel*. Il survit à la mort de l'auteur. Il est *inaliénable*, l'auteur ne peut y renoncer, ni même le transmettre. Il est juridiquement *imprescriptible*. L'exercice de ce droit est *absolu*; l'auteur peut en user à discrétion sauf lorsqu'il y a plusieurs auteurs ou abus pour nuire à autrui, ou encore détournement.

La première prérogative est le droit de *divulgation* : l'auteur décide seul de rendre ou non publique sa création. Il peut choisir ensuite d'y inscrire son nom ou ses qualités ; c'est le droit de *paternité*. S'il désire rester anonyme, cela n'autorise en rien l'appropriation par d'autres personnes. L'œuvre ne peut être altérée sans un accord exprès de l'auteur; il faut respecter son *intégrité*. Enfin, un auteur peut exprimer des regrets face à sa création et demander le retrait

[34] Art. L113-2 al. 3 CPI.
[35] Art. L113-2 al. 1 CPI.

de son œuvre. Il fait alors valoir son *droit de retrait* ou *repentir*. L'auteur doit alors indemniser l'éditeur de l'œuvre.

À la mort de l'auteur les droits de retrait et de repentir disparaissent. Restent le droit au nom et à la paternité. À la mort de l'auteur, le droit de divulgation se voit placé sous le contrôle du juge afin de faire respecter la volonté du créateur.

LES DROITS PATRIMONIAUX

Ils sont *limités* dans le temps : soixante-dix ans après la mort du créateur. Ils peuvent faire l'objet de contrats car ils sont *cessibles* ; à condition toutefois de ne pas violer la liberté de décision de l'auteur.

L'auteur dispose du droit exclusif d'autoriser ou d'interdire la reproduction ou la représentation de son œuvre.

Le droit de reproduction se définit comme « la fixation matérielle de l'œuvre par tous procédés qui permettent de la communiquer au public de manière indirecte ». Ainsi toute copie de l'œuvre, quel que soit le support, ne peut être faite sans l'autorisation de l'auteur.

Le droit de représentation s'entend, lui, comme la mise en contact direct de l'œuvre avec le public. La communication de l'œuvre au public est une représentation, le public pouvant être les chercheurs du laboratoire, comme toute autre personne destinataire (amphi d'étudiants, colloque...).

Ces droits exclusifs souffrent des exceptions étroitement limitées par le législateur. Elles sont principalement :

- l'exception de copie privée qui se limite à l'usage personnel et privé du copiste ;
- le droit de citation (voir fiche *Droit de citation*).

Il existe un droit à la copie privée, mais il se limite à l'usage personnel et privé du copiste. Les usages professionnels ou collectifs (mêmes internes) sont donc proscrits.

Le droit de suite se trouve un peu en marge, car il s'applique pour les œuvres d'art plastique. Lors de la revente de l'œuvre, 3 % du prix va au créateur s'il est vivant, sinon à ses héritiers.

3. LES DROITS VOISINS

Souvent, l'artiste ne peut développer seul sa création. Pour la communiquer au public, il a besoin d'auxiliaires pour assurer l'effort financier, le producteur de phonogrammes ou de vidéogrammes, ou pour donner vie à son œuvre, l'artiste-interprète. Des droits leur sont accordés pour une durée de cinquante ans, à partir de la première communication.

DROITS DU PRODUCTEUR

Le producteur a le droit d'autoriser ou d'interdire la reproduction directe ou indirecte de l'œuvre qu'il a produite. Il peut en contrôler aussi la forme de communication (diffusion, vente, échange, location).

Dans le cas d'un phonogramme produit à des fins commerciales, il ne peut s'opposer à la communication directe de l'œuvre dans un lieu public (sauf sonorisation de spectacle) ou sur une radio. En contrepartie, il reçoit une rémunération fixée par la loi. Ce sont généralement les sociétés de gestion

collective (SACEM...) qui collectent les sommes correspondantes auprès des utilisateurs puis qui les reversent.

DROITS DE L'ARTISTE-INTERPRETE

L'artiste-interprète a droit au respect de son nom, de sa qualité, de son interprétation[36].

Il a le droit d'autoriser ou d'interdire la fixation, la reproduction et la communication au public de sa prestation. Il est fréquent qu'il apporte ses droits à une société de gestion collective (ADAMI, SPEDIDAM...) qui délivre l'accord pour toute utilisation de la prestation et qui perçoit les sommes dues à l'artiste en cotrepartie.

[36] Art. L212-2 CPI.

DONNEES PERSONNELLES ET ANONYMISATION

La collecte d'informations personnelles devenant chaque jour plus simple, la vie privée doit être vraiment protégée. Les méthodes de profil de personnalité sont utilisées par bon nombre d'entreprises pour mieux connaître, soit leurs clients, soit aussi parfois leurs employés. La source d'informations disponible sur Internet ne cesse de croitre. L'anonymisation constitue un mode important de protection. Les corpus oraux sont souvent grands consommateurs de données ; il faut donc concilier les impératifs légaux avec les exigences de la recherche.

Les textes fondamentaux encadrant la protection des données personnelles ne définissent pas la notion d'anonymisation car toute donnée peut devenir sensible, selon la finalité de son traitement. Il nous faut donc reprendre les concepts-clés des différents textes afin d'avoir une vision plus précise de l'anonymisation.

LES DONNEES A CARACTERE PERSONNEL

La loi du 6 janvier 1978 s'articule autour de la notion de donnée nominative. La Convention 108 du Conseil de l'Europe de 1981 lui préfère celle de données personnelles et la directive 95/46/CE choisit l'expression « *donnée à caractère personnel* ». Le projet de transposition de la directive reprend d'ailleurs ce dernier terme. Au-delà des différences de termes, il faut souligner la généralité des formules. Ainsi le considérant 26 de la directive dispose :

> *« …pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne ».*

Les traitements ne sont à déclarer à la CNIL que lorsqu'ils utilisent des données à caractère personnel. Mais il n'y a *pas de critère précis* qui puisse être dégagé. Les décisions de la CNIL s'appuient sur le type d'informations utilisées, mais surtout sur la logique qui va présider à leur traitement. L'article 2 de la directive définit la notion de données à caractère personnel et donne quelques exemples, sans pour autant délimiter le champ d'application. Pourtant, à la lecture du rapport du Sénat sur le projet de transposition, le lecteur pourrait avoir l'impression inverse, puisque l'auteur écrit :

> *« La directive prévoit des critères permettant de délimiter le champ des données concernant une personne identifiable… ».*

Toutes les données, quelle que soit leur forme ou leur support, peuvent tomber dans le champ d'application du cadre légal « informatique et libertés ». Le considérant 14 mentionne explicitement le son, l'image et la voix. On peut aujourd'hui y ajouter, en raison des progrès de la génétique l'ADN, ou l'iris de l'œil.

L'identification d'une personne peut se faire de manière directe ou indirecte. Le caractère personnel d'une donnée dépend des moyens de tri, de rapprochement qui pourraient être mis en œuvre. Cela conduit donc à une évolution constante du champ des données personnelles, la technique mettant à la disposition du plus grand nombre des outils de plus en plus performants.

Certains auteurs avancent qu'il suffit qu'il y ait une *probabilité suffisante de rapprochement* avec une personne pour qu'une donnée acquière un caractère personnel indirect. Les analyses ne sont pas toujours explicites mais il n'est pas possible de négliger cet argument[37]. Dans le domaine statistique, la CNIL a imposé des seuils au-delà desquels des rapprochements d'agrégats de données – pourtant individuellement anonymes – sont interdits.

Le caractère personnel d'une information dépend de l'objet qu'elle décrit, du contexte dont elle provient, mais aussi de la personne qui la reçoit. Pour pouvoir identifier un individu ou un groupe, nous avons besoin des informations, mais nous n'y arriverons pas sans un élément de connaissance propre qui déclenchera le mécanisme d'association. *Le récepteur constitue un élément important* de l'équation. Le phénomène croissant de marchandisation des données attire les convoitises de toutes sortes d'individus, pas toujours à même de les exploiter en dehors d'une transaction commerciale. Il nous semble donc important d'inclure dans une réflexion les capacités des personnels qui traitent les données. Pour un informaticien qui gère un système traitant des données génétiques, les données auxquelles il accède, sont-elles pour lui des données à caractère personnel ?

LA NOTION DE TRAITEMENT DES DONNEES A CARACTERE PERSONNEL

L'article 2b de la directive européenne définit ainsi le traitement des données :

« toute opération ou ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé, et notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction. »

La longueur de la définition illustre avant tout le champ des possibilités ouvert par l'outil informatique, tout en allant plus loin car c'est à tous les types de traitements qu'il est ici fait allusion. La distinction entre traitement automatisé et non automatisé n'a plus cours. Il en va de même pour la notion de fichier, le législateur européen met sur le même plan les fichiers informatiques et manuels : *il suffit que les données soient organisées suivant une structure définie*[38].

Le traitement n'implique pas forcément une manipulation du fichier, un simple stockage suffit à le faire entrer dans le champ d'application. La difficulté vient une nouvelle fois de la très grande portée de la définition.

Nous retrouvons l'interrogation que nous avions soulevée : la difficulté qu'éprouve le législateur à éviter la systématisation trop grande des notions qu'il veut défendre. Une donnée n'acquiert pas forcément de caractère personnel par sa nature, tout dépend de celui qui l'utilise. Nombreux sont ceux qui soulignent à juste titre qu'il est impossible d'admettre des définitions trop vastes, sous peine de les rendre inapplicables, et que mieux vaudrait se concentrer sur des types définis qui remettent en cause des valeurs fondamentales[39].

---

[37] Lamy *Droit de l'informatique et des Réseaux* 508 et suivants.
[38] Considérant 27, de la directive 95/46
[39] Frayssinnet, J. dir. (2001) *Droit de l'Informatique et de l'Internet*, Paris, PUF, §127 : 85-86.

L'ANONYMISATION

L'*anonymisation* sert à qualifier l'opération par laquelle se trouve supprimé dans un ensemble de données, recueilli auprès d'un individu ou d'un groupe, tout élément qui permettrait l'identification de ces derniers. Le nom propre n'est donc pas le seul élément qu'il faille prendre en compte. On pourrait parler de « *dépersonnalisation* » des données comme dans la loi fédérale allemande sur la protection des données à caractère personnel du 23 mai 2001.

Lorsqu'on réfléchit à l'anonymisation, il convient de connaitre les éléments à traiter, mais aussi les opérations que vont subir les données.

La difficulté soulevée par la question de l'anonymisation apparait plus clairement après avoir rapidement passé en revue quelques principes clés des lois encadrant l'informatique. Il ne s'agit pas tant de savoir comment effectuer le travail d'anonymisation, mais plutôt de définir quelles données doivent être anonymisées, pour qui, et dans quel contexte.

L'exemple des pratiques autorisées pour la recherche médicale fournit quelques pistes de réflexion.

La condition première est d'avoir un responsable ainsi qu'une ou plusieurs finalités précises. Les données transmises ne peuvent l'être que si elles sont destinées à des membres du même milieu professionnel, soumis aux mêmes règles déontologiques. Le plus souvent celui qui reçoit les données doit pouvoir travailler sur des données anonymes.

L'anonymat doit être irréversible, et la CNIL est seule habilitée à autoriser la fourniture de données non anonymisées après examen du projet scientifique. La publication ou un autre mode d'exploitation des résultats ne peut donner lieu *en aucune manière* à une possible identification des personnes.

L'obligation d'obtenir un consentement préalable peut être levée si retrouver les personnes concernées s'avère difficile. S'il n'a pas été recueilli immédiatement, le consentement doit être obtenu avant le premier traitement. Les demandes de dérogation sont du ressort exclusif de la CNIL.

Voici les divers procédés qu'elle préconise :

- Le codage : les données personnelles sont cryptées par des clés cryptographiques générées par des logiciels informatiques.
- Les bases de données séparées : Le réseau SESAME-VITALE utilise bien entendu le cryptage des données. Mais pour garantir un maximum de confidentialité, deux types de bases de données ont été distingués. Des bases primaires contiennent toutes les données mais elles ne sont pas connectées au réseau, elles servent de sécurité et disposent de tables de concordance pour lever l'anonymat après autorisation. D'autres bases de données assurent le fonctionnement quotidien du réseau, mais seules les données nécessaires sont présentes.

Une autre voie existe : Les limitations techniques. La loi québécoise « concernant le cadre juridique des technologies de l'information » propose de protéger l'anonymat non pas en modifiant les données, mais en limitant les possibilités de recherche, voire en les adaptant à la personne qui consulte la base selon des critères bien précis (sa profession, une autorisation, sa présence dans le fichier, etc.)

Cette dernière perspective offre pour la constitution et l'exploitation de corpus oraux la possibilité de faire coïncider les obligations légales avec les nécessités

du travail de recherche. *Toute donnée étant potentiellement sensible, une anonymisation systématique s'avère de plus en plus complexe ; elle peut même mettre en danger l'intérêt de certaines recherches.* En effet, des détails concernant les personnes comme par exemple le nom, ou le lieu d'habitation peuvent constituer un élément important du corpus, et des résultats qui peuvent en être obtenus. C'est pourquoi la possibilité de ménager des niveaux d'accès selon des critères stricts (ex : chercheur ou non, présence d'autorisation, but de la consultation, etc.) semble une alternative efficace.

Il existe d'autres procédés à inventer. En effet, l'article 11-2 de la nouvelle loi ouvre la possibilité de faire certifier des techniques nouvelles par la CNIL. Ce n'est pas au chercheur de présenter son procédé, mais à l'institution à laquelle il appartient.

Il faut bien sûr que le type de données collectées ait fait l'objet d'une réflexion quant à son intérêt pour l'étude entreprise ; sous peine de mettre à mal les garanties mises en place et au risque de ne pas obtenir l'accord des autorités compétentes.

Pour le droit à l'image se référer à :

Pierrat, E. (2002) *Reproduction interdite : le droit à l'image expliqué aux professionnels et à ceux qui souhaitent se protéger*, Paris, Maxima Laurent du Mesnil.

Isgour, M. & Vinçotte, B. (1998) *Le droit à l'image,* Bruxelles, Larcier.

Serna, M. (1997) *L'image des personnes physiques et des biens*, Paris, Economica.

Bécourt, D. (2004) *Image et vie privée*, Paris, L'Harmattan.

Bloch, P. dir. (2002) *Image et Droit,* Paris, L'Harmattan.

LE DROIT DE CITATION

Citer des œuvres est un acte important qui s'impose particulièrement dans tout travail scientifique et la constitution et l'utilisation de corpus n'échappent pas à la mise en œuvre du droit de citation. Même si le droit d'auteur prend en compte la citation, l'interprétation qui est faite de ce droit suscite bien souvent des interrogations. Beaucoup d'idées reçues entourent le droit de citation ; comme par exemple l'existence d'un pourcentage défini entre l'extrait choisi et l'œuvre dont il est extrait. Il convient donc de traiter ici le cadre général du droit de citation et les applications particulières à certaines catégories d'œuvres ou certains modes de citation.

LE CADRE GENERAL DU DROIT DE CITATION

Le droit de citation est une exception au droit de reproduction. En effet, tout ou partie d'une œuvre originale ne peut être reproduit par quelque moyen que ce soit sans autorisation de l'auteur (voir fiche *Œuvres protégées*). Ainsi l'article L122-5 3° permet les analyses et courtes citations. Trois conditions doivent alors être observées :

- **La citation doit être justifiée** : il faut un but (critique, scientifique…) et elle doit être incorporée à un développement lié à ce but (démonstration, exposé). Sinon on en revient au cas du recueil qui constitue une œuvre en lui-même ; sauf pour ce qui appartient au domaine public.
- **La citation doit être courte** : les extraits ne peuvent reprendre l'essentiel de l'œuvre dont ils sont issus. L'œuvre qui incorpore des citations doit pouvoir « survivre » à leur suppression. L'appréciation de la brièveté est fonction du rapport entre les citations et l'œuvre citante dans laquelle elles sont incorporées.
- **La citation doit respecter le droit moral de l'auteur** : cela signifie la mention explicite de l'auteur de l'œuvre citée (*droit de paternité*) ; mais aussi la préservation de l'intégrité de l'œuvre, tant dans la forme que dans l'esprit. Enfin, si une œuvre n'a pas été divulguée, la citation est interdite.

Toutes les règles qui viennent d'être énoncées correspondent parfaitement aux œuvres écrites. Il faut aussi s'intéresser aux questions particulières des autres types d'œuvres et voir dans quelles conditions le code de la propriété intellectuelle prévoit des aménagements.

LES CAS PARTICULIERS

- **Les œuvres graphiques ou appartenant aux arts plastiques** : « une jurisprudence constante exclut la reproduction intégrale d'une œuvre au titre de la citation. Quant à la reproduction partielle de l'œuvre (partie du dessin, tableau ou photo) elle porte atteinte à l'intégrité de l'œuvre et ne peut se faire sans l'autorisation de l'auteur. En conséquence, l'exception (le droit) de citation ne peut être invoquée. »
- **Les œuvres musicales** : il n'y a pas d'exclusion de principe de la citation dans le domaine musical. Les règles encadrant leur citation sont les mêmes que pour les autres types d'œuvres.

- **Les bases de données** : pour le droit, une base de données constitue un objet spécifique. Elle bénéficie d'ailleurs d'un droit spécifique (voir fiche *Base de données*). Concernant la citation, la jurisprudence lui a aussi reconnu un régime spécifique. Depuis l'arrêt *Microfor* de 1988, il est admis qu'une base de données peut être constituée exclusivement d'extraits d'œuvres sans qu'il y ait d'autres apports. Dans ce seul cas, les deux exigences posées dans le cadre général (citation justifiée, brièveté) disparaissent. Toutefois subsiste l'obligation de mentionner de façon explicite l'auteur de l'œuvre et l'origine de celle-ci.
- **Le lien hypertexte** : l'usage du lien hypertexte constitue la base de la navigation Internet. Insérer un lien vers un autre document peut s'assimiler à une citation si les trois règles de bases sont respectées, et surtout si l'objet cité n'est pas illicite ou si sa reprise n'est pas interdite par son auteur. La fourniture de lien vers un objet contrefait devient, par exemple, de la complicité. Si la liberté de citation n'apparait pas clairement, il est là aussi préférable d'entrer en contact avec l'auteur.

LE CONSENTEMENT

Le consentement de la personne concernée, dans la plupart des situations de traitement de données personnelles, constitue le plus souvent une manière d'assurer sa protection. Toutefois, il n'est pas toujours possible d'obtenir ce consentement lors de la collecte des données. De plus, il est important d'indiquer que l'obtention du consentement n'exonère pas le responsable du traitement de ses obligations à l'égard des personnes concernées (voir fiche *Responsable du traitement*).

LE PRINCIPE DU CONSENTEMENT

Le consentement doit être éclairé. Pour ce faire, le responsable du traitement doit, en principe, procéder ou faire procéder à une *information préalable* de la personne avant de recueillir son consentement. Le nouvel article 32 de la loi de 1978 énonce clairement les informations qui doivent être fournies :

- l'identité du responsable du traitement et, le cas échéant, celle de son représentant ;
- la finalité poursuivie par le traitement auquel les données seront soumises ;
- le caractère obligatoire ou facultatif des réponses ;
- les conséquences éventuelles, à cet égard, d'un défaut de réponse ;
- les destinataires ou catégories de destinataires des données ;
- l'existence d'un droit d'accès, de rectification voire d'opposition à la collecte ;
- les transferts de données à caractère personnel envisagés à destination d'un État non membre de la Communauté Européenne.

En principe, le consentement doit être exprès. La personne affirme clairement qu'elle accepte que les données personnelles la concernant fassent l'objet d'un traitement. Même si le législateur ne l'impose pas, le consentement écrit est considéré comme une bonne pratique. Dans des situations particulières, d'autres formes peuvent être choisies, l'important est de pouvoir faire la preuve de la volonté de la personne concernée (ex. : enregistrement d'un accord verbal).

Qui consent ? Toute personne physique. Lorsqu'il s'agit d'une personne déclarée incapable (qu'elle soit majeure ou mineure), l'information doit parvenir au représentant légal. Pour le cas d'enfants, il faut l'autorisation des parents ou du dépositaire de l'autorité parentale.

Il est vrai que l'écrit tient une place importante dans le formalisme du consentement. Quand, dans certaines situations, cela se révèle impossible à mettre en œuvre sous peine de fausser les résultats de la recherche engagée, des solutions alternatives existent.

Même si elle s'est exprimée de façon expresse, toute personne ayant consenti au recueil des données et à leur traitement dispose, pour des motifs légitimes, d'un droit de rétractation ou d'opposition à ce que des données à caractère personnel la concernant continuent à faire l'objet d'un traitement.

DES ALTERNATIVES AU CONSENTEMENT

Le consentement de la personne concernée ne suffit pas toujours à la protéger contre des utilisations abusives des données qui la concernent. Il faut aussi veiller à ne pas faire de l'exigence du consentement une obligation administrative qui ferait perdre de vue ce qui compte : la protection de la personne. C'est pourquoi les textes récents ont mis en place des alternatives au consentement ou même des garanties spécifiques pour le traitement de certains types de données (voir fiche *Responsable du traitement*). Parmi les alternatives pouvant s'appliquer au traitement des corpus oraux on citera :

> *«la réalisation de l'intérêt légitime poursuivi par le responsable du traitement ou par le destinataire, sous réserve de ne pas méconnaître l'intérêt ou les droits et libertés fondamentaux de la personne concernée. » (art. 7, 5°).*

Toutefois, l'intérêt légitime de la recherche et du traitement aura à être démontré. En pratique l'application de cette alternative a pour conséquence de dispenser le responsable du traitement de demander un consentement exprès à condition, bien entendu, de ne pas méconnaitre l'intérêt ou les droits et libertés fondamentaux de la personne concernée.

EXEMPLES D'AUTORISATIONS

Voici à titre *d'exemples* deux formulaires d'autorisation, tirés d'expériences de chercheurs en France (ICAR) et aux États-Unis (Ervin-Tripp).

Les formulaires d'autorisation diffèrent notamment en ce qui concerne la présentation des options proposées à l'informateur. Celles-ci concernent essentiellement les contextes d'exploitation des données et la forme des données montrables en public (différents supports, anonymisés ou non).

Ces exemples ne peuvent constituer un modèle à reprendre tel quel, seul un travail sur l'explicitation de la démarche et les objectifs d'exploitation de chaque projet permet de construire un formulaire d'autorisation adéquat.

De manière générale, il est fortement conseillé d'adapter le formulaire aux visées particulières de l'enquête, notamment aux objets que l'on désire recueillir et étudier, aux types d'acteurs sociaux concernées par l'enquête et aux conditions d'exploitation et de diffusion du corpus.

1) Exemple de formulaire type de demande d'autorisation mis au point au laboratoire ICAR (UMR 5191 CNRS)

[papier avec entête officiel]

Autorisation

pour l'enregistrement audio/vidéo et l'exploitation des données enregistrées

Présentation de l'enquête

[Peut se présenter sous forme de brochure séparée laissée aux enquêtés]

[Préciser l'institution d'où émane la recherche, la personne qui dirige/qui est responsable du projet, les chercheurs concernés sur le terrain.

Préciser le thème général du projet, le type de corpus qui est recueilli de manière générale, le type d'enregistrement qui est recueilli auprès de ces informateurs en particulier, son traitement et utilisation prévus.

Souligner les apports du projet, valoriser la collaboration de l'informateur, expliciter les bénéfices éventuels qu'il peut en tirer et les risques éventuels qu'il peut courir.]

Ces recherches ne sont possibles que grâce au consentement des personnes qui acceptent d'être enregistrées. Nous vous demandons par conséquent votre autorisation à procéder aux enregistrements.

Autorisation (biffer les paragraphes qui ne conviennent pas)

Je soussigné(e)

- autorise par la présente NN et NN à enregistrer en audio/vidéo le [préciser le type d'événement enregistré].

- autorise l'utilisation de ces données, sous leur forme enregistrée aussi bien que sous leur forme transcrite et anonymisée (cf. *infra*) :

a) à des fins de recherche scientifique (mémoires ou thèses, articles scientifiques, exposés à des congrès, séminaires) ;

b) à des fins d'enseignement universitaire (cours et séminaires donnés à des étudiants avancés, à partir du niveau maitrise, en sciences du langage et en sciences sociales) ;

c) pour une diffusion large dans la communauté des chercheurs, sous la forme d'éventuels échanges et prêts de corpus à des chercheurs, moyennant la signature d'une convention de recherche ;

d) pour une diffusion sur un site Internet dédié à la recherche.

- prends acte que pour toutes ces utilisations scientifiques les données ainsi enregistrées seront anonymisées, cela signifie :

a) que les transcriptions de ces données utiliseront des pseudonymes et remplaceront toute information pouvant porter à l'identification des participants ;

b) que les bandes audio qui seront présentées à des conférences ou des cours (généralement sous forme de très courts extraits ne dépassant pas la minute) seront « bipées » lors de la mention d'un nom, d'une adresse ou d'un numéro de téléphone identifiables (qui seront donc remplacés par un « bruit » qui les effacera) ;

c) en revanche, pour des raisons techniques, le projet ne peut pas s'engager à anonymiser les images vidéo mais s'engage à ne pas diffuser d'extraits compromettant les personnes filmées.

- souhaite que la contrainte supplémentaire suivante soit respectée :

Lieu et date:

Signature :

[Prévoir un double ou un autre document qui sera laissé à la personne, comportant une adresse de contact et éventuellement une adresse Internet où consulter les résultats publiés du projet].

2) Exemple de demande d'autorisation de Susan Ervin-Tripp, Univ. de Californie, Berkeley

Researcher name

LETTER OF CONSENT

PHOTOGRAPHIC, AUDIO, AND/OR VIDEO RECORDS RELEASE CONSENT FORM

As part of this project we have made a photographic, audio, and/or video recording of you while you participated in the research. We would like you to indicate below what uses of these records you are willing to consent to. This is completely up to you. We will only use the records in ways that you agree to. In any use of these records, names will not be identified.

1. The records can be studied by the research team for use in the research project.

Photo ..... Audio ..... Video .....

[Please use initials]

2. The records can be shown to subjects in other experiments.

Photo ..... Audio ..... Video .....

[Please use initials]

3. The records can be used for scientific publications.
Photo ..... Audio ..... Video .....
[Please use initials]
4. The written transcript can be kept in an archive for other researchers.
Photo ..... Audio ..... Video .....
[Please use initials]
5. The records can be used by other researchers.
Photo ..... Audio ..... Video .....
[Please use initials]
6. The records can be shown at meetings of scientists interested in the study of.....
Photo ..... Audio ..... Video .....
[Please use initials]
7. The records can be shown in classrooms to students.
Photo ..... Audio ..... Video .....
[Please use initials]
8. The records can be shown in public presentations to nonscientific groups.
Photo ..... Audio ..... Video .....
[Please use initials]
9. The records can be used on television and radio.
Photo ..... Audio ..... Video .....
[Please use initials]

I have read the above description and give my consent for the use of the records as indicated above.

Date .....
Signature .....
Signature of Guardian, if Applicable .....
Native language(s) .....
Where native language learned (city or region) .....
Languages used on the tape
Where language(s) used on tape were learned .....
Age at which each language used on tape was learned .....
Education Occupation .....
Name Age ..... Sex.....

BASES DE DONNÉES, OBJET D'UN DROIT « SUI GENERIS »

Avec l'utilisation des nouvelles technologies, la création d'un corpus oral aboutit, le plus souvent, à la création d'une « base de données » renfermant toutes les informations recueillies, transformées et produites au cours des différentes phases du travail de recherche. Peu d'activités n'utilisent pas de bases de données.

L'alinéa 2 de l'article L112-3 du Code de Propriété Intellectuelle (CPI) définit la notion de base de données :

> *« On entend par base de données un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique et individuellement accessibles par des moyens électroniques ou par tout autre moyen ».*

Ainsi, pour constituer une base de données, il faut des données mais aussi une structure pour les ordonner. Les données ont un statut propre, indépendant de celui de la base de données. Elles peuvent être des œuvres protégées par le droit d'auteur, elles peuvent être des données personnelles. En fonction de chaque statut, elles relèveront des cadres juridiques correspondants.

La base de données, quant à elle, relève des œuvres susceptibles d'être protégées par le droit d'auteur. A côté de ce régime de protection dont bénéficie l'auteur de la base de données, l'investisseur (ou producteur) de la base dispose d'un droit dit « sui generis » qui le protège contre des exploitations et utilisations abusives des données de la base. Ainsi, l'auteur et le producteur bénéficient de droits différents et ces régimes ne s'excluent pas. Ces deux régimes de protection peuvent bénéficier à une seule et même personne qui serait à la fois auteur et producteur. Les bénéficiaires peuvent être aussi des personnes différentes.

LE DROIT « SUI GENERIS » DES BASES DE DONNEES, PROTECTION DE L'INVESTISSEMENT

Qui bénéficie de ce droit ? En quoi consiste le régime de protection mis en place ?

LA TITULARITE DES DROITS

Dans le Code de la propriété intellectuelle le bénéficiaire du « droit sui generis »[40] est appelé le « producteur ». Selon l'article L341-1 du CPI le producteur est celui « *qui prend l'initiative et le risque des investissements correspondants… »*. Il ne s'agit donc *pas* forcément du *concepteur* de la base, mais plutôt de celui, ou de ceux qui ont pris *l'initiative*, les décisions clés et qui, avec les investissements requis, ont permis la réalisation de la base.

L'article L341 du CPI précise qu'il faut un *investissement* financier, matériel ou humain *substantiel* pour bénéficier du droit sui generis. Ce dernier critère permet de distinguer une base de données d'une simple compilation (simple reprise d'éléments contenus dans une autre base). Ainsi, une simple reprise des annuaires de France-Télécom ne peut faire l'objet d'une protection par ce régime particulier. Il faut aussi mettre en évidence le volume de travail, le cout des interventions lors de la création de la base comme dans les mises à jour.

[40] Art. L 341-1, 342-2, 342-1 à 342-5 du CPI.

La jurisprudence refuse la protection du droit « sui generis » à une revue d'annonces légales au motif que la revue ne justifie pas d'investissements substantiels dans leur obtention et dans leur traitement.

Les « producteurs » de corpus oraux ne pourront bénéficier de la protection du droit « sui generis » que s'ils font la démonstration de la réalité de l'investissement substantiel réalisé. Dans ce cas, ils bénéficieront des droits correspondants.

LES DROITS DU PRODUCTEUR

Le droit principal concerne l'*interdiction* pour l'utilisateur légitime d'extraire de façon substantielle les données contenues dans la base. Par *substantielle*, le code vise autant la qualité que la quantité des données extraites (art. L 341-1 et 341-2 du CPI). Ainsi, le terme *substantiel* s'apprécie au cas par cas. Des informations rares – bien qu'en petit nombre – peuvent tomber sous le coup de l'interdiction. Une extraction, même non substantielle, peut se voir interdite si elle a un caractère *répété* ou *systématique*. Le but avoué est ici d'empêcher le pillage de bases de données par des concurrents mal intentionnés.

Il faut garder à l'esprit que ces possibilités d'interdiction sont un droit et non pas une obligation. Le producteur peut autoriser – moyennant contrepartie – ces extractions. Le droit sui generis du producteur prend effet à compter de l'achèvement de la fabrication de la base de données et expire quinze ans après le 1er janvier de l'année civile qui suit celle de cet achèvement (art. L 342-5 du CPI). Si à la fin de la période, il y a un nouvel investissement substantiel, la protection se voit renouvelée. Les atteintes au droit sui generis sont sanctionnées pénalement. En cas d'infraction, les peines sont de deux ans de prison et de 150 000 euros d'amende. Pour les personnes morales, l'emprisonnement se transforme en interdiction d'exercice.

LA COEXISTENCE AVEC LES AUTRES DROITS ET LES LIMITES

Les droits du producteur trouvent leur première limite :

- Dans le droit des utilisateurs légitimes à extraire ou réutiliser une partie non substantielle du contenu de la base.
- Le législateur a prévu, dans certains cas[41], un statut dérogatoire à l'extraction à des fins privées d'une partie de la base, que cette extraction soit qualitativement ou quantitativement substantielle. Il faut noter que la directive européenne de 1996 prévoyait, derrière cette dérogation, le cas de l'enseignement ou de la recherche scientifique à but non commercial. Cette dernière disposition n'a pas été retenue par le législateur. En fait, les deux conditions imposées limitent fortement ce statut dérogatoire : seuls sont visés les contenus de bases de données non électroniques, et cette extraction doit respecter les droits d'auteurs ou les droits voisins sur les œuvres ou éléments incorporés dans la base.
- Le producteur doit aussi, le plus souvent, assurer l'*accès* aux informations tout en garantissant leur *licéité* ainsi que leur *fiabilité*. Les informations doivent donc être mises à jour et avoir été obtenues de manière légale (les droits éventuels liés à ces informations ne peuvent pas être ignorées).

[41] Art. 342-3 2°.

RESPONSABLE DU TRAITEMENT

Tout traitement de données doit avoir un responsable. Sa mission est d'éviter ou de circonvenir les risques inhérents à la gestion et à l'utilisation des données recueillies. La loi définit qui est le responsable et lui fixe donc des obligations.

L'article 3-1 de la loi du 6 janvier 1978 dispose :

> *« Le responsable d'un traitement de données à caractère personnel est, sauf désignation expresse par les dispositions législatives ou réglementaires relatives à ce traitement, la personne, l'autorité publique, le service ou l'organisme qui détermine ses finalités et ses moyens. »*

Qui est-il ? C'est une personne physique qui détient le pouvoir de décision sur les finalités et les moyens à mettre en œuvre.

LES PRINCIPES GENERAUX

Le responsable du traitement se doit donc de veiller à la qualité des données, au respect des finalités indiquées, au respect du principe de licéité et aux conditions de conservation.

LA QUALITE DES DONNEES

Pour pouvoir être traitées, les données doivent avoir été recueillies selon un ensemble de principes qui garantissent la protection des personnes. Une donnée doit être :

- **adéquate, pertinente et non excessive**. Toutes les données faisant l'objet d'un traitement doivent être en lien avec la finalité poursuivie. La CNIL se montre particulièrement vigilante sur ce point. L'INSEE s'est souvent vu refuser ses questionnaires ou être obligé de les revoir car les données collectées étaient jugées trop nombreuses ou inutiles par rapport à la finalité annoncée. Plus on acquiert de données sur un même individu, plus le risque est grand de voir ce traitement surveillé étroitement, voire refusé par la CNIL.
- **exacte**. Les données doivent être exactes et mises à jour. Ceci renvoie au droit d'accès, d'opposition et de rectification ouvert à chaque personne concernée par le traitement.

LE RESPECT DES FINALITES INDIQUEES

La finalité du traitement sert à justifier celui-ci. Il s'agit de répondre à la question du but de la mise en œuvre d'un ou plusieurs traitements. De même il peut y avoir plusieurs finalités. Le responsable du traitement - selon la définition - détermine la finalité. Il doit donc annoncer *par avance* le but du traitement qu'il s'apprête à réaliser. Le responsable du traitement qui justifierait après le traitement les finalités poursuivies manquerait à ses obligations légales et serait susceptible d'être sanctionné pénalement.

LE RESPECT DU PRINCIPE DE LICEITE

Toute donnée collectée doit avoir été recueillie loyalement. Cela suppose une *information préalable*, une *demande écrite de consentement* (voir fiche *Consentement*), l'explication quant à la *finalité du traitement*, le *nom du responsable du traitement*, ainsi que les *conséquences en cas de refus*. La notion de loyauté renvoie au contexte dans lequel s'est effectuée la collecte.

LES CONDITIONS DE CONSERVATION

Confidentialité et conservation limitées dans le temps. Il appartient au responsable du traitement d'assurer la confidentialité et le respect des règles de communication de ces données hors du cadre défini pendant toute la durée de conservation de ces données. La durée de conservation varie selon la finalité du traitement effectué. Les données peuvent être conservées au-delà de la durée prévue initialement quand elles présentent un intérêt pour des fins historiques, statistiques ou scientifiques (art. 36). Cette possibilité de conservation n'entraine pas la possibilité d'exploitation ni de diffusion, les conditions d'accès aux données étant réglées par la loi sur les archives.

LES FORMALITES PREALABLES : DECLARATION ET AUTORISATION

DECLARATION

En principe, pour les catégories les plus courantes de traitement dont la mise en œuvre n'est pas susceptible de porter atteinte à la vie privée ou aux libertés, la formalité requise est une déclaration à la CNIL. Dans les situations répétitives, cette déclaration peut être simplifiée. La CNIL délivre sans délai un récépissé et, dès réception de celui-ci, le responsable peut mettre en œuvre le traitement. La déclaration comporte l'engagement que ce traitement satisfait aux exigences de la loi (respect du principe de licéité, voir *supra*).

AUTORISATION

Si les corpus contiennent des données sensibles, le responsable du traitement devra demander une autorisation à la CNIL qui dispose d'un délai de deux mois pour se prononcer (délai susceptible d'être renouvelé une fois). L'absence de réponse de la CNIL dans les délais doit être interprétée comme un rejet de la demande d'autorisation.

LE PATRIMOINE IMMATERIEL ET L'UNESCO

Ces fiches ont pour objet de présenter les questions posées par les documents (déclarations, conventions, autres…) de l'UNESCO. En effet, la constitution de corpus oraux et la recherche sur les langues participent à la protection du patrimoine culturel de l'humanité qui est une des grandes missions de cette organisation internationale. La constitution de grands corpus oraux peut servir de documentation générale sur des langues et contribuer à l'élaboration des outils de diffusion de langues peu (ou pas) écrites[42].

L'UNESCO s'est intéressée aux différentes formes de régulation des recherches portant sur le patrimoine culturel : régulation éthique, déontologique et juridique.

ÉTHIQUE ET DEONTOLOGIE DE LA RECHERCHE

Dans la recommandation de 1989 (article E.g) la communauté scientifique internationale est encouragée « *à se doter d'une éthique appropriée à l'approche et au respect des cultures traditionnelles* ». Le chercheur doit être animé d'un souci de respect à l'égard de ses collaborateurs occasionnels (sujets de recherche), dont il devra rechercher la confiance, et à l'égard des traditions de ceux qu'il étudie.

Par ailleurs, l'exigence pour les chercheurs de se doter d'un code de déontologie a été posée notamment lors de la Conférence de Washington sur l'évaluation globale de la recommandation de 1989 relative à la sauvegarde de la culture traditionnelle et du folklore[43]. La cinquième recommandation faite à l'UNESCO l'invite à : *« encourager les groupements internationaux (chercheurs, professionnels de la culture…) à créer et à adopter des codes déontologiques qui assurent que des démarches appropriées et respectueuses sont suivies vis-à-vis de la culture traditionnelle et du folklore.* »

LES RECOMMANDATIONS RELATIVES A L'ENCADREMENT JURIDIQUE DES TRAVAUX SUR LES LANGUES

Invitant le lecteur à s'intéresser aux cadres normatifs dans lesquels les recherches sur les langues sont menées et souhaitant ouvrir sur des exemples de législations nationales, ces fiches proposent des pistes de réflexion sur les questions posées par la recherche sur les langues en voie de disparition. Bien entendu, il n'était pas possible de traiter de toutes les situations locales et ces exemples sont là pour inciter les chercheurs à se renseigner sur les droits nationaux susceptibles de s'appliquer dans les pays où sont menées les recherches.

Une des questions qui se posent lorsqu'on appréhende les travaux sur la langue est celle de la détermination de son statut juridique en tant qu'élément du patrimoine culturel : fait-elle partie du domaine public et, partant, libre de tout droit ou, au contraire, s'agit-il d'un bien appropriable, et, dans ce cas, quelles sont les conséquences pour le travail des chercheurs ?

Ces différentes interrogations nous conduisent à analyser dans les documents de l'UNESCO et les législations de quelques pays africains ce qui est dit sur le statut des langues (fiche I). De ce statut découleront les conditions dans

---

[42] Voir *supra* 2-1.

[43] Conférence précitée, voir note 26.
http://www.folklife.si.edu/resources/Unesco/actionplan_french.htm.

lesquelles peuvent être menées les recherches et la constitution des corpus oraux (fiche II). La question d'un droit à la protection de la vie privée à travers les recommandations de l'UNESCO et dans quelques pays africains complètera cette présentation (fiche III).

LA DETERMINATION DU STATUT JURIDIQUE DE LA LANGUE

L'UNESCO et la reconnaissance explicite de la langue comme élément du patrimoine culturel immatériel de l'Humanité

La caractéristique des textes de l'UNESCO est de cerner la langue d'un point de vue collectif en ceci qu'elle fait partie du patrimoine culturel de l'Humanité. On peut à cet égard se référer à la description qui en est donnée par sa section du patrimoine immatériel. Il en ressort que :

> *« Les langues sont la plus grande création et expression du génie de l'humain. Elles ne sont pas uniquement des outils complexes et raffinés de communication. Elles constituent un élément déterminant de l'identité humaine et, à ce titre, représentent un noyau primordial du patrimoine culturel de l'Humanité. »*

Trois textes principaux contribuent à cette appréhension de la langue comme élément du patrimoine culturel :

> *Tout d'abord, la Recommandation sur la sauvegarde de la culture traditionnelle et populaire*[44]*. Selon l'article A de cette Recommandation, la culture traditionnelle et populaire est : « l'ensemble des créations émanant d'une communauté culturelle fondées sur la tradition, exprimées par un groupe ou des individus et reconnues comme répondant aux attentes de la communauté en tant qu'expression de l'identité culturelle et sociale de celle-ci, les normes et les valeurs se transmettant oralement, par imitation ou par d'autres manières. Ses formes comprennent, entre autres, la langue, la littérature, la musique, la danse, les jeux, la mythologie, les rites, les coutumes, l'artisanat, l'architecture et d'autres arts. »*
>
> *Ensuite, la Déclaration universelle de l'UNESCO sur la diversité culturelle*[45]*, qui considère que la culture est : « l'ensemble des traits distinctifs spirituels et matériels, intellectuels et affectifs qui caractérisent une société ou un groupe social et qu'elle englobe, en outre les arts et les lettres, les modes de vie, les façons de vivre ensemble, les systèmes de valeurs, les traditions et les croyances. » Le cinquième point de son Plan d'Action vise à : « sauvegarder le patrimoine linguistique de l'Humanité et soutenir l'expression, la création et la diffusion dans le plus grand nombre possible de langues. »*

[44] Recommandation sur la sauvegarde de la culture traditionnelle et populaire, 15 novembre 1989.

[45] Déclaration universelle sur la diversité culturelle, 17 octobre 2001.

> *Enfin, la Convention pour la sauvegarde du patrimoine culturel immatériel*[46]*, qui est la consécration du patrimoine immatériel. Par patrimoine culturel immatériel, il faut entendre selon l'article 2 de ladite Convention : « les pratiques, représentations, expressions, connaissances et savoir-faires - ainsi que les instruments, objets, artefacts et espaces culturels qui leur sont associés - que les communautés, les groupes et, le cas échéant, les individus reconnaissent comme faisant partie de leur patrimoine culturel. Ce patrimoine culturel, transmis de génération en génération, est recréé en permanence par les communautés et les groupes en fonction de leur milieu, de leur interaction avec la nature et de leur histoire, et leur procure un sentiment d'identité et de continuité, contribuant ainsi à promouvoir le respect de la diversité culturelle et de la créativité humaine. »*

Il se manifeste dans les domaines suivants :

- les traditions et expressions orales, y compris la langue comme vecteur du patrimoine culturel immatériel ;
- les arts du spectacle ;
- les pratiques sociales, rituels et événements festifs ;
- les connaissances et pratiques concernant la nature et l'univers;
- les savoir-faires liés à l'artisanat traditionnel.

LANGUES ET FOLKLORE

Les composantes sus-citées du patrimoine culturel (y compris la langue) qui sont l'objet des recherches en linguistique, reçoivent aussi la qualification de folklore ou d'expressions du folklore. C'est d'ailleurs cette terminologie qui a été proposée comme modèle pour les législations nationales. Pour s'en convaincre, on peut se référer aux Dispositions types de législation nationale sur la protection des expressions du folklore contre leur exploitation illicite et autres actions dommageables[47]. Mais cette catégorie ne doit pas faire illusion car elle semble se fondre - même si les logiques ne sont pas les mêmes - dans les catégories précitées. Ainsi, un auteur[48], pour définir le folklore, se réfère entièrement à la convention de 1989 sur la culture traditionnelle et populaire. Par expressions du folklore, il faut entendre, au sens de l'article 2 des Dispositions types :

> *« les productions se composant d'éléments caractéristiques du patrimoine artistique traditionnel développé et perpétué par une communauté ou par des individus reconnus comme répondant aux aspirations artistiques traditionnelles de cette communauté, en particulier*
>
> *les expressions verbales telles que les contes populaires, la poésie populaire et les énigmes ;*

[46] Convention pour la sauvegarde du patrimoine culturel immatériel, 17 octobre 2003.

[47] Élaborées conjointement par l'Unesco et l'OMPI et approuvées par un comité d'experts gouvernementaux en 1985.

[48] Folarin, S. (2002) « Conservation, préservation et protection juridique du folklore en Afrique », *Bulletin du droit d'auteur*, XXXII/4 :41.

> *les expressions musicales telles que les chansons et la musique instrumentale populaires ;*
>
> *les expressions corporelles telles que les danses et les spectacles populaires ainsi que les expressions artistiques des rituels. »*

C'est de la conjonction des trois textes principaux mentionnés ci-dessus que résulte la définition de la langue comme élément du patrimoine culturel immatériel de l'Humanité. Mais que recouvre cette notion ?

LA NOTION DE PATRIMOINE CULTUREL IMMATERIEL DE L'HUMANITE

Nous ne pouvons ici développer les notions de patrimoine culturel et de patrimoine culturel immatériel. Nous renvoyons, en note, à des études sur le sujet[49].

Toutefois nous dirons quelques mots sur la notion de patrimoine culturel de l'Humanité. Sans entrer dans le débat sur ce que recouvre la notion d'Humanité, nous mentionnerons l'importance d'une coopération entre tous les acteurs concernés pour la sauvegarde du patrimoine :

> *« …les États parties reconnaissent que la sauvegarde du patrimoine culturel immatériel est dans l'intérêt général de l'Humanité et s'engagent, à cette fin, à coopérer aux niveaux bilatéral, sous-régional, régional et international.»*[50]

On remarque aussi, par ailleurs, que les expressions du folklore, qui font pourtant partie du patrimoine culturel de l'Humanité, sont l'objet d'une « appropriation nationale ». Les langues ne peuvent-elles pas, alors, relever à la fois du patrimoine de l'Humanité et par ailleurs être prises en compte dans la protection des patrimoines culturels nationaux ?

LES ÉTATS AFRICAINS ET LA RECONNAISSANCE IMPLICITE DE LA LANGUE COMME ELEMENT DU PATRIMOINE CULTUREL NATIONAL A TRAVERS LA NOTION DE FOLKLORE

On retrouve la notion de patrimoine culturel immatériel au niveau régional dans l'Accord de Bangui sur la propriété intellectuelle[51]. Comme dans les législations nationales, la mention se fait par référence au folklore.

L'OAPI ET LE PATRIMOINE CULTUREL IMMATERIEL

Le titre II de l'annexe VII de l'Accord de Bangui porte sur la protection et la promotion du patrimoine culturel. Aux termes de l'article 67 al. 1, « ***le patrimoine culturel est l'ensemble des productions humaines matérielles et immatérielles caractéristiques d'un peuple dans le temps et dans l'espace***.»

Cette définition a le mérite d'être complète en ce qu'elle cerne les deux aspects du patrimoine culturel, l'aspect matériel et l'aspect immatériel qui inclut indubitablement la langue. Même si la langue n'est pas expressément visée dans le texte, on peut la retrouver implicitement dans la référence à l'oralité.

---

[49] Cornu, M. (2003) « Droit des biens culturels et des archives » : 1, sur le patrimoine culturel immatériel.
http://www.unesco.org/culture/heritage/intangible/html_fr/index_fr.shtml

[50] Convention sur le patrimoine immatériel, article 19.2.

[51] Elaboré par l'OAPI (Organisation Africaine de la Propriété Intellectuelle), adopté en 1977 et révisé le 22 février 1999.

En effet, le folklore qui est une production du patrimoine culturel comprend : *« les productions littéraires de tout genre et de toute catégorie orale ou écrite, contes, légendes, proverbes, épopées, gestes, mythes, devinettes*[52].»

LES LEGISLATIONS NATIONALES

La langue, et partant le patrimoine immatériel, ne sont pas expressément et directement envisagés par la plupart des législations africaines[53].Toutefois, ils y sont intégrés par référence aux lois sur le droit d'auteur, qui soulignent dans une formule assez générique que : *« le folklore appartient à titre originaire au patrimoine culturel national.*[54] »

Les langues ne sont prises en compte que comme des « vecteurs » de contenus spécialisés (chant, conte, proverbe, énigme, etc.) qui sont des expressions du folklore et sont à ce titre protégés au titre du droit d'auteur.

En conséquence, lorsque les corpus réunissent ces contenus du folklore, il faut demander les autorisations requises aux titulaires des droits d'auteur.

LA RECHERCHE SUR LES LANGUES SUSCEPTIBLES DE RENTRER DANS LE CHAMP DE PROTECTION MIS EN PLACE PAR L'UNESCO

La recherche scientifique est un moyen de sauvegarde de la langue. Ainsi, l'identification de la culture traditionnelle et populaire participe à la protection du patrimoine culturel immatériel permettant de *« créer des systèmes d'identification et d'enregistrement (collecte, indexation, transcription) ou développer des systèmes déjà existants au moyen de guides, guides de collecte, de catalogues types, etc., eu égard à la nécessité de coordonner les systèmes de classement utilisés par différentes institutions*[55] ».

LA SAUVEGARDE DU PATRIMOINE IMMATERIEL

L'UNESCO fait ainsi figure de pionnière en matière de promotion de la recherche pour la sauvegarde du patrimoine immatériel. En plus de son programme spécifique aux langues en danger, elle déclare, dans le cadre de la musique traditionnelle du monde, devoir adapter son action aux besoins des chercheurs. Elle soutient la préservation des archives sonores et des centres de documentation et elle encourage la recherche.

L'article D.e de la Recommandation de 1989 demande aux États de « *promouvoir la recherche scientifique se rapportant à la préservation de la culture traditionnelle et populaire* ». De même l'article 13.c de la Convention de 2003 dispose que les États doivent s'efforcer « *d'encourager les études scientifiques, techniques et artistiques ainsi que les méthodologies de recherche pour une sauvegarde efficace du patrimoine culturel immatériel, en particulier le patrimoine culturel immatériel en danger* ». Cette promotion passe par l'adoption de diverses mesures juridiques, techniques, administratives et financières appropriées.

---

[52] Article 68 al. 2.a.

[53] Sauf la loi ivoirienne du 28 juillet 1987, qui vise dans son article 3 les œuvres du folklore.

[54] Par exemple l'article 8 de la loi ivoirienne du 25 juillet 1996 ; l'article 5-1 de la loi camerounaise du 19 décembre 2000 ; l'article 82 de la loi tchadienne du 2 mai 2003. La loi de la République démocratique du Congo, dans son article 6, dispose que le folklore est « l'un des éléments fondamentaux du patrimoine culturel traditionnel ».

[55] Recommandation sur la culture traditionnelle et populaire, article B.b.

Orchestrées par l'UNESCO, diverses actions visent plus spécialement à la sauvegarde du patrimoine linguistique de l'Humanité. Elles concernent, en général, l'encouragement de la diversité linguistique – dans le respect de la langue maternelle – la promotion de la diversité linguistique[56] avec un programme spécifique concernant les langues en péril. « *Une langue est en péril lorsque ses locuteurs commencent à la délaisser, réservant son utilisation à des contextes de moins en moins nombreux, et ne la transmettant plus de génération en génération*[57]. »

Les chiffres et estimations sur ce phénomène sont quelque peu alarmants. Selon l'UNESCO, il existe environ 6 000 langues dans le monde, dont plus de la moitié sont en danger. On estime par ailleurs qu'une langue meurt tous les quinze jours. Il est estimé de même « *qu'une langue ne peut survivre qu'à la condition de compter au moins 100 000 locuteurs. Or, sur les 6 700 langues actuelles, la moitié compte moins de 10 000 locuteurs*[58]. » Tel est le cas de la langue zapara qui est parlée couramment par seulement cinq personnes[59].

Les projets de recherche concernant les langues en péril doivent viser deux objectifs principaux.

Il s'agit d'une part d' «épargner l'humanité de la perte qui peut découler de l'extinction d'une langue en danger. Cette approche met en exergue l'archivage, qui consistera à collecter autant de documentation que possible sur la langue, et à procéder à une description linguistique aussi complète que le temps le permettra. »

Il s'agit, d'autre part, de « revitaliser la langue en encourageant son utilisation dans l'alphabétisation et dans l'enseignement primaire[60].»

En suivant cette logique des textes de l'UNESCO, la finalité de la recherche doit s'inscrire dans une optique de sauvegarde du patrimoine culturel, y compris les langues. À cet égard, il est important de rappeler que la recommandation de 1989 porte sur la sauvegarde de la culture traditionnelle et populaire. Il en va de même pour la convention de 2003 concernant le patrimoine culturel immatériel. Cet objectif affiché de sauvegarde du patrimoine immatériel induit diverses mesures. Ainsi, selon l'article 2 al 3 de la convention de 2003 : « *on entend par 'sauvegarde' les mesures visant à assurer la viabilité du patrimoine culturel immatériel, y compris l'identification, la documentation, la recherche, la préservation, la protection, la promotion, la mise en valeur, la transmission, essentiellement par l'éducation formelle et non formelle, ainsi que la revitalisation des différents aspects de ce patrimoine.* »

LE DROIT D'ACCES AUX MATERIAUX DU PATRIMOINE IMMATERIEL

La promotion du patrimoine impose de laisser aux chercheurs un droit d'accès à ce patrimoine. Ce droit d'accès implique pour le chercheur de pouvoir collecter les matériaux nécessaires à son travail mais aussi de pouvoir travailler à partir de ceux déjà collectés et conservés. Il ressort de la lecture de

---

[56] Respectivement les points 5 ; 6 et 10 du plan d'action de la Déclaration universelle sur la diversité culturelle.

[57] Unesco, Kit d'information de la Section du patrimoine immatériel, Division du patrimoine culturel, Secteur de la culture.

[58] La mort des langues, http://www.tlfq.ulaval.ca/axl/Langues/2vital_mortdeslangues.htm.

[59] Sources, No 106 – juillet août 2001, p.6, Unesco.

[60] http://www.acalan.org, mission et vision de l'Acalan.

l'article 13d.ii de la Convention de 2003 que l'État doit adopter des mesures juridiques, techniques, administratives et financières appropriées visant à « *garantir l'accès au patrimoine* ». Dans la Recommandation de 1989, ce droit est affirmé de façon particulière. En effet, cette Recommandation fait de la recherche, la finalité de la conservation des matériaux du patrimoine culturel. Autrement dit, les matériaux doivent être conservés pour que des recherches puissent être menées. Aux termes de l'article B : « *la conservation concerne la documentation relative aux traditions se rapportant à la culture traditionnelle et populaire et a pour objectif, en cas de non-utilisation ou d'évolution de ces traditions, que les chercheurs et les porteurs de la tradition puissent disposer de données leur permettant de comprendre le processus de changement de la tradition* »

LE DROIT A LA PROTECTION DES MATERIAUX COLLECTES

La Recommandation de 1989 (article F) prévoit, de façon explicite, la protection des « intérêts des collecteurs en veillant à ce que les matériaux recueillis soient conservés dans les archives, en bon état et de manière rationnelle ». Les chercheurs peuvent se prévaloir d'un tel droit dans la mesure où leur travail consiste en partie dans la collecte de données sur le terrain. Toutefois, l'exercice de ce droit implique aussi que les données recueillies soient déposées aux archives prévues à cet effet.

L'ARCHIVAGE DES MATERIAUX COLLECTES

L'accès aux matériaux du patrimoine immatériel peut se faire de plusieurs façons. Il peut se réaliser dans le cadre d'institutions de documentation (ou musées) que les États devront tout d'abord mettre en place. Dans le texte de l'article 13b de la Convention de 2003, il est question d'organismes compétents pour la sauvegarde du patrimoine culturel immatériel sur un territoire. La recommandation de 1989 établit une liste de mesures (sept au total[61]). Parmi lesquelles, on peut citer :

- la mise en place de services nationaux d'archives où les matériaux de la culture traditionnelle et populaire collectés puissent être stockés dans des conditions appropriées ;
- la création de musées ;
- l'octroi de moyens en vue d'établir des copies d'archives et de travail de tous les matériaux de la culture traditionnelle et populaire.

Une fois de tels organismes institués, l'État doit en faciliter l'accès et les matériaux doivent être réellement mis à disposition.

La responsabilité du service d'archives peut se trouver engagée. La Recommandation de 1989 (article F.iv) invite à « *reconnaître que les services d'archives ont la responsabilité de veiller à l'utilisation des matériaux recueillis* ».

LE DROIT A L'INFORMATION

Ce droit peut être perçu comme une modalité particulière du droit d'accès comme il appert dans la diffusion de l'information. Ainsi, l'Académie Africaine des Langues (Acalan)[62] se fixe pour objectif de « *faciliter la documentation et l'échange d'information par la mise en place d'une base de données, la collecte*

[61] Article B alinéas a à g.

[62] http://www.acalan.org, mission et vision de l'Acalan (Académie africaine des langues).

et l'archivage des documents, la publication et consacrer une bonne partie de ses ressources à l'impulsion de la recherche et à la coordination des activités de recherche. » L'UNESCO a recommandé, en 1989, la mise en place d'une unité centrale d'archives aux fins de la prestation de certains services (indexation centrale, *diffusion de l'information* relative aux matériaux de la culture traditionnelle et populaire et aux normes applicables aux activités la concernant, y compris l'aspect préservation)[63].

LA RETRIBUTION DES PERSONNES SOLLICITEES

En occultant les aspects de propriété intellectuelle qui peuvent résulter de l'utilisation des expressions du folklore ou de toute expression langagière qualifiable d'œuvre de l'esprit, la rétribution se pose pour toutes les personnes auditionnées pour la recherche, surtout lorsqu'on se trouve dans une optique de protocole expérimental. Les textes ne prévoient pas cette éventualité. L'exigence sur ce point est d'ordre éthique. Selon G. Durnon, le montant doit être fixé en commun accord avec les intéressés et doit être équitable et s'il existe des obstacles (milieu traditionnel opposé à une telle pratique), le chercheur peut participer à une œuvre d'intérêt collectif.

LE DROIT DE DIFFUSION DES RESULTATS DE LA RECHERCHE

L'UNESCO ne vise pas directement la diffusion des résultats de la recherche. Cependant, les publications scientifiques participent à la *large diffusion* des éléments constituant le patrimoine qu'encourage l'UNESCO, la seule limite consistant dans le fait *d'éviter, lors de cette diffusion, toute déformation* pouvant porter atteinte à son intégrité.

LE PARTAGE DES RESULTATS DE LA RECHERCHE

Quand il s'agit des travaux de recherche effectués par les spécialistes d'un État membre dans un autre État membre, la Recommandation de l'UNESCO de1989 dispose dans son article Gd que les États devraient : « *garantir aux États membres sur le territoire desquels ont été effectués des travaux de recherches le droit d'obtenir de l'État membre concerné copie de tous documents, enregistrements vidéo, films et autres matériels*. »

L'idée, ici, est de participer, à l'issue de la recherche, à l'enrichissement des archives des institutions locales : « *en fournissant des copies des documents sonores ou audiovisuels recueillis au cours de la recherche*[64]». Ce qui semble recommandé, c'est de faire bénéficier les communautés sollicitées des retombées positives de la recherche. Cela se justifie si on part du postulat que chaque peuple a un droit sur sa propre culture. C'est ce qu'affirme en substance la Recommandation de 1989 dans son article D. On remarque par ailleurs l'existence de fortes revendications émanant des communautés dans lesquelles sont effectuées les recherches. Comme le note un rapport de l'UNESCO[65] : « *speakers increasingly demand control over the terms and*

[63] Article C.b de la Recommandation Unesco de 1989 sur la sauvegarde de la culture traditionnelle et populaire.

[64] Durnon, G. (1981) *op. cit.* :51 ouvrage précité, p.51. Cela peut se faire par le biais d'une « procédure de restitution des biens culturels aux pays d'origine qui est appliquée depuis plusieurs décennies au département d'ethnomusicologie du Musée de l'Homme à Paris. »

[65] Language Vitality and Endangerment,
http://portal/unesco.org/culture/en/ev.php-URL_ID=9105&URL_DO=DO_TOPIC&URL_SECTION=201.html

conditions that govern research; furthermore, they claim rights to the outcomes and future uses of the research". Une prise en compte de ces droits met à la charge du chercheur une obligation d'implication des populations dans la mise en œuvre de la recherche. Aux termes du rapport précité : « *any research in endangered language communities must be reciprocal and collaborative. Reciprocity here entails researchers not only offering their services as a quid pro quo for what they receive from the speech community, but being more actively involved with the community in designing, implement and evaluating their research project*. »

LA PROTECTION DES PERSONNES CONCERNEES PAR LA RECHERCHE

La constitution des corpus oraux donne lieu à la collecte de nombreuses informations sur les personnes : nom, prénom, âge, appartenance ethnique, sexe, statut social, lieu de résidence et de naissance, image et voix (enregistrement audio et vidéo, photographie), etc.

LES PERSONNES CONCERNEES

Ces données concernent diverses personnes. Il pourra s'agir selon les cas d'un chanteur, d'un compositeur, d'un traducteur, d'un interprète, d'un conteur, ou d'un locuteur ordinaire. La Recommandation de 1989 traite de l'informateur qu'il faut protéger en tant que porteur de la tradition. Une autre qualification existe en ce qui concerne les personnes : celle des « trésors humains vivants ». Les « trésors humains vivants » sont « *des personnes qui possèdent à un très haut niveau les connaissances et les savoir-faires nécessaires pour interpréter ou créer des éléments spécifiques du patrimoine culturel immatériel que les États membres ont choisi comme témoignages de leurs traditions culturelles vivantes et du génie créateur des groupes, des communautés et d'individus présents sur leur territoire.* » C'est le cas par exemple des membres de la famille Dökala en Guinée, qui assurent l'enseignement de l'histoire familiale, locale, régionale conformément à l'héritage légué par les anciens. « *Ce sont eux qui détiennent au plus haut niveau les valeurs authentiques de la civilisation mandingue* »[66]. Le système des trésors humains vivants a été adopté en 1993 et est censé être une partie essentielle de la mise en œuvre de la Convention sur le patrimoine immatériel.

LE CHAMP DE LA PROTECTION

Le champ de la protection des personnes concerne, selon l'article F.i de la Recommandation de l'UNESCO de 1989, la vie privée et la confidentialité. Le droit à la confidentialité d'une information dont bénéficie une personne peut être perçu d'un point de vue négatif en tant qu'*obligation de secret* qui pèse sur une autre (généralement un professionnel). Visant à mettre le bénéficiaire à l'abri de divulgations, le droit à la confidentialité se situe dans le prolongement du droit au respect de la vie privée.

En ce qui concerne ce droit, au-delà de la première difficulté à laquelle on se heurte - résidant dans le défaut de définition légale de la notion même de vie privée, en droit français comme dans les législations africaines - une seconde

[66] Namankoumba Kouyaté, Méthodes traditionnelles de transmission de l'oralité : l'exemple du Sosso-Bala. Conférence : « Evaluation globale de la recommandation de 1989 sur la sauvegarde de la culture traditionnelle et populaire ; pleine participation locale et coopération internationale », Washington, 1997
http://www.folklife.si.edu/resources/Unesco/kouyate.htm

difficulté résulte d'une certaine divergence de conception. Cl. Ouoba écrit à ce sujet que *« si l'on estime que le secret et le privé sont indispensables à l'épanouissement individuel de chaque citoyen dans les sociétés occidentales, il serait un peu trop idéaliste de transposer une telle appréciation dans les sociétés où le partage et la communion sont les bases de l'existence même*[67]*.»* Abondant dans le même sens, A. Sow Sidibé ajoute que « *alors que la conception occidentale est fondée sur l'individualisme, celle négro-africaine repose sur des valeurs communautaires qui privilégient le groupe*[68] ».

Le droit à la vie privée est garanti, directement ou indirectement, dans la grande majorité des États africains tant par le biais de l'adhésion à des conventions internationales que par des législations spécifiques[69]. Au titre du droit au respect de la vie privée se trouve aussi consacré un droit à l'image. « *Représentation physique ou morale, elle (l'image) appartient en propre à l'individu, et sa reproduction ou sa divulgation, en somme sa violation, ne peut se faire sans son accord*[70].» Toute reproduction de l'image d'une personne doit se faire avec son consentement, sauf exception (notamment en matière d'information). Ainsi que le recommandait G. Durnon, dans le cadre de la collecte des musiques : *« la prise en vue nécessite l'assentiment des intéressés*[71]. »

UN REGIME PARTICULIER POUR LES DONNEES TRAITEES A DES FINS DE RECHERCHE

Ainsi l'article 68[72], relatif à la recherche scientifique, dispose que l'utilisation des données est soumise à une autorisation préalable du concerné ou de ses ayantdroits et de la commission nationale. Et selon l'article 49 : *« les données à caractère personnel, traitées pour des finalités particulières, peuvent être communiquées en vue d'être traitées une autre fois pour des fins historiques et scientifiques, à condition d'obtenir le consentement de la personne concernée, de ses héritiers ou de son tuteur ainsi que l'autorisation de l'Instance Nationale de Protection des Données à Caractère Personnel.* » Il s'agit ici de la reconnaissance du principe de finalité compatible pour le traitement des données à caractère personnel à des fins historiques et scientifiques lorsque les données avaient été collectées pour une autre finalité.

Même si le contexte et les enjeux ne sont pas les mêmes dans la recherche biomédicale et en recherche linguistique, on peut citer la déclaration du Réseau africain sur l'éthique, le droit et le VIH : *« (…) la recherche doit être effectuée*

---

[67] Ouoba, Cl. (2002) *Le droit à la vie privée au Burkina Faso, Conception, réalité juridique et socioculturelle*, Thèse de l'Université de Grenoble 2 : 50.

[68] Sow Sidibé, A. « Le secret médical aujourd'hui », revue électronique *Afrilex* 2 : 25. http://www.afrilex.u-bordeaux4.fr/

[69] A titre d'exemple : ONU, Déclaration universelle des droits de l'homme, 10 décembre 1948, Charte africaine des droits de l'homme et des peuples, 26 juin 1986, Burkina Faso (Loi No 101-2004/an du 20 avril 2004 portant protection des données à caractère personnel), Tunisie (Loi organique No 2004-63 du 24 juillet 2004 portant protection des données à caractère personnel).

[70] Ouoba, Cl. *op.cit.*: 37.

[71] Durnon, G. (1981) *Guide pour la collecte des musiques et instruments traditionnels*, Editions de l'Unesco : 95.

[72] Tunisie : Loi organique n° 2004-63 du 24 juillet 2004 portant protection des données à caractère personnel.

sur la base d'un consentement libre et éclairé de la personne sans intrusion dans sa vie privée et sans coercition (…) »[73].

L'INFORMATION DES PERSONNES

Au-delà du fait qu'il s'agit d'une obligation légale et éthique du chercheur et qu'y satisfaire peut s'avérer difficile voire inapproprié dans le cadre de certaines recherches, l'information peut présenter un avantage. Elle peut être un moyen pour solliciter la collaboration des personnes concernées. G. Durnon écrit à ce sujet que « *on se rend compte de l'intérêt qu'elles portent aux recherches menées quand, au cours de l'enquête, on s'attache à expliquer qui nous sommes, les raisons de notre présence, l'objet et le but de nos recherches, ce qu'il adviendra des objets et informations recueillis sur place, de l'aide que nous sollicitons et ce qui peut être proposé en retour*[74]. »

Quelques exemples d'organismes africains :

- Acalan, Académie africaine des langues,
- Celhto, Centre d'étude linguistique et historique pour la tradition orale (Niamey, Niger),
- Cerdotola, Centre régional de documentation sur les traditions orales et les langues africaines (Yaoundé, Cameroun).

Principales sources bibliographiques

Ouoba, C. (2002), *Le droit à la vie privée au Burkina Faso, Conception, réalité juridique et socioculturelle*, Thèse, Université de Grenoble 2.

Durnon, G. (1981) *Guide pour la collecte des musiques et instruments traditionnels*, Paris, Éditions UNESCO.

Recommandation UNESCO de 1989.

Recommandation UNESCO de 2003.

Cornu, M. (2005) « A propos de l'adoption du code du patrimoine, Quelques réflexions sur les notions partagées », *Recueil Dalloz* 22 : 1452-1458

[73] Réseau africain sur l'éthique, le droit et le VIH, Déclaration de Dakar, juillet 1994, principe de l'éthique dans la recherche.
[74] G. Durnon, *op. cit.* : 41.

PRISE DE SON ET ENREGISTREMENT SUR LE TERRAIN[75]

PRINCIPES

En matière de prise de son, il n'existe pas de solution unique répondant à tous les besoins. Les principaux critères à prendre en compte sont :

- la nature de la source à enregistrer (plusieurs sources d'émission du son ou une seule : un ou plusieurs locuteurs) ;
- le contexte, et les perturbations sonores ou le parasitage qu'il peut produire ;
- la durée des entretiens et le besoin d'autonomie du matériel qui peut en découler.

Les moyens financiers dont on dispose restreignent souvent le choix en équipement. Cependant, à moyens égaux, la part consacrée à ce chapitre tend à être négligée. Au cours des quarante dernières années, la démocratisation du matériel d'enregistrement a généralement conduit à une désaffection envers le matériel haut de gamme et de fait à une moindre exigence de qualité (exemple : substitution de la cassette audio à la bande ¼ de pouce et au Nagra). Aujourd'hui comme hier, acquérir le matériel adéquat peut représenter un investissement conséquent dans un projet de collecte sonore.

Par ailleurs, tout matériel nécessite un temps d'appropriation. La prise de son requiert un certain apprentissage. Il existe des formations pratiques de quelques jours aux bases du métier de preneur de son, qui peuvent permettre d'une part de tirer tout le parti des ressources dont on dispose, et d'autre part de libérer le collecteur de soucis techniques lors de l'entretien.

Les qualités couramment attendues d'un matériel d'enregistrement sont les suivantes :

- facilité d'utilisation (pour éviter les erreurs de manipulation en cours d'enregistrement) ;
- autonomie (batterie suffisante) ;
- robustesse ;
- légèreté ;
- ergonomique (poids, taille, bouton, lisibilité des vumètres...) ;
- capacité du support d'enregistrement (pour éviter les interruptions dues aux changements de face ou de support) ;
- niveau d'entrée audio réglable (de manière à éviter sous-modulation – autrement dit un son trop faible – et surmodulation – autrement dit une saturation) ;
- sortie casque réglable ;
- en numérique, possibilité d'enregistrer dans le format recherché, et en particulier dans un format linéaire (non compressé) et pérennisable (voir fiche *Supports pour enregistrer et archiver le son*) ;
- interopérabilité et rapidité de transfert vers une station informatique (qui servira de plate-forme d'édition et de gravure).

D'autres caractéristiques moins évidentes sont à rechercher pour une réalisation de qualité :

- bruit de fonctionnement de l'appareil le plus faible possible ;

[75] Fiche rédigée par Luc Verrier et Alain Carou (BnF).

- qualité des circuits analogiques (notamment préamplification micro) ;
- câblage et connectique professionnels (symétriques : les 2 fils conducteurs sont entourés d'une tresse métallique qui les protège des parasites) ;
- alimentation fantôme 48 volts pour micro statique ;
- qualité des convertisseurs analogique/numérique (bande passante, dynamique, bruit).

Des conditions spécifiques peuvent nécessiter la prise en compte d'autres éléments :

- discrétion de l'équipement (petite taille) ;
- matériel dit « tropicalisé » (adapté aux conditions climatiques extrêmes : chaleur, froid, humidité) ;
- traitement du signal, notamment pour le travail en conditions difficiles (filtre coupe-bas pour le vent, limiteur pour l'enregistrement de sources au niveau sonore aléatoire) ;
- système d'édition intégré, pour permettre un dérushage immédiat ;
- système de gravure intégré, pour gagner en autonomie et en sécurité sans perdre en portabilité.

TYPES D'ENREGISTREURS

ENREGISTREUR SPECIALISE SUR MEMOIRE FLASH, MICRO DRIVE OU DISQUE DUR

Cette technique très fiable, de haute capacité et ouverte à tout type de format numérique, se démocratise actuellement :

- mémoire flash en baisse (1 Go coute moins de 100 euros début 2006) ;
- émergence des Micro Drive (4 Go), plus chers et plus gourmands en énergie ;
- apparition de disques durs 1.8'' (80 Go), issus de la technologie des ordinateurs portables.

Le support de stockage peut être amovible. Le transfert vers un ordinateur ou un système de stockage externe (recommandé) est très rapide. Le média de stockage, s'il est amovible, peut être directement raccordé à un ordinateur ou à une autre unité de stockage (voire à un graveur autonome) via un adaptateur et/ou une connexion informatique incorporée (USB, FireWire, SCSI).

Ces appareils possèdent généralement des entrées et sorties audio numériques (SPDIF, AES) permettant un raccordement et un transfert des données sans passer par le domaine analogique (donc sans perte).

Les composants analogiques sont similaires à ceux des DAT Pro, la partie mécanique (fragile) en moins. On peut disposer de plus de 2 canaux sur certains modèles professionnels.

Certains modèles disposent d'un système d'édition intégré.

Au sommet de cette catégorie se classent les enregistreurs sur disque dur, tels que le Nagra V, considéré comme la « Rolls » de l'enregistrement de terrain et digne remplaçant des Nagra analogiques (pour un cout équivalent). Ce type de modèle se décline également en multi-pistes (HHB, Cantar).

BALADEUR « MP3 » (VARIANTE BON MARCHE DU PRECEDENT)

Maintenant très répandu (iPod, iRiver…), ce type d'appareil est d'une utilisation très simple pour la lecture, moins pour l'enregistrement. L'autonomie est élevée (25 h), ainsi que la capacité de la mémoire (100 Go).

Le microphone intégré est d'une qualité médiocre, acceptable comme dictaphone. La qualité des circuits analogiques et l'ergonomie sont les éléments qui laissent le plus à désirer. On peut adjoindre au baladeur un pré-ampli micro/convertisseur pour pallier ses carences.

L'ergonomie est très limitée (enchainement de menus). Attention à bien disposer d'un modèle offrant un format d'enregistrement ouvert, mais aussi à paramétrer correctement le format désiré.

Ce matériel bon marché est pour le moment du « gadget » promis à une durée de vie commerciale courte : il n'y a pas d'entretien durable assuré. Avant de miser sur cette technologie, il est donc recommandé d'attendre l'arrivée prochaine d'appareils semi-professionnels plus spécialisés.

Autre innovation récente en développement : l'enregistrement sur PDA, offrant une interface couleur et les fonctionnalités d'une micro-station d'édition.

ORDINATEUR « PORTABLE »

Simple d'utilisation et générant peu de frais si on dispose déjà d'un portable pour d'autres usages, cette solution offre d'intéressantes facilités. Cependant, elle se classe plutôt dans les solutions « transportables » que « portables ». L'enregistrement se fait directement sur le disque dur de l'ordinateur ou autre solution de stockage externe, ce qui simplifie le transfert et permet de le faire dans tout type de format ouvert. Le logiciel d'édition permet l'enregistrement et le montage, voire la gravure. La stabilité de la configuration logicielle doit avoir été testée avant utilisation.

Le bruit et le rayonnement électromagnétiques générés par l'ordinateur peuvent pénaliser la qualité du signal. Les cartes-son intégrées sont souvent de qualité médiocre, et mieux vaut éviter d'utiliser l'entrée micro intégrée. Il est préférable de faire l'acquisition d'un module externe pré-ampli/micro/convertisseur, branché via une interface informatique USB ou FireWire (le module peut être simplement stéréo, mais aussi multi-canaux si besoin).

DAT

Cette technologie est en fin de vie, mais reste souvent utilisée. Elle contraint à un transfert rapide des bandes numériques, qui ne doivent en aucun cas être archivées telles quelles vu leur fragilité. Attention : la récupération d'index lors du transfert est quasi-impossible.

Techniquement, c'est une solution semi-professionnelle à professionnelle tout à fait éprouvée. Les préamplis intégrés sont de qualité. Un des talons d'Achille est la fiabilité de la mécanique d'entrainement de la bande (« machine tournante », par différence avec les modèles présentés ci-dessus), et désormais les couts d'entretien, de plus en plus onéreux.

MINIDISC

C'est une solution très bon marché, mais en passe d'être détrônée par les baladeurs-enregistreurs MP3. Le format numérique d'enregistrement a été longtemps exclusivement compressé et propriétaire (ATRAC). Le Hi-MiniDisc, lancé en 2004, permet dorénavant d'enregistrer dans un format linéaire mais

toujours propriétaire (OpenMG). Avec le logiciel SonyStage, il est possible de transférer rapidement des données en OpenMG vers un PC. Le format OpenMG n'est cependant en aucun cas à considérer comme un format d'archivage, du fait de la dépendance par rapport à l'offre technologique Sony.

Sony a récemment mis à la disposition de ses utilisateurs un utilitaire de conversion d'OpenMG vers WAV. On ignore à ce jour si la transformation est entièrement transparente.

L'ergonomie est limitée (enchaînement de menus, affichage de taille réduite). Les niveaux d'entrée ne sont pas réglables sur tous les modèles. La connectique est non professionnelle, vulnérable. Le choix de micros est restreint, sans adaptateur dédié.

CASSETTE AUDIO

Autre solution très bon marché, la cassette audio dispose toujours d'un marché (pays africains notamment) et a donc encore plusieurs années assurées. Sa robustesse et sa fiabilité sont éprouvées. Cependant, le matériel de niveau professionnel se raréfie (reste notamment Marantz PMD222).

Les limites de la cassette sont connues : durée par face, qualité moyenne, navigation difficile dans le document…

Mieux vaut éviter d'utiliser les réducteurs de bruit (Dolby), l'appareil servant par la suite de lecteur ayant en général des réglages différents de l'enregistreur.

La microcassette (utilisée dans les dictaphones) présente une qualité et une espérance de vie insuffisantes pour être encore utilisée.

Essai de synthèse comparative :

	Enregistreur dédié Flash, MicroDrive, disque dur	Baladeur MP3	Ordinateur portable
Technologie	actuelle	actuelle	actuelle
Prix	1 000/10 000 €	150-450 €	Ordinateur +200 € d'équipement spécifique
Facilité d'utilisation	semi-pro et pro	grand public	grand public ou semi-pro
Ergonomie	+	-	+
Capacité	selon disque dur	selon disque dur	selon disque dur
Formats d'enregistrement	wav, BWF, MP2, MP3	wav, MP2, MP3…	potentiellement tous
Interopérabilité	+	-	+
Qualité des convertisseurs A/D	++	-	+

	Nagra analogique	DAT	Cassette audio	MiniDisc
Technologie	fin de vie	fin de vie	fin de vie	fin de vie
Prix	1 000 € (occasion)	1 500 €	700 € (si neuf)	150 €
Facilité d'utilisation	pro	semi-pro	grand public	grand public
Ergonomie	+	+	+	-
Capacité	30 min	2 h	1-2 h	80 min
Formats d'enregistrement	analogique	PCM 16/32 à 48	analogique	ATRAC, Open MG (Hi-MD)
Interopérabilité	sans objet	-	sans objet	-
Qualité des convertisseurs A/D	sans objet	+	sans objet	-

CONSEILS PRATIQUES

MEDIAS DE COLLECTE VIERGES ET DISQUES DURS

Acheter des médias de qualité et déjà éprouvés.

Veiller à stocker les médias vierges dans un environnement de même qualité que l'archive (les dégradations physico-chimiques intervenant que le média soit enregistré ou non). Éloigner des sources de magnétisme et de chaleur.

Éviter un stockage excessif.

Protéger les cassettes, MD, DAT contre le réenregistrement (ergot de protection). Les médias magnétiques (cassettes, DAT, bandes) perdent en fiabilité au fil des cycles effacements/réenregistrements.

Les médias magnéto-optiques (MD) et mémoires flash sont réenregistrables quasiment sans limite dans la pratique (100 000 cycles écriture/lecture). La vulnérabilité réside dans la partie sensible des mémoires flash (connecteur), à manipuler avec précaution.

Les disques durs sont garantis par les fabricants pour fonctionner régulièrement et n'offrent pas les mêmes garanties en cas de stockage dormant.

CHOIX DES MICROS

Deux critères fondamentaux pour le choix d'un micro, selon l'usage auquel on le destine, sont sa sensibilité et sa directivité :

- un manque de sensibilité devra être compensé par un gain de préampli plus important, ce qui augmentera d'autant le bruit de fond (souffle) ;
- un micro couvre un espace sonore plus ou moins large, de 360° (omnidirectionnel) à 30° (micros « canons »), en passant par les micros directionnels (hypercardioïde, semi-canon).

Exemple : on choisira un micro canon pour des prises de sons précises (chant d'un oiseau), et un couple de micros cardioïdes pour des ambiances et l'enregistrement de musiques.

A côté des micros dynamiques (utilisés par les journalistes et pour la scène, permettant d'encaisser de forts niveaux) et des micros statiques (les plus respectueux des timbres, mais plus fragiles et délicats sur le terrain) existe une gamme de micros grand public : ainsi les électret (MiniDisc, alimentés par pile) et autres micros avec préampli intégré (pour usage spécifique : cravate, perche), généralement pourvus de connectique grand public (mini jack).

Conseils :

Bon rapport qualité prix : le micro Sennheiser K6 avec une capsule ME64 ou 66 (semi canon directif) assez polyvalent (ajouter une bonnette Rycotte pour les prises de son en extérieur).

Micro main type Shure SM58 ou LEM D021 : excellent micro de reportage pour les interviews en milieu bruyant, mais nécessitant d'être placé le plus près possible de la bouche...

Prévoir les accessoires nécessaires :

- pied de micro (encombrant mais adapté) ;
- pied de table (plus transportable mais peut poser des problèmes) ;
- perche, permet une optimisation rapide de la distance (solution cinéma, nécessite de la pratique et une certaine acuité) ;
- système HF ou micros sans fil: chers mais très pratiques (spectacle, cinéma, TV…) ;
- bonnette (anti-vent, plop voix…) ;
- suspension élastiques pour micro (isolation mécanique des vibrations) ;
- filtre adaptateur coupe-bas (vent, plop).

REGLAGES

Faire des essais avant de lancer l'enregistrement. Bien se préparer (longueur de câble, batterie chargée, bloc secteur, cassettes vierges de durée suffisante…)

En numérique, vérifier que l'appareil est bien paramétré : bon format et bonne résolution.

Attention à éviter la saturation (dépassement du niveau maximum) : écouter, observer les indicateurs de niveau. En cas de saturation audible, si les vumètres n'indiquent pas le maximum, c'est que le préampli est saturé en entrée : enclencher l'atténuateur d'entrée micro. Régler le niveau d'enregistrement afin qu'il n'y ait pas de dépassement du 0 dBFs seuil critique (moyenne –10 dBFs).

Attention au problème de « Larsen » (sifflement suraigu) qui peut être lié au réglage du volume du casque. Le microphone re-capte le son émis par le casque : baisser, voire couper le signal qui va au casque lorsque celui-ci n'est pas sur la tête.

Il est essentiel d'écouter ce que l'on enregistre de façon régulière pendant l'enregistrement !

Attention : certains systèmes « ne conservent pas » l'enregistrement s'il est interrompu accidentellement. Les solutions les plus évoluées (Nagra V) permettent d'avoir l'équivalent d'une lecture analogique « après enregis-

trement » (i.e. possibilité de contrôler ce qui est effectivement enre-gistré sur le disque dur).

QUELQUES CONSEILS TECHNIQUES

Un micro à la main (dynamique) doit être tenu à environ 20 cm de la bouche. Si possible, laisser quelques secondes de silence entre chaque question.

Idéal : micro-cravate couplé avec micro d'ambiance.

Si on dispose de deux entrées micros : 1. interviewé 2. les questions (on peut utiliser une petite mixette pour plus de sources).

Faire attention aux bruits de manipulation du micro et des câbles.

Ne pas coller le micro près de l'appareil (bruits mécaniques).

Penser à prendre quelques minutes en fin d'interview pour capter un « silence plateau » ou une ambiance (geste de souplesse au niveau du montage).

Une annonce en début d'enregistrement (sujet, interviewer, interviewé, date, lieu…) est un moyen simple et pérenne de garantir l'identification du contenu de l'enregistrement.

Le support d'enregistrement n'est pas forcément un support qualifié pour l'archivage (voir fiche *Supports pour enregistre et archiver le son*).

TRANSFERT ET EDITION

Cette étape est cruciale même si elle parait simple au premier abord. On utilisera une carte son avec entrée et sortie numérique optique (ADAT), SPDIF ou AES (permet un transfert sans perte et immune aux bruits parasites). La carte son ou l'interface audio et le logiciel d'édition doivent fournir un large choix de fréquences d'échantillonnage et de résolution (44.1 à 96 kHz, 16 et 24 bits), et permettre l'acquisition et la conversion de différents formats (wave, bwf, mp2, mp3…). Attention aux niveaux en numérique ou en analogique (norme d'alignement analogique-numérique : 0dBvu = -18dBFs).

Le logiciel d'édition effectue sur le son les mêmes opérations qu'un éditeur de texte (Word par exemple) : couper/copier/coller, enregistrer sous différents formats, accélérer ou ralentir le son… Il peut permettre aussi de réaliser certains filtrages (coupe-bas, ronflette, correction…) afin de fournir un document de qualité facilement écoutable. Une indexation pourra être faite afin de permettre une meilleure navigation au sein du document.

Pour finir, on réalise avant gravure une « image » du support d'archivage, incluant les données audio et les métadonnées.

SUPPORTS POUR ENREGISTRER ET ARCHIVER LE SON[76]

Assurer la conservation à long terme du son numérique

SUPPORTS D'ENREGISTREMENT, SUPPORTS D'ARCHIVAGE

L'arrêt progressif de la fabrication des matériels professionnels analogiques (à bandes et à cassettes) conduit à exclure l'archivage sous une forme autre que numérique. Tout enregistrement réalisé aujourd'hui sur support analogique impliquera à court terme un investissement non négligeable en temps et en argent pour le convertir et l'archiver dans un format numérique.

D'autre part, tous les supports d'enregistrement numérique ne réunissent pas les qualités attendues d'un support d'archivage.

Les qualités généralement attendues d'un support d'enregistrement sont sa capacité, sa maniabilité, éventuellement la possibilité d'indexation. Souvent, l'autonomie et la robustesse du matériel d'enregistrement associé, ainsi que son prix, sont les critères décisifs du choix.

Un support d'archivage numérique doit quant à lui réunir de tout autres qualités :

- Garantie de pouvoir trouver du matériel de lecture à moyen terme : large diffusion de la technologie, fabrication par plusieurs constructeurs différents.
- Possibilité de coder l'audio dans un format « ouvert » de qualité satisfaisante : la relecture du fichier ne peut être garantie à moyen et long terme si la syntaxe du format est secrète, ou en d'autres termes si la relecture de l'archive est dépendante de l'offre commerciale d'un industriel.
- Existence d'outils pour contrôler l'état de l'enregistrement sur le support : en effet, la lecture d'un support numérique ne nous apprend rien de son état de conservation, sauf lorsque l'information qu'il contenait devient illisible. La perte n'est pas proportionnelle à la dégradation comme dans le domaine analogique, mais obéit à un effet de seuil (« tout ou rien »). Il est indispensable de disposer d'outils d'évaluation de l'état du support pour engager les copies d'information en temps utile.
- Robustesse (capacité de conserver l'information dans son intégrité pendant plusieurs années). Outre ces quatre garanties fondamentales, sans lesquelles il n'est pas d'archivage numérique viable, deux autres sont également à rechercher, particulièrement dans l'optique d'une gestion de masse : simplicité des opérations de copie ; protection contre l'effacement accidentel.

Un support d'enregistrement commode et bon marché, si on considère uniquement la phase de collecte, peut se révéler bien plus couteux si l'on intègre dans le calcul la dimension archivage à long terme (en particulier s'il doit y avoir conversion du format natif à un autre format). Voici quelques exemples :

- L'usage du **MiniDisc** implique l'enregistrement dans un format propriétaire Sony. Même si les supports magnéto-optiques sont réputés d'une bonne tenue dans le temps et donc aisés à conserver

[76] Fiche rédigée par Luc Verrier et Alain Carou (BnF).

sur un plan purement physique, le MiniDisc ne répond pas aux conditions 1 (nombre de constructeurs très limité, technologie menacée à court terme), 2 (format de stockage propriétaire), 3 (pas d'outil de contrôle existant) et 5 (vitesse d'extraction bridée).

- **Le DAT** permet l'enregistrement en PCM 16 bits/48 Khz. Cependant, l'archivage sur DAT est déconseillé depuis plusieurs années en raison de la fragilité de ce support (condition 4 non remplie). Les conditions 1, 3 et 5 sont également non remplies.
- **Le CD enregistrable** une fois (CD-R) répond aux conditions 1 (technologie universelle), 2 (compatibilité tous formats de fichiers), 3 (existence d'outils d'analyse abordables), 5 et 6. Pour satisfaire à la condition 4, en revanche, des règles strictes sont à observer.

De manière générale, aucun support d'archivage n'offre aujourd'hui de garantie de pérennité sur le long terme, du fait *primo* de leur dégradation naturelle, *secundo* de l'obsolescence plus ou moins rapide des technologies de lecture. L'archivage numérique consiste donc non pas à trouver le support éternel, mais à mettre en œuvre une méthode rationnelle et réaliste de contrôle de support, de veille technologique et de migration (copies ponctuelles et copies en masse) en fonction des nécessités. Alors que dans le monde analogique, chaque génération de copie est source de perte qualitative, le nombre de copies est indifférente dans le monde numérique, du moment qu'elles sont effectuées à temps.

ARCHIVER SUR CD ENREGISTRABLE

Support et format doivent être clairement distingués. Le **support** CD-R permet l'inscription de données dans plusieurs **formats**, notamment le CD audio, lisible sur un lecteur de salon et limité à une résolution audio 16 bits/44.1 kHz ; et le CD-ROM, qui autorise le stockage de tout type de fichier, audio ou autre.

Le CD-R (appelé également CD-WORM, c'est-à-dire enregistrable une seule fois) peut être considéré comme un bon support d'archivage sur étagères (« off-line ») pour une durée de plusieurs années. Cependant, c'est depuis plusieurs années un marché essentiellement grand public, où les industriels cherchent à baisser leurs prix, y compris aux dépens de la qualité. Peu de marques réunissent donc les caractéristiques requises pour satisfaire potentiellement à un objectif d'archivage.

Quelques critères peuvent aider à s'y retrouver :

- **Capacité** : la norme « Orange Book » définit une capacité équivalant à 73 minutes de CD audio. Il est aujourd'hui quasiment impossible de trouver des CD-R de cette durée. Il est fortement recommandé de s'en tenir à ceux qui l'excèdent le moins, à savoir ceux de 80 minutes (Par précaution, on s'abstiendra de remplir le CD jusqu'au bout).
- **Couche métallique** : trois métaux sont employés (aluminium, argent, or). L'or présente la réflectivité la plus élevée, donc les meilleures chances de retrouver correctement l'information à la lecture. Des problèmes de corrosion ont été constatés avec l'argent. La qualité de la métallisation s'est révélée un point critique ces dernières années : un CD présentant une apparence

grêlée, piquée ou cloquée avant ou après gravure ne doit pas être archivé.

- **Couche de pigment** : c'est la couche qui est transformée par le passage du laser graveur. Trois pigments existent : phtalocyanine, cyanine et azo. La phtalocyanine présente une stabilité intéressante. L'identification du pigment enregistrée en en-tête du CD-R et parfois fournie par le logiciel de gravure ne doit pas être considérée comme fiable.
- **Vitesse de gravure optimale** : les CD optimisés pour des vitesses basses (jusqu'à 12x) obtiennent généralement de meilleurs résultats que les autres.
- **Production triée** : le revendeur doit pouvoir garantir que la production a été pré-triée par le fabricant, de manière à éliminer les ratés de fabrication du lot. Cela se traduit en principe par des discontinuités dans les numéros de série des CD achetés.

Cela dit, ces paramètres n'offrent pas des garanties suffisantes. La qualité d'un CD-R gravé est caractérisée par une série de paramètres, délivrés par un analyseur, parmi lesquels on retiendra :

- Le taux d'erreurs corrigibles (BLER, BERL, E22) et incorrigibles (E32).
- Les qualités de pré-traçage de la piste : Push Pull.
- La qualité de la modulation : I3, I11.
- La précision des transitions on/off du laser de gravure : symétrie, jitter.
- La majorité des analyseurs courants n'indiquent que les taux d'erreur. Les préconisations de l'Association internationale des archives sonores et audiovisuelles (IASA) fixent les valeurs à ne pas dépasser lors du contrôle qualité post-gravure :

BERL	< 5
BLER moyen	< 2
BLER max	< 10
E22	= 0
E32	= 0

Les caractéristiques de la gravure varient en fonction de la vitesse de gravure et du graveur utilisé. Aucune marque de CD-R ne peut donc être recommandée pour elle-même indépendamment du graveur avec lequel on la couple. De même, les variations dans la composition, le mode de fabrication et la rigueur du tri obligent à réexaminer très régulièrement ses choix.

Il existe deux grandes familles d'analyseurs sur le marché :

Software (PlexTools, CD Inspector, QA 201…) : ce sont les moins couteux, car ils exploitent les résultats fournis par un lecteur externe. Mais leur fiabilité dépend de celle du lecteur dont on se sert, ce qui représente le plus souvent une inconnue. Il est recommandé au moins, si on sert de ce type d'outils, de comparer les résultats obtenus en se branchant sur deux lecteurs différents.

Hardware (CATS, EC2 …) : ce sont des appareils d'analyse complets, platine de lecture incluse. Des modèles fiables existent à partir de 6 000 euros. L'analyse en vitesse réelle (1x) est vivement recommandée.

Les limites du CD-R deviennent manifestes pour de grandes masses de données : le contrôle sur analyseur et la copie sont des tâches fortement consommatrices de temps, du fait de la non-robotisation de ces tâches ; l'investissement initial est très faible, mais le coût unitaire du CD-R (support vierge + temps-personne) est peu compétitif en comparaison des solutions de stockage de masse.

ARCHIVER EN MASSE SUR BANDES OU DISQUES DURS

L'archivage à base de bandes magnétiques offre aujourd'hui le meilleur rapport entre qualité des données et prix de revient pour la gestion de grandes masses d'informations. Ces supports en cartouches peuvent être déployés dans des robotiques qui en assurent le chargement en lecteurs-enregistreurs. Ceux-ci sont couplés à des serveurs sur disque où les données sont stockées le temps d'une consultation. L'accès et le contrôle du support sont donc largement automatisables.

Plusieurs technologies sont en concurrence (LTO, Storagetek, AIT…), très fiables mais soumises à des cycles d'obsolescence très courts du fait de la densification croissante des capacités (les volumes stockables sur un support doublent tous les dix-huit mois à deux ans). Le renouvellement du parc de lecteurs-enregistreurs et la migration des données sur une nouvelle génération de supports sont à prévoir tous les quatre à six ans, selon les prévisions des fabricants.

Plus couteux, le stockage entièrement en-ligne sur disques durs nécessite une extrême sécurisation de son architecture (bien plus développée que dans les classiques schémas de réplication de données RAID).

DANS TOUS LES CAS, LE STOCKAGE NUMERIQUE EST GENERATEUR DE COUTS RELATIVEMENT IMPORTANTS TOUT AU LONG DE LA VIE DE L'INFORMATION A CONSERVER.

SUPPORTS POUR ENREGISTRER ET ARCHIVER LA VIDEO[77]

Numériser la vidéo pour la sauvegarder

VIE ET MORT DES TECHNOLOGIES VIDEO ANALOGIQUES

Au même titre que les technologies d'enregistrement audio analogiques (et à plus court terme encore peut-être), les technologies vidéo analogiques sont aujourd'hui en voie d'extinction. Devenues sans usage dans le domaine de la production, supplantées qu'elles sont par le numérique et les facilités qu'il offre, évincées par le DVD dans le domaine de l'édition du fait de leur qualité inférieure, les bandes vidéo et les vidéocassettes n'auront bientôt plus de valeur que pour les archives qui les auront collectées patiemment, mais qui n'ont pas d'autre voie que de les numériser pour continuer à en rendre le contenu accessible. Pas davantage que dans le monde de l'audio ou des documents informatiques, les besoins des archives n'ont suffi ni ne suffiront dans l'avenir à prolonger la vie d'une technologie vidéo. Il faut organiser au plus vite, si ce n'est déjà fait, l'entrée dans la sphère numérique des documents existant uniquement sous des formats analogiques. Ce au rythme le plus rapide possible, car les couts de transfert du document augmenteront aussi vite que se raréfieront le matériel de lecture en état de marche et les compétences humaines pour le maintenir.

PLUSIEURS DEGRES D'URGENCE TECHNIQUE

Depuis plusieurs années, il est devenu difficile et couteux de faire transférer des formats totalement obsolètes, tels que 2 pouces et 1 pouce (broadcast), mais aussi EIAJ, VCR (institutionnel), V2000 et Betamax (grand public). L'U-Matic et sa variante le BVU, support très représenté dans les archives institutionnelles et culturelles, est gravement menacé de par sa dégradation intrinsèque et la disparition du matériel de lecture. Apparus dans les années 80 et largement diffusés, les formats plus récents Betacam SP (broadcast) et VHS (grand public) entrent à leur tour dans la zone rouge. L'arrêt de la fabrication du matériel VHS de niveau professionnel en est un signe important. Une fois obsolète, le matériel acquis ne peut être entretenu qu'un temps limité. La disponibilité des pièces détachées est généralement garantie pendant moins de dix ans après la cessation de fabrication.

Ces degrés d'urgence déterminent ainsi des niveaux de priorité technique, qui sont ensuite à croiser avec les priorités documentaires.

OBJECTIF : LE STOCKAGE DE MASSE DANS DES STANDARDS OUVERTS

Numériser, mais dans quel format de données ?

Côté format, le choix d'un standard « ouvert » s'impose, quel que soit le média dont on envisage la numérisation (image fixe, image animée, écrit, son). Autrement dit, les règles d'interprétation informatique des données numériques en signal vidéo doivent être publiques. Un format « propriétaire » (secret industriel) sera difficile, voire impossible à restituer le jour où le matériel qui lui est associé aura disparu.

[77] Fiche rédigée par Alain Carou et Dominique Théron (BnF).

L'HYPOTHESE BETA NUMERIQUE

A ce titre, le transfert sur Betacam numérique (format propriétaire Sony) ne peut être qu'une solution transitoire, une étape de transfert avant le passage à un format numérique pérenne. En fin de vie de cette technologie, il sera nécessaire de relire en vitesse réelle les supports pour passer à un autre format. L'évaluation du cout réel d'une sauvegarde en passant par le Beta numérique (support déjà couteux à la base) doit donc intégrer le cout d'une migration ultérieure. Cette option aura cependant un intérêt dans deux cas :

- Si l'on veut stocker à terme dans un format numérique sans compression (voir § suivant), mais que monter une chaine de numérisation de masse de ce type soit difficile à court terme, le passage par le Beta numérique représente une mesure conservatoire.
- Si l'archive n'est pas du tout prête à un archivage de masse rationnel de fichiers numériques : le Beta numérique permet alors de rester dans une logique traditionnelle de supports sur étagères, le temps de préparer la mutation nécessaire.

COMPRESSION OU NON ?

La compression vidéo repose sur l'élimination de détails pas ou peu perçus par l'œil et l'utilisation des redondances d'une image à l'autre : l'économie faite dans la description des images successives entraine une réduction du volume de données d'un facteur 10, 20 ou 50.

Le choix de la compression s'impose aujourd'hui pour toute opération de numérisation en masse si l'on veut rester dans des échelles de cout raisonnables. Le principe de réalité entre ainsi en conflit avec la règle déontologique, strictement observée dans le monde des archives sonores, qui imposerait une numérisation sans compression. Pour des usages spécifiques de recherche, qui requièrent un maximum de définition et des analyses image par image (exemples : une opération chirurgicale, une interprétation musicale filmée en plan large), un minimum de compression, voire pas de compression du tout, est une option à examiner sérieusement.

Le taux de compression (défini en nombre de bits par seconde) est à choisir en fonction de la qualité du format d'origine : 6 Mbits/sec suffisent amplement pour le VHS, 12 Mbits/sec paraissent nécessaires pour le Betacam SP.

LES NORMES MPEG

Les normes MPEG-2 et 4 répondent à l'exigence de compression et de format ouvert. A l'heure actuelle, le MPEG-2 reste le standard dominant. Le développement de la norme MPEG-4 (qualité analogue, voire supérieure, pour un débit moindre) est cependant à suivre.

La conformité des numériseurs à la norme qu'ils sont supposés produire doit être contrôlée, dans la mesure où elle n'est pas systématiquement assurée (exemple : respect de la résolution de l'écran 720x576). Des outils logiciels d'analyse du flux MPEG existent pour cela.

METADONNEES

Un document numérique, quel qu'il soit, n'est pas pérennisable sans un minimum de métadonnées associées. Les métadonnées minimales sont celles qui permettront l'identification du contenu, la description complète de son mode de production (description de la chaine de numérisation) et les caractéristiques techniques du format qui permettront d'engager des actions de pérennisation (par exemple la migration vers un autre format) en cas de

risque. Pour être exploitables informatiquement, ces métadonnées doivent obéir strictement à une formalisation (par exemple dans le langage de balise XML). Afin de limiter au maximum les saisies manuelles, perte de temps pour les techniciens, les métadonnées devront être générées automatiquement en exploitant les informations déjà connues préalablement (celles issues notamment du travail de préparation documentaire).

D'autres métadonnées pourront par ailleurs être ajoutées à loisir selon l'usage : vignettes périodiquement extraites du document comme aide à la consultation ; image numérisée de jaquettes ou de fiches papier associées au document vidéo ; indexation temporelle du contenu ; ou encore (dans le futur) reconnaissance de la voix permettant une recherche « plein texte », etc.

ORGANISATION DE LA CHAINE DE NUMERISATION

La numérisation se décompose en plusieurs étapes mais a pour règle de base la meilleure relecture possible du document d'origine.

PREPARATION DOCUMENTAIRE ET PHYSIQUE DES ELEMENTS

Le travail commence par une identification du support, du standard couleur (ou NB), de la durée et, si possible, du contenu. Idéalement, un magnétoscope permettant de relire les bandes concernées doit donc être à la disposition de la personne chargée de cette préparation. Mais les archives anciennes (antérieures à 1975) et /ou broadcast (antérieures à 1985), sont des bandes magnétiques sur flasques qui peuvent réclamer l'aide d'un prestataire seul équipé pour cela. Une fois les analyses, tris et classements effectués, une liste dans un tableur devient l'outil de base. Attention à ne pas sous-estimer l'organisation des informations dans cette liste qui servira à alimenter diverses bases de données par la suite.

Il faut alors nettoyer la bande, avec des machines spécialisées quand elles existent, ou grâce à un passage sur un magnétoscope de réforme et un essuyage manuel quand il n'y a pas d'autre solution.

CHAINE DE TRANSFERT

Vient ensuite le magnétoscope : bon état général, têtes de lectures neuves ou récentes, niveaux audio et vidéo correctement réglés sont la base. Le standard couleur, s'il n'est pas en PAL d'origine (mais en Secam ou NTSC), doit impérativement être transcodé dans de bonnes conditions grâce à un appareil spécifique.

Un autre élément incontournable de la chaine de lecture est le TBC (correcteur de base temps), appareil voué à compenser les instabilités et fluctuations temporelles présentes sur le signal vidéo. Le recours à des machines contemporaines des sources à traiter, le plus souvent analogiques, permet de résoudre un certain nombre de problèmes dépassant les normes actuelles qu'un TBC numérique contemporain sera incapable de traiter.

Cette phase de lecture du document ne saurait être complète sans évoquer les différents outils de contrôle nécessaires : oscilloscope, vecteurscope (PAL), moniteur vidéo et audio de bonne qualité.

NUMERISATION, COMPRESSION

Suivent la conversion analogique-numérique proprement dite, et la compression. La performance des cartes d'encodage vidéo en matière de compression varie d'un modèle à l'autre. Il est indispensable de tester leur fiabilité en examinant leur capacité à gérer des images très mouvantes (par

exemple la surface de l'eau, ou une danse à un carnaval) sans générer de défauts (carrés figés, pixellisation).

CONTROLE QUALITE

Un contrôle qualité de la numérisation et des métadonnées s'impose avant la sauvegarde sur support d'archivage. Dans l'intervalle, le fichier reste sur un serveur-tampon (baie de disques sécurisée).

Il porte sur la conformité des noms de fichiers, des métadonnées et sur la qualité du résultat livré. L'attention du vérificateur devra se porter sur les « pertes de synchronisation » (perte de l'image et du son), les problèmes de « tracking » (suivi de piste, réglable sur le scope), la présence des canaux son et le réglage mono/stéréo.

A l'organisation du contrôle puis du versement dans l'archive numérique finale (cf. *infra*) doivent répondre impérativement des capacités serveur et réseau adaptées.

CORRECTION DU SIGNAL

Un traitement du signal peut être souhaitable. Quand on a affaire à des supports de qualité médiocre, un débruitage en amont de la numérisation est indispensable pour permettre une compression MPEG correcte. En effet, le bruit (parasitage aléatoire du signal) représente en numérique une masse considérable d'informations à gérer en plus du signal utile.

D'autres opérations peuvent intervenir en aval de la numérisation, avec des outils très performants. Il convient de dissocier restauration linéaire avec des réglages moyens (colorimétrie par exemple) s'appliquant à tout le document, et restauration plan à plan. Le facteur « temps passé » d'opérateurs spécialisés est discriminant entre la restauration de documents seulement destinés à l'archivage et la restauration de documents ou d'extraits voués à une diffusion commerciale. Dans le cas d'une intervention lourde, il devrait être décidé, pour des raisons déontologiques, d'archiver la copie droite (avant restauration) en plus du résultat final.

En tout cas, la restauration ne doit pas être considérée comme un substitut à une lecture de qualité, c'est une opération complémentaire.

VERSEMENT DANS L'ARCHIVE NUMERIQUE

Il n'existe pas de technologie de stockage numérique pérenne. Le stockage numérique est donc affaire de migrations (copies) en masse périodiques et contrôlées. Rien à voir avec la lourdeur de la copie d'analogique en numérique : la copie de numérique à numérique peut être automatisée et ultra-rapide.

Les technologies de bandes d'archivage offrent les niveaux de sécurité (élevé) et de cout (bas) recherchés. Super DLT, Super AIT, LTO sont des choix correspondant aux besoins de la vidéo, avec des capacités de stockage sur bande allant actuellement de 100 à 400 Go. Des robotiques ou, à moindre échelle, des systèmes auto-loader avec un lecteur permettent de réaliser les opérations de lecture, contrôle d'état des supports et copie sans manipulation humaine. Ces technologies sont soumises à un cycle d'obsolescence rapide, qui contraindra à des migrations de masse tous les cinq ou six ans environ. D'où la nécessité impérieuse de disposer d'une visibilité financière à moyen terme, au-delà de l'opération de passage au numérique, pour garantir la pérennité des investissements engagés et – surtout – l'accès aux fonds qui ne seront bientôt plus accessibles du tout sous leur forme analogique d'origine.

Quelques références complémentaires en ligne (essentiellement en anglais) :

Identifier les formats vidéo à vue d'œil et connaitre le niveau de risque technique :

www.video-id.com

Conserver l'accès aux données numériques :

bibnum.bnf.fr/conservation/infopreservation_fr.pdf

Le format de métadonnées de préservation METS :

www.loc.goc/standards/mets

Numériser sans compression la vidéo scientifique, une démarche pionnière de la Phonogrammarchiv de Vienne (à lire dans un souci prospectif, mais encore difficile à mettre en œuvre dans les limites économiques habituelles) : www.pha.oew.ac.at/phawww/literatur/iasa21_2003.pdf

CODAGES ET FORMATS

Pour les ressources enregistrées, leurs annotations linguistiques et documentaires.

Dans le monde informatique, les données sont codées en suivant des codages explicitement définis et organisés logiquement dans des formats de fichier. Ces derniers sont eux-mêmes stockés sur des supports ayant leur propre organisation physico-logique.

LES GRANDS PRINCIPES GUIDANT LE CHOIX D'UN CODAGE OU D'UN FORMAT

La distinction la plus importante est celle qui est faite entre *propriétaire* et *non-propriétaire*. Un codage propriétaire est un codage qui appartient à une personne ou une société qui en garde secrète la description. Il s'agit en règle générale d'une stratégie commerciale. Un tel codage est à bannir pour la conservation à long terme dans la mesure où les données ainsi codées risquent de disparaitre avec le secret de leur description. Seul un codage non propriétaire et libre permet une conservation dans de bonnes conditions.

Un autre aspect important, étroitement lié à l'aspect non-propriétaire, est la *standardisation* ou la *normalisation*. On peut définir un standard comme un accord entre des fabricants industriels qui défendent leur intérêts (souvent commerciaux), alors qu'une norme est un accord passé au sein d'un État (normes nationales : par exemple l'AFNOR) ou entre des États (normes internationales : par exemple l'ISO). Les organismes de normalisation prévoient aussi des mécanismes d'entretien et de conservation des normes créées, ce qui n'est pas forcément le cas des organismes de standardisation. On privilégiera les normes internationales dans la mesure où elles représentent la meilleure garantie de maintien de la connaissance indispensable à une interprétation correcte des données.

Un autre aspect auquel il faut prêter attention est la possibilité, pour un codage, d'utiliser des techniques protégées par des brevets, ce qui peut en limiter l'usage pendant un certain temps et/ou sur une certaine zone géographique. Par exemple, le groupe de travail « Moving Pictures Experts Group » qui gère, sous les hospices de l'ISO, les standards de compression, de décompression, de codage... pour l'image animée et pour le son, a notamment défini un standard connu sous le nom de « MP3 » ou « MPEG audio Layer 3 ». Ce codage, qui pour l'usager semble libre parce qu'utilisé dans des outils eux-mêmes gratuits, est en fait couvert par un brevet détenu par les sociétés Fraunhofer IIS et Thomson, et n'est ni libre ni gratuit.

LES DONNEES AUDIO

CODAGES

Un codage au sens étroit du terme désigne le type de correspondance que l'on souhaite établir entre chaque valeur du signal analogique et le nombre binaire qui représentera cette valeur. Il existe différents types de codages :

PCM : (Pulse Coded Modulation) c'est la valeur réelle de la mesure qui est représentée ;

Différentiel : c'est la différence entre le niveau du signal à l'instant de l'échantillonnage et le niveau qu'il avait lors de l'échantillonnage précédent qui est représenté.

Prédictif : il prévoit la valeur suivante d'après l'historique des valeurs échantillonnées. Le codage mesure seulement la différence entre la valeur prévue et la valeur réelle.

Adaptatif : il adapte la résolution (nombre de bits) au type de variation sonore détecté.

Le codage le plus simple et le plus répandu est certainement le codage PCM, même si ce n'est pas le plus économique en espace de stockage ou en temps de transfert. En dehors de ces choix de codage, la qualité de l'enregistrement dépendra du matériel de prise de son ou de numérisation, de la situation d'enregistrement, ainsi que des caractéristiques de numérisation : fréquence d'échantillonnage, résolution de l'échantillon et nombre de canaux[78].

Il est aussi d'usage de parler de codages pour les algorithmes de compression que l'on peut appliquer aux données. Ces algorithmes proposés dans des programmes appelés *codec* aboutissent généralement à une perte d'information (MACE, MPEG, u_law, etc.), c'est-à-dire que le résultat de la décompression des données n'est pas identique à l'original. En général, une bonne compression de parole ou de musique propose de supprimer en priorité les informations que la physiologie de l'oreille humaine ne permet pas d'entendre (voir fiche *Codages*). Ces algorithmes ont pour but de diminuer la taille des fichiers ou d'accroître le débit des transferts. Pour de la conservation de document il est bien évident que l'on ne se tournera pas vers de telles solutions. Pour la même raison, on évitera l'utilisation d'outils comme les enregistreurs miniDisc qui appliquent à la source un algorithme de compression.

FORMATS

Un format de fichier définit les règles d'écriture et l'organisation des données encodées. Ces règles sont utilisées par les logiciels pour écrire/enregistrer et pour lire/écouter. Les formats de fichier audio sont assez nombreux (RIFF/wav, AIFF, AU, MP3…). Ils peuvent éventuellement être liés à certains codages (par exemple, le format MP3 est lié au codage MPEG). Comme pour les codages, le choix d'un format reposera sur son aspect propriétaire ou non, normalisé ou non. Une attention particulière sera apportée à l'aspect libre du format. En effet, certains formats sont liés à des techniques soumises à des brevets qui ne pourront pas forcément être acquittés des utilisateurs successifs. De plus l'emploi de ces formats est souvent limité aux seules solutions que le fabricant logiciel qui détient le brevet propose, en général uniquement pour les plates-formes porteuses commercialement (MS-Windows, MacOS, etc.). A plus long terme, si vous ne trouvez pas comment normaliser vos données, vous risquez de ne plus pouvoir les lire (les fabricants de logiciels ne sont pas tenus d'en assurer la maintenance).

[78] Pour ces caractéristiques, nous conseillons d'adopter celles des CD-Audio (bon compromis qualité/quantité), c'est-à-dire : 44 100 Hz, 16 bits, mono ou stéréo (en fonction des conditions de l'enregistrement). L'organisme IASA (International Association of Sound and Audiovisual Archives) préconise actuellement, pour la conservation des données audio, des caractéristiques plus élevées : 96 KHz, 24 bits, format BWF).

LES ANNOTATIONS LINGUISTIQUES

CODAGES

Les annotations linguistiques sont composées de définitions d'objets linguistiques (les mots, les morphèmes, les tours de parole, etc.) et de commentaires sur ces objets. On distingue généralement dans les commentaires, les transcriptions qui donnent une version écrite de l'oral (en utilisant un certain nombre de conventions de notation comme celles de l'API), des autres annotations que sont les traductions, les gloses, les indications de mise en scène, etc. qui utilisent, elles, une métalangue (la plupart du temps, il s'agit de la langue de l'annotateur). Toutes ces annotations requièrent la mise en place d'un système de codage des caractères. Les conventions d'écriture des langues précisent aussi le sens de l'écriture, l'ordre des éléments à utiliser lors d'un tri, les équivalences de casse, l'utilisation de la ponctuation, etc. Depuis 1990 (date de la version 1.0), nous disposons d'un code « universel » qui fédère l'ensemble des codes existants. Ce code (Unicode[79]) est synchronisé sur la norme ISO-10646 qui a le même objectif. Il est déjà largement utilisé et a été adopté notamment par la Toile. Il permet donc de coder des documents multilingues mélangeant des écritures aux caractères et aux propriétés différents, et ceci de manière indépendante de la plate-forme informatique utilisée, ce qui facilite l'échange et le partage des documents. Dans la mesure où il n'y a pas d'autres propositions de codage en concurrence, Unicode est devenu incontournable.

Le reste des annotations concerne les objets de l'analyse. Ces objets doivent à la fois être définis et utilisés de manière identifiable dans les documents. Les codages utilisés en linguistique sont très liés aux théories employées et sont très peu formalisés, de sorte qu'il n'y a pratiquement pas d'implémentation informatique. La plupart du temps, il s'agit tout au plus d'ontologies. A notre connaissance, le travail le plus abouti et accepté comme un standard est certainement la TEI (Text Encoding Initiative). Elle a pour vocation le codage de la structure logique d'un certain nombre de types de documents utilisés dans la littérature, la linguistique, etc., comme par exemple les poèmes, les pièces de théâtre, les dictionnaires, les transcriptions de la parole. Ces propositions de codage ne sont pas forcément adéquates à toutes les analyses possibles, mais il est judicieux, au moment de choisir un codage, de se situer par rapport à celles existantes. Il sera aussi utile de suivre les avancées du groupe de travail de l'ISO/TC 37/SC4 qui porte sur la gestion des ressources linguistiques, qui est aujourd'hui en cours d'élaboration et qui concernera autant le codage des annotations linguistiques que celui des métadonnées documentaires.

FORMATS

Les deux familles principales de format de fichier pour structuration de l'information sont les bases de données relationnelles et les langages de balisage de textes. Nous ne parlerons pas d'une troisième grande famille représentée par l'ensemble des systèmes propriétaires, qu'ils puisent leurs justifications historiquement ou commercialement, ni des outils dont le but n'est pas la structuration de l'annotation mais sa présentation typographique, sa mise en page (logiciels de traitement de texte).

---

[79] Site web du Consortium Unicode (http://www.unicode.org).

Les bases de données sont généralement utilisées pour traiter des données de calcul alors que les systèmes de balisage de texte le sont pour les données textuelles. Ces deux mondes sont beaucoup plus entrelacés que par le passé.

La plus grande révolution a certainement été l'arrivée en 1998 du langage de balisage de texte XML. Ce dernier est un avatar de SGML, lui même normalisé ISO-8879 en 1986. XML est à la fois plus simple et plus moderne que son ancêtre. Il est bien intégré dans le web. Il s'agit en fait de tout un ensemble de technologies (Xpath pour l'identification et la navigation dans une arborescence XML, Xlink et Xpointer pour l'expression des liens, XSL pour la définition de feuilles de styles, Xquery comme langage de requête, DOM comme interface de programmation…). L'ensemble de ces technologies est géré par le consortium W3. Son adoption par les fabricants de logiciel a été très rapide et XML est considéré maintenant comme un standard incontournable pour la structuration, la gestion et l'échange des ressources. Du point de vue des bases de données relationnelles, il est surtout utilisé comme un format d'échange permettant de passer d'un système à un autre. Plus récemment, l'apparition de bases de données natives XML a rendu plus floue la distinction entre ces deux mondes.

Un des grands principes de XML est la séparation de la structure logique de la structure physique (par exemple sa mise en page). Une autre propriété de XML est qu'il permet de définir une syntaxe formelle pour la description de la structure logique des documents que l'on souhaite créer. C'est ce qu'a fait la TEI en définissant une ou des DTD[80].

LES METADONNEES

Les métadonnées servent à décrire des ressources (enregistrements, annotations). Ces descriptions peuvent contenir des informations sur la nature physique des ressources (durée de l'enregistrement, format de fichier, etc.), sur les droits associés, sur la situation d'enquête (lieu, date, participants, etc.). Ces métadonnées correspondent aux renseignements que l'on pourrait trouver dans une notice bibliographique de bibliothèque. Il existe un certain nombre de renseignements communs avec ce type de notice, mais les caractéristique propres des corpus oraux, ainsi que les préoccupations particulières des personnes qui les étudient, ont conduit à la définition de champs tels que l'âge du locuteur ou les conditions d'enregistrement, que l'on aura plus de mal à faire entrer dans une notice classique de bibliothèque. Les métadonnées servent principalement à deux choses : à cataloguer et à échanger. Pour que les échanges soient possibles, il convient de normaliser à la fois la forme des métadonnées mais aussi la procédure d'échange.

CODAGES

Plusieurs codages ont été proposés et sont utilisés pour la description des enregistrements et de leurs annotations. La TEI propose d'écrire toutes ces informations dans un en-tête assez détaillé. Pour les ressources du web, Dublin-Core[81], normalisé ISO-15836 en 2003, propose un jeu de quinze étiquettes qui sont notamment utilisées dans les en-têtes des fichiers HTML. Il existe bien sûr les codages pratiqués par les bibliothèques tels que les standards Marc, US-Marc, etc. qui se sont adaptés pour coder les nouveaux

[80] Document Type Definition.

[81] Site web du Dublin Core Metadata Initiative (http://dublincore.org).

supports informatiques. Il existe aussi des communautés qui ont proposé des recommandations comme par exemple OLAC[82] (basé sur du Dublin-Core enrichi et spécifié pour l'adapter aux ressources linguistiques), ou IMDI[83].

FORMATS

Quelle que soit la manière dont les métadonnées sont encodées (préconisation Dublin-Core, OLAC, TEI ou IMDI), la tendance générale est à l'utilisation de XML comme format d'échange. Le libre choix est laissé aux gestionnaires des métadonnées de les structurer directement en XML, en utilisant une base de données ou toute autre solution. Le choix d'une solution repose sur les critères énoncés précédemment.

LES PROTOCOLES D'ECHANGE

Le protocole Z39.50 est une norme ANSI/NISO, gérée actuellement par la « Library of Congress ». Sa vocation est la recherche automatisée d'informations bibliographiques dans des bases de données réparties. Le but originel était l'interconnexion des systèmes ouverts (OSI). En fait, la plupart des implémentations existantes ont superposé ce protocole sur TCP-IP plutôt que sur les couches définies dans le modèle OSI. De nombreuses bibliothèques universitaires utilisent ce protocole pour échanger leurs notices bibliographiques. Leur nombre est actuellement en forte croissance.

L'OAI (Open Archive Initiative) est une organisation plus récente qui définit entre autres un protocole relativement simple pour la récolte de métadonnées dans les archives ou « réservoirs » de données. A l'origine, il s'agissait de permettre l'interopérabilité entre les différentes archives de pré-prints et e-prints qui avaient chacune leur langage de requête. Ce protocole comprend un petit nombre de requêtes qu'il est possible d'adresser à un détenteur d'archives. Par exemple, on peut obtenir d'un détenteur d'archives son identification, la liste de ses identifiants de ressources, la liste des encodages qu'il utilise pour ses metadonnées, etc. Ce protocole fixe aussi la syntaxe XML des réponses que peut émettre un fournisseur d'archives. Un certain nombre de règles de politesse doivent être implémentées par le fournisseur, comme l'envoi de codes particuliers pour signaler les erreurs de syntaxe des requêtes. Ce protocole a l'avantage d'être simple à mettre en œuvre, et d'utiliser XML pour le formatage. L'objectif de l'OAI est la standardisation de la procédure de collecte des métadonnées afin de permettre à des fournisseurs de services (par exemple des moteurs de recherche) d'effectuer leur travail sur des métadonnées préalablement centralisées. En effet, la recherche directe à travers un ensemble réparti de fournisseurs, comme c'est le cas avec le protocole Z39.50, pose des problèmes de performance lorsque certains nœuds du réseau ralentissent ou bloquent la poursuite.

[82] Open Language Archives Community (http://www.language-archives.org).

[83] EAGLES/ISLE Metadata Initiative (http://www.mpi.nl/IMDI/).

BIBLIOTHEQUE NATIONALE DE FRANCE

LES STATUTS DE LA BIBLIOTHEQUE NATIONALE DE FRANCE :

Dans son article 2, le décret no 94-3 du 3 janvier 1994 « portant création de la Bibliothèque nationale de France » indique que celle-ci :

> *« a pour missions […]*
>
> *de collecter, cataloguer, conserver et enrichir dans tous les champs de la connaissance, le patrimoine national dont elle a la garde, en particulier le patrimoine de langue française*[84] *ou relatif à la civilisation française. A ce titre,*
>
> – elle exerce, en vertu de l'article 5, alinéa 2, de la loi du 20 juin 1992[85] […] les missions relatives au dépôt légal confiées par cette loi et les décrets pris pour son application à la Bibliothèque nationale ; elle gère, pour le compte de l'État, dans les conditions prévues par la loi du 20 juin 1992 susvisée, le dépôt légal dont elle est dépositaire. Elle en constitue et diffuse la bibliographie nationale.
> – elle rassemble, au nom et pour le compte de l'État, et catalogue des collections françaises et étrangères d'imprimés, de manuscrits, de monnaies et médailles, d'estampes, de photographies, de cartes et plans, de musique, de chorégraphies, de documents sonores, audiovisuels et informatiques. […]
>
> *d'assurer l'accès du plus grand nombre aux collections […].A ce titre : - elle conduit des programmes de recherche en relation avec le patrimoine dont elle a la charge […] ; elle coopère avec d'autres bibliothèques et centres de recherche et de documentation français ou étrangers, notamment dans le cadre des réseaux documentaires ; […] elle permet la consultation à distance […] ; elle mène toutes actions pour mettre en valeur ses collections. »*

De cet ensemble d'activités, on retiendra six missions fondamentales de l'activité de la BnF : enrichir le patrimoine national dont elle a la garde, cataloguer ses collections, les conserver, les communiquer au public, les valoriser et enfin coopérer avec d'autres établissements. En ce qui concerne l'enrichissement des collections, nous renvoyons à la partie du chapitre quatre consacrée à la BnF. Rapportées au département de l'Audiovisuel, en charge du patrimoine sonore, vidéographique, multimédia et électronique au sein de la BnF, les cinq autres missions se déclinent comme suit.

---

[84] Nous soulignons.

[85] Remplacée depuis par les articles L131-1 à L133-1 relatifs au dépôt légal du Code du patrimoine (*Journal officiel* du 24 février 2004).

LA CONSERVATION DES DOCUMENTS SONORES (ET AUDIOVISUELS)

Conservant une collection d'un million d'enregistrements sonores, 120 000 documents vidéographiques et 70 000 documents multimédias et électroniques, le département de l'Audiovisuel a mis en place un plan de sauvegarde de ses collections, visant au transfert progressif des supports fragiles ou obsolètes sur support numérique. En ce qui concerne le son, les cylindres, les disques à gravure directe " Pyral ", les supports magnétiques (bandes, cassettes…) sont reportés en priorité sur support numérique (sur mémoire de masse informatique et sur CD-R). L'ensemble de la vidéo analogique (VHS, U-Matic…) a été intégrée dans le même plan de sauvegarde et a été numérisée.

Le département de l'Audiovisuel est membre de l'IASA (International Association of Sound and Audiovisual Archives, http://www.iasa-web.org dont il suit et relaye en France les préconisations en termes de conservation des supports audiovisuels. Il est également membre actif de l'Association Française des détenteurs de documents Audiovisuels et Sonores (AFAS) qui organise régulièrement des journées d'étude sur les questions de numérisation (voir sur le site de l'association : http://afas.mmsh.univ-aix.fr/).

LE TRAITEMENT DOCUMENTAIRE

L'ensemble des collections du département de l'Audiovisuel fait l'objet d'un traitement documentaire informatisé en format INTERMARC. Le catalogue du département de l'Audiovisuel est intégré au catalogue général de la Bibliothèque, BN-OPALE PLUS, et peut être consulté en ligne à l'adresse : http://www.bnf.fr. A noter toutefois qu'en raison de la complexité de certains fonds, un service de recherche à distance permet de répondre aux questions des usagers : audiovisuel@bnf.fr.

LA CONSULTATION DES DOCUMENTS SONORES ET AUDIOVISUELS

Deux salles audiovisuelles ont été programmées sur le site François-Mitterrand-Tolbiac de la Bibliothèque : l'une au niveau Tout public (salle B), en Haut de Jardin, l'autre au niveau « Recherche » en Rez de Jardin (salle P). Celle-ci est équipée de 54 places audiovisuelles et de 17 cabines d'écoute et de visionnage. Accessible aux chercheurs, sur accréditation, elle offre à la consultation l'ensemble de la collection patrimoniale audiovisuelle du département. Un système audiovisuel constitué de régies manuelle ou robotisée, de serveurs numériques, de postes de consultation permet la communication de l'ensemble de ces documents.

LA VALORISATION DU PATRIMOINE SONORE

A l'heure actuelle le département de l'Audiovisuel offre une trentaine d'heures d'enregistrements sonores issues de ses collections à l'écoute en ligne sur le site de la Bibliothèque : http://www.bnf.fr, notamment dans les programmes « Gallica »,

« Gallica-Voyage en France », http://gallica.bnf.fr/voyagesenfrance/ ;

« Gallica-Voyage en Afrique », http://gallica.bnf.fr/VoyagesEnAfrique/ ;

« Anthologie », http://gallica.bnf.fr/Anthologie/.

Cette offre est évidemment destinée à croitre avec, dans un premier temps, le projet de mise en ligne de l'ensemble des enregistrements produits par les Archives de la Parole entre 1911 et 1914.

LA COOPERATION AU PLAN NATIONAL ET INTERNATIONAL

L'alinéa 4 de l'article 3 du décret du 3 janvier 1994 précise que la Bibliothèque nationale de France peut « coopérer, en particulier par la voie de convention ou de participation à des groupements d'intérêt public, avec toute personne publique ou privée, française ou étrangère, et notamment avec les institutions qui ont des missions complémentaires des siennes ou qui lui apportent leur concours ». Cette coopération prend place au sein du « Département de la Coopération » de la BnF qui travaille en étroite relation avec les départements de collections dont le département de l'Audiovisuel. Les « pôles associés » sont une illustration de cette coopération. A l'heure actuelle, dans le domaine de l'archive sonore, quatre centres (Conservatoire occitan, Dastum, Maison méditerranéenne des sciences de l'homme, Métive) affiliés à la Fédération des Associations de Musiques et de Danses Traditionnelles (FAMDT) sont ainsi pôles associés de la BnF et perçoivent une aide au traitement documentaire de leurs fonds. Aujourd'hui, la coopération s'oriente également vers des actions de numérisation partagée, de mise en place de projets de catalogues collectifs, etc.

LES ARCHIVES : LEGISLATION

LES ARCHIVES DE FRANCE

La direction des Archives de France est une direction du Ministère de la Culture et de la Communication qui assure la mise en œuvre et le contrôle de la loi 79-18 du 3 janvier 1979 sur les archives, aujourd'hui codifiée dans le code du patrimoine (ordonnance du 20 février 2004). Elle coordonne toutes les attributions confiées par la loi à l'administration des archives, à l'exception de celles qui concernent les archives des ministères des affaires étrangères et de la défense, et des services et établissements qui en dépendent ou qui y sont rattachés.

Depuis l'entrée en vigueur de la loi n° 83-663 du 22 juillet 1983, la direction des Archives de France ne gère plus directement les archives départementales, placées désormais sous l'autorité des conseils généraux, mais elle garde sur elles un contrôle scientifique et technique.

La direction des Archives de France comprend (arrêté du 25 mars 2002) :

- l'Inspection générale ;
- la délégation aux célébrations nationales ;
- le département du réseau institutionnel et professionnel ;
- le département de la politique archivistique et de la coordination interministérielle ;
- le département de l'innovation technologique et de la normalisation ;
- le département des publics ;
- le bureau des affaires générales et de la documentation.

LES ARCHIVES AU SENS DU CODE DU PATRIMOINE (LIVRE II)

Les archives sont définies à l'article L 211-1 comme suit :

> *« Les archives sont l'ensemble des documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé, dans l'exercice de leur activité. La conservation de ces documents est organisée dans l'intérêt public tant pour les besoins de la gestion et de la justification des droits des personnes physiques ou morales, publiques ou privées, que pour la documentation historique de la recherche. »*

La loi distingue deux catégories d'archives : les archives publiques et les archives privées.

LES ARCHIVES PUBLIQUES

« Sont considérées comme archives publiques :

Les documents qui procèdent de l'activité de l'État, des collectivités locales, des établissements et entreprises publics ;

Les documents qui procèdent de l'activité des organismes de droit privé chargés de la gestion des services publics ou d'une mission de service public ;

Les minutes et répertoires des officiers publics ou ministériels (article L 211-4). »

« Les archives publiques, quel qu'en soit le possesseur, sont imprescriptibles » (art. L 212-1).

« Les conditions de leur conservation sont déterminées par décret en Conseil d'État » (art. L 212-2).

« Les archives publiques font l'objet de procédures de sélection et de certaines règles précises d'élimination. »

A l'expiration de leur période d'utilisation courante par les services, établissements et organismes qui les ont produits ou reçus, les documents mentionnés à l'article 211-4 font l'objet d'un tri pour séparer les documents à conserver et les documents dépourvus d'intérêt administratif et historique, destinés à l'élimination :

« La liste des documents destinés à l'élimination ainsi que les conditions de leur élimination sont fixées en accord entre l'autorité qui les a produits ou reçus et l'administration des archives ».

LES ARCHIVES PRIVEES

« Les archives privées sont l'ensemble des documents définis à l'article 1er qui n'entrent pas dans le champ d'application de l'article 211-4 » (art. 211-5)

C'est le mode de production et non pas le type de support ou le sujet qui définit l'appartenance à l'une ou l'autre catégorie.

Exemple : l'enregistrement d'une séance du Conseil général est un document d'archives public tandis que l'enregistrement d'une interview d'un personnage politique à la radio est un document d'archives privé.

MODALITES DE CONSULTATION

Les *modalités de consultation* diffèrent selon la catégorie : pour *les archives publiques,* la communication est encadrée par la loi

« Article 6. – Les documents dont la communication était libre avant leur dépôt aux archives publiques continueront d'être communiqués sans restriction d'aucune sorte à toute personne qui en fera la demande. »

[...]

Tous les autres documents d'archives publiques pourront être librement consultés à l'expiration d'un délai de trente ans ou des délais spéciaux prévus à l'article L 213-2.

Article L 213-2 Le délai au-delà duquel les documents d'archives publiques peuvent être librement consultés est porté à :

> *« a) Cent cinquante ans à compter de la date de naissance pour les documents comportant des renseignements individuels de caractère médical ;*
>
> *b) Cent vingt ans à compter de la date de naissance pour les dossiers de personnel ;*
>
> *c) Cent ans à compter de la date de l'acte ou de la clôture du dossier pour les documents relatifs aux affaires portées devant les juridictions, y compris les décisions de grâce, pour les minutes et répertoires des notaires ainsi que pour les registres de l'état civil et de l'enregistrement ;*
>
> *d) Cent ans à compter de la date du recensement ou de l'enquête, pour les documents contenant des rensei-*

gnements individuels ayant trait à la vie personnelle et familiale et, d'une manière générale, aux faits et comportements d'ordre privé, collectés dans le cadre des enquêtes statistiques des services publics ;

e) Soixante ans à compter de la date de l'acte pour les documents qui contiennent des informations mettant en cause la vie privée ou intéressant la sûreté de l'État ou la défense nationale, et dont la liste est fixée par décret en Conseil d'État. »

POUR LES ARCHIVES PRIVEES

Article 213-6. Lorsque l'État et les collectivités locales reçoivent des archives privées à titre de don, de legs, de cession, de dépôt révocable ou de dation au titre de l'article 1131 et du I de l'article 1716 bis du code général des impôts, les administrations dépositaires sont tenues de *respecter les conditions de conservation et de communication qui peuvent être mises par les propriétaires.*

Leur consultation est donc définie par le propriétaire et spécifiée dans le contrat de don ou dépôt.

Cas particulier : Les Archives privées présentant pour des raisons historiques un intérêt public peuvent être classées comme **archives** *historiques*, sur proposition de l'administration des archives, par arrêté du Ministre chargé de la culture (art. L 212-15). L'article L 212-20 spécifie que « les archives classées comme archives historiques sont imprescriptibles » mais, article 12 que « le classement de documents comme archives historiques n'emporte pas transfert à l'État de la propriété des documents classés » (art. L 212-16).

Enfin, « toute destruction d'archives classées est interdite », article L 212-27/a.

PLACE DES ARCHIVES ORALES AU SEIN DES ARCHIVES NATIONALES ET DES SERVICES D'ARCHIVES DEPARTEMENTAUX ET MUNICIPAUX

L'article premier de la loi sur les archives ne fait pas de distinction par support ou par domaine. Les corpus oraux enregistrés sur support audio ou vidéo ne constituent donc pas une catégorie à part. Ils peuvent selon leur mode de production être des archives publiques ou des archives privées.

Les modes d'intégration dans les collections, peuvent se faire de façon passive ou active :

- L'institution reçoit les versements des administrations dans le cadre de l'exercice de la loi mais c'est le détenteur d'archives orales privées qui, seul, prend l'initiative du versement. Ce dernier a le choix entre le don, le legs, le dépôt révocable, de la cession de droit.
  C'est lui qui décide des conditions de consultation. S'il ne manifeste aucune volonté particulière, les règles des archives publiques seront appliquées aux archives orales privées.
- Le service d'archives peut prendre l'initiative d'un programme de collectes et produire, pour compléter ou se substituer à des archives absentes, des enregistrements de type interviews, témoignages, récits de vie.

MUSEES DE FRANCE : LEGISLATION

Le Code du Patrimoine consacre son Livre IV aux Musées pour lesquels la loi n° 2002-5 du 4 janvier 2002 a créé l'appellation « musées de France » :

> *« Article Premier. – L'appellation « musée de France » peut être accordée aux musées appartenant à l'État, à une autre personne morale de droit public ou à une personne morale de droit privé à but non lucratif ».*
>
> *Est définit « comme musée, au sens du présent livre, toute collection permanente composée de biens dont la conservation et la présentation revêtent un intérêt public et organisé en vue de la connaissance, de l'éducation et du plaisir du public. »*

Les « musées de France » ont pour missions permanentes de :

- Conserver, restaurer, étudier et enrichir leurs collections ;
- Rendre leurs collections accessibles au public le plus large ;
- Concevoir et mettre en œuvre des actions d'éducation et de diffusion visant à assurer l'égal accès de tous à la culture ;
- Contribuer aux progrès de la connaissance et de la recherche ainsi qu'à leur diffusion.

L'application de la loi passe par l'instauration d'un Haut Conseil des musées de France défini à l'article 3. Cette appellation peut être retirée. Si les collections des musées de France sont imprescriptibles, elles doivent, avant leur inscription sur l'inventaire des musées, recevoir l'avis scientifique de commissions spécifiques.

Les textes, la loi et les décrets et arrêtés pris pour l'application de la loi 2002-5 du 4 janvier 2002, favorisent l'organisation de réseau et une politique de dépôts d'œuvres d'un musée à l'autre. Les Directions Régionales des Affaires Culturelles (DRAC) sont chargées de veiller, en région, au contrôle technique de l'application des textes.

Les musées, régis antérieurement par l'Ordonnance de 1949, de par leur contenu et leur mode d'organisation sont d'une infinie variété. Entre l'établissement public du Louvre et un écomusée, pionnier mais de taille modeste, comme celui de la Roudoule dans les Alpes-Maritimes, peu de ressemblance, si ce n'est qu'il s'agit d'un de musée de France dans les deux cas.

INATHEQUE DE FRANCE

Sources de mémoire

L'INSTITUT NATIONAL DE L'AUDIOVISUEL

Créé en 1975, l'Ina est un établissement public à caractère industriel et commercial, chargé de conserver et exploiter le patrimoine audiovisuel français.

L'INATHEQUE DE FRANCE

La loi du 20 juin 1992 instituant un dépôt légal pour la radio et la télévision, représente une date essentielle dans l'histoire de l'audiovisuel français. Pour la première fois, à travers cette loi, l'audiovisuel, tout comme l'écrit, est considéré comme une source majeure d'archives et de mémoire.

Pour mettre en œuvre cette nouvelle mission, l'Ina crée, le 1er janvier 1995, l'Inathèque de France.

SES MISSIONS :

- Assurer la constitution et la conservation du patrimoine audiovisuel national.
- Organiser la consultation des œuvres et documents à des fins de recherche.
- Publier la bibliographie exhaustive des documents conservés au titre du Dépôt Légal.
- Favoriser la production et la diffusion des savoirs sur les images, les sons et les médias afin d'enrichir le débat public.

CONSERVATION ET ENRICHISSEMENT DOCUMENTAIRE

Ce sont 45 chaines de télévision et 17 diffuseurs radio qui sont suivis 365 jours par an.

Ce seront à terme 100 chaines collectées.

Chaque année 380 000 heures de programmes de télévision et 150 000 heures de radio sont identifiées et cataloguées dans les bases de données de l'Ina.

70 000 émissions de télévision et de radio font l'objet d'une description de contenu, sous la forme de mots-clés et de résumés, complétée par tout autre élément d'information nécessaire à l'exploitation de ces documents par les chercheurs.

CONSULTATION

L'Inathèque de France accueille les étudiants et les chercheurs dans son Centre de consultation situé au rez-de-jardin de la Bibliothèque nationale de France.

Le centre dispose de 56 places équipées d'un poste de consultation multimédia (SLAV : Station de Lecture AudioVisuelle) qui permet à la fois la consultation des bases de données de l'Ina, la gestion de corpus de travail, l'écoute ou le visionnage des émissions, leur analyse à l'aide d'outils adaptés.

Plus d'un million d'heures de télévision et de radio sont consultables.

La consultation s'exécute dans le respect du Code de la Propriété intellectuelle et artistique de sorte qu'aucune copie des enregistrements ne peut être effectuée, même à des fins pédagogiques et universitaires.

PROGRAMME « ARCHIVAGE » DU LACITO

Le LACITO (Laboratoire de Langues et Civilisations à Tradition Orale) est un laboratoire du CNRS dont les chercheurs (linguistes, anthropologues et ethno-musicologues) travaillent depuis plus d'une trentaine d'années à la description de langues pour la plupart sans écriture. De leurs enquêtes de terrain, ils ramènent des enregistrements audio, plus rarement vidéo, ainsi que des transcriptions, des traductions, etc. faites sur place avec l'aide de locuteurs. Ces enregistrements et analyses constituent les matériaux de base qui vont servir aux chercheurs pour poursuivre leurs recherches au retour de leur mission.

Le chercheur durant son enquête sera amené à expliquer les buts de sa mission, et tentera d'instaurer une « relation de confiance » entre lui et ses informateurs. Cette confiance est d'autant plus importante que les chercheurs sont parfois amenés à faire d'autres missions sur le même terrain. Elle peut être difficile à obtenir et facile à perdre, y compris par l'intervention ultérieure d'autres catégories d'enquêteurs (missionnaires, etc.) auxquelles les enquêtés risquent d'assimiler le chercheur.

L'*information préalable*, tout comme la *demande d'autorisation*, sont en général, compte tenu de la nature des cultures étudiées, faites sous un mode oral. Dans la pratique des enquêtes, ce n'est que très récemment que les chercheurs se préoccupent de garder une trace de cette autorisation (par exemple sous la forme d'un enregistrement audio). Ce qui prévalait jusqu'à présent, et qui prévaut encore aujourd'hui, c'est la relation de confiance qui lie enquêteurs et enquêtés.

L'information donnée par les chercheurs sur l'utilisation des enregistrements dépend fortement du niveau de culture de leurs interlocuteurs. Il est en effet parfois difficile de faire comprendre les implications de la mise à disposition d'un enregistrement audio sur la Toile à des personnes qui n'ont jamais entendu parler de l'informatique et qui n'ont jamais vu d'ordinateur. De plus, pour les enquêtes pratiquées il y a trente ans ou plus, aucun des chercheurs n'imaginait à l'époque les nouvelles utilisations qu'il pourrait faire de ses données. Dans ces conditions, bien sûr, l'information préalable ne pouvait être complète. Par ailleurs, le suivi des informateurs n'est pas aisé dans tous les pays et retrouver des locuteurs afin de les informer des changements de *finalité* n'est pas toujours possible.

Les enregistrements, jusqu'à récemment, servaient principalement aux chercheurs qui les avaient collectés. Des copies pouvaient en être faites pour des collègues, mais il n'existait ni catalogue, ni organisation pour le stockage, la conservation et la copie. Quand un chercheur disparaissait, toutes ses données accumulées risquaient donc de disparaitre avec lui.

Vers la fin de années 90, un programme s'est mis en place au LACITO pour lutter contre la disparition des données d'enquête. Dans ce programme « Archivage », les enregistrements analogiques sont numérisés afin de les préserver du vieillissement. Ils sont aussi catalogués afin de pouvoir les retrouver. Les notes de terrain (transcription, traductions, etc.) sont elles aussi numérisées et cataloguées. Enfin, des liens sont établis dans le catalogue pour ne pas perdre la relation qui existe entre enregistrements et notes de terrain.

Le programme « Archivage » a deux buts principaux qui sont :

- la préservation et la pérennisation des données d'enquête,
- leur diffusion.

La préservation est assurée par la numérisation des sources. Celle-ci se fait en utilisant des formats et des codages ouverts et libres. Les enregistrements sont numérisés sans compression en qualité CD-Audio dans un format WAV. Les notes de terrain sont structurées avec le langage de balisage de texte XML/Unicode en utilisant une syntaxe inspirée de la TEI. Les transcriptions sont codées la plupart du temps avec l'Alphabet Phonétique International. L'ensemble de ces ressources (fichiers audios et fichiers d'annotations) sont cataloguées au sein d'un document XML. Chacune d'elles est décrite à l'aide de métadonnées codées avec des étiquettes Dublin-Core suivant les spécifications préconisées par OLAC. Les ressources sont stockées sur des CD-ROM (un contrat est actuellement en discussion pour que la BnF soit le conservateur de ces données), elles sont aussi stockées sur un serveur où elles sont régulièrement recopiées sur des supports de sauvegarde.

La diffusion est assurée par un site Internet qui héberge à ce jour quelques 130 documents dans une trentaine de langues (principalement des langues de Nouvelle-Calédonie, du Népal et du Caucase). Une interface de consultation a été définie afin de consulter de manière synchronisée les documents d'enregistrement et leurs annotations. L'accès à l'ensemble de ces données se fait en général par la consultation du catalogue. Cela peut se faire localement en utilisant l'interface du site des archives, soit en consultant des moteurs de recherche spécialisés, puisque l'archive est une « archive ouverte », c'est-à-dire qu'elle utilise le protocole OAI-PMH pour communiquer avec l'extérieur. Tous les outils de consultation comme ceux qui ont été développés pour la création et la diffusion des ressources sont des *logiciels libres*.

CLAPI

Corpus de Langue Parlée en Interaction, laboratoire ICAR

Depuis trente ans, le laboratoire ICAR (ex-GRIC) (UMR 5191 du CNRS) mène à Lyon des recherches sur les interactions. En s'appuyant sur cette longue tradition, le laboratoire a développé une banque de données de Corpus de Langue Parlée en Interaction, CLAPI, dans le but d'assurer la sauvegarde et la gestion des corpus anciennement produits dans le laboratoire et de stimuler la production de nouveaux corpus en accord avec les exigences théoriques et technologiques du laboratoire actuel.

La base CLAPI compte en octobre 2005 :

- 600 h d'enregistrements audio et en partie vidéo, dont 350 h numérisées (2,5 Mo de mots) ;
- 20 h de transcriptions alignées avec le signal sonore et au format XML. (125.000 mots) ;
- 70 corpus (dont 35 décrits par la fiche de métadonnées) ;
- corpus d'interactions dans des situations sociales très variées (de la conversation quotidienne à des activités très spécifiques de travail, ou à des situations institutionnelles diversifiées) ;
- corpus de français et d'autres langues (comme p.ex. les langues régionales et l'arabe) et de situations natifs/non-natifs.

La base CLAPI ne se limite pas à accueillir des corpus ; elle est avant tout fondée sur un savoir-faire développé par une équipe, notamment dans les domaines suivants :

- *Le terrain* : le recueil de données en situation « naturelle » repose sur une approche ethnographique qui prépare les enregistrements et permet d'identifier les lieux et les moments les plus pertinents et propices à la constitution du corpus. Celle-ci ne se limite donc pas à une simple « capture » audio ou vidéo, même si l'enregistrement qui en est issu constitue le fondement du travail ultérieur de traitement et d'analyse.
- *L'enregistrement des données* : les enregistrements - audio et vidéo - visent des situations sociales d'interaction très diverses, recueillies dans un cadre dit « naturaliste », au sens de non orchestré et non provoqué par le chercheur. Des dispositifs d'enregistrement ont été développés qui concilient la capture multi-source et la préservation de la « naturalité » de l'interaction. En outre sont recueillis les documents, artefacts et autres objets manipulés ou produits pendant l'interaction.
- *La transcription* : la convention ICOR a été élaborée, qui assure la représentation, à différents niveaux de granularité, des phénomènes spécifiques à l'oral en interaction tout en garantissant l'homogénéité nécessaire à l'exploitation automatique et à l'interopérabilité ; la convention ICOR se limite pour l'instant aux phénomènes verbaux : la notation du multimodal est à l'étude.
- *L'identification des corpus et les métadonnées*: pour CLAPI a été mis au point un ensemble fonctionnel de descripteurs adaptés aux corpus de LPI (75 rubriques).
- *Les dimensions juridique, déontologique et éthique* : CLAPI a été l'occasion de concevoir des documents juridiques (autorisation de

recueil et de diffusion, convention de dépôt dans la base CLAPI, charte d'hébergement, convention de prêt) en application des dispositions relatives à la protection de la vie privée, à la propriété intellectuelle et au droit des bases de données ; ainsi que d'instaurer des pratiques d'anonymisation sur le signal, les transcriptions et le contenu des descripteurs.

- *L'intégration de corpus dans la base* : la base CLAPI a été conçue pour accueillir aussi bien des corpus anciens que récents, pour sauvegarder des corpus à valeur patrimoniale et historique comme pour inspirer la création de nouveaux corpus répondant aux exigences contemporaines sur le plan technique et scientifique.
- *L'hébergement sécurisé* : un système d'accès informatisés différenciés a été mis au point pour gérer les consultations et les interventions au sein de la base.
- *L'accès aux corpus* : CLAPI a stimulé les expériences sur la négociation de la diffusion des données interactionnelles, et les moyens humains qu'elle requiert, dans le respect des contraintes légales et en accord avec les auteurs des corpus. ICAR a pris l'option de rendre interrogeables en ligne librement, par les outils de la plate-forme, des extraits de corpus choisis par leurs auteurs (en octobre 2005, 15 extraits de corpus soit 3h30, dont 2h15 avec signal).
- *La diffusion de ces savoir-faires* : autour de CLAPI a été mise sur pied l'organisation de journées d'études, de formations internes/externes et aussi des propositions d'assistance technique et d'expertise.
- La conception et les développements d'outils de traitement et d'analyse des corpus :
  - traduction des transcriptions originales, quelle que soit la convention utilisée, au format XML, pour offrir l'homogénéité nécessaire aux analyses automatiques (20h à ce jour),
  - reconnaissance automatique des variantes graphiques d'une même forme générées par l'usage de « l'orthographe adaptée »,
  - phénomènes modélisés à ce jour : productions verbales/tours de parole et leur attribution aux locuteurs, timing, pauses (courtes, longues, quantifiées), chevauchements, tokens/ formes, commentaires/observations),
  - concordancier avec alignement texte/signal,
  - recherche de co-occurrences,
  - interface graphique permettant d'effectuer des requêtes personnalisées multi-critères qui combinent descripteurs et phénomènes,
  - bilan quantitatif des phénomènes par transcription,
  - repérage automatique des répétitions dans une transcription.

La plate-forme CLAPI est consultable à l'adresse suivante : http://clapi.univ-lyon2.fr

PFC

Le projet Phonologie du Français Contemporain (PFC) vise à constituer un vaste corpus de phonologie du français contemporain. Il a démarré sur l'initiative conjointe de Jacques Durand (ERSS, CNRS – Université de Toulouse le Mirail), Bernard Laks (MoDyCo, CNRS – Université Paris X Nanterre), Chantal Lyche (Université d'Oslo). À terme, PFC constituera la plus grosse base de données orales portant sur le français et l'une des plus grosses bases toutes langues confondues.

Le projet part de la constatation qu'il est nécessaire de poursuivre le travail de description entrepris depuis au moins un siècle par tous les spécialistes de la communication parlée pour :

- fournir une meilleure image du français parlé dans son unité et sa diversité, dans la réalité de ses usages attestés et dans sa diversité géographique, sociale et stylistique ;
- mettre à l'épreuve les modèles phonologiques et phonétiques sur le plan synchronique et diachronique ;
- favoriser les échanges entre les connaissances phonologiques et les outils du traitement automatique de la parole ;
- permettre la conservation d'une partie importante du patrimoine linguistique du monde francophone, et ce en contrepoint aux corpus déjà constitués ;
- permettre la constitution de meilleurs matériaux pédagogiques pour la description du français.

À partir d'un protocole d'enquête uniforme et en prenant appui sur des méthodes d'analyse et des outils développés en commun, le projet a pour ambition d'offrir une vision globale et unitaire de la phonologie du français contemporain, dont les principales caractéristiques sont :

- la diversité socio-géographique : une cinquantaine de points d'enquête prévus dans l'espace francophone à partir de groupes issus de réseaux denses. À chaque point d'enquête, 10 locuteurs en moyenne sont interviewés selon un protocole unique et constant ;
- la diversité de registres : 4 situations prises en compte dans le protocole ;
- la diversité des phénomènes phonologiques envisagés (inventaires phonologiques, schwa, liaison…) ;
- l'importance quantitative : environ 500 locuteurs, soit entre 800 et 1 000 h d'enregistrements, dont une centaine de locuteurs prévus pour la Belgique et les autres pays francophones.

En même temps qu'il offre à chaque chercheur la possibilité de suivre ses propres hypothèses et de construire ses propres objets sur la base sonore numérisée, PFC offre un ensemble de préanalyses particulièrement utiles en fournissant des transcriptions stables et vérifiées. Sur les fichiers sonores de ce corpus, est effectué un travail de transcription et d'alignement du texte sur le signal. Les principes de transcription s'inspirent des travaux et des expériences antérieures dans la transcription de gros corpus de français parlé (GARS à Aix-Marseille, VALIBEL à Louvain-la-Neuve), notamment pour la gestion des tours de parole, la transcription des pauses, des reprises, des incises et des erreurs.

S'agissant du schwa (le e muet) et des phénomènes de liaison, PFC offre en plus un ensemble d'analyses finales quantifiées avec un détail et une précision

jamais atteintes : pour chaque locuteur, codage complet et exhaustif, aligné sur le signal, de 3 et 5 minutes de parole en conversation guidée et libre, ainsi que de la lecture de deux éléments codifiées (un texte et une liste de mots).

Le traitement de ces données est basé sur les dernières avancées de la recherche en phonologie, phonétique et sociolinguistique. La structure informatique les stockant et les diffusant se situe, dans la problématique actuelle de diffusion de contenus sur Internet (particulièrement le langage de structuration XML et l'interopérabilité des données).

La difficulté spécifique d'un corpus sonore est, à l'heure actuelle, de pouvoir faire des liens entre une demande documentaire (les locuteurs du même âge ou de la même enquête) et telle partie d'un fichier sonore : trouver les « pointeurs » qui permettent, à partir d'un descripteur, formulé par un ou plusieurs termes, de rentrer à un endroit précis du fichier sonore. C'est un des enjeux du projet PFC : permettre la consultation d'un vaste corpus sonore, plus développé qu'une simple lecture des données.

PFC propose alors une structure de consultation des données recueillies et homogénéisées, via les protocoles Internet. Une base de données fortement structurée et relationnelle est ainsi accessible avec un simple navigateur. L'interface d'interrogation permet des requêtes larges et fines sur ces données avec un croisement inédit entre les données documentaires textuelles et les données sonores numérisées.

Notre objectif majeur est de construire un corpus favorisant différents niveaux d'approche, adapté à différents publics (étudiants, enseignants, chercheurs, ingénieurs). La variété des exploitations possibles est très grande grâce à la mise à disposition d'une ressource à la masse critique importante et aux données standardisées et donc interopérables. L'enseignant ayant besoin d'un tutoriel comparatif de français oral pour des publics, même jeunes, comme l'ingénieur devant construire un système de reconnaissance vocale, pourront se baser utilement sur cette ressource.

http://www.projet-pfc.net

DELIC

L'équipe DELIC (DEscription Linguistique sur Corpus), créée en 1999, a développé un projet qui s'appuie notamment sur l'élaboration et l'exploitation morphosyntaxique de corpus oraux (et aussi écrits). Elle a hérité du corpus du GARS (Groupe Aixois de Recherches en Syntaxe), ce qui l'a confrontée aux difficultés que soulève la récupération de corpus un peu anciens et a développé divers projets de constitution de nouveaux corpus. Cette présentation de l'équipe sera centrée sur les problèmes rencontrés pour témoigner de l'expérience acquise.

LA CONSTITUTION DES CORPUS

La partie ancienne (le corpus du GARS) a été développée pendant une vingtaine d'années[86] et compte 1 700 000 mots restaurés. En l'espace de 20 ans, de nombreux changements sont intervenus (modification des supports d'enregistrement, variations dans les conventions, sensibilisation aux problèmes juridiques, etc.). Tout un travail de restauration a donc été nécessaire.

Les nouveaux corpus tiennent bien évidemment compte de l'expérience acquise :

- les enregistrements sont effectués sur MD (minidisque), puis le son est numérisé ;
- le matériel d'enregistrement (MD et micro) permet de disposer, en général, d'enregistrements d'une bien meilleure qualité ;
- des autorisations sont remplies pour chaque nouvel enregistrement ;
- les conventions ont été revues pour écarter quelques phénomènes qui n'étaient pas utilisés dans les analyses (par exemple, les allongements ne sont plus notés) et restent stables depuis quelques années.

L'équipe a aussi pu bénéficier de l'expérience des membres du GARS qui étaient avertis de certains problèmes liés au recueil des données et elle continue à former des étudiants pour collecter et transcrire les enregistrements.

Pour les transcriptions, deux techniques sont utilisées : soit « à l'ancienne » avec écouteur et papier (ou saisie sur clavier), soit à l'aide du logiciel Transcriber (disponible sur le Net). Les transcriptions sont ensuite vérifiées par des personnes averties. Ce travail d'édition qui demande beaucoup de soin et de temps est indispensable.

LES PROBLEMES DE CONSERVATION

Pour la partie ancienne du corpus, les enregistrements sur cassettes sont quelquefois d'une qualité médiocre. Un certain nombre d'entre eux ont été numérisés, mais ce travail demande un investissement très lourd. D'autre part, dans certains cas, les enregistrements ont été égarés, et pour un grand

[86] Blanche-Benveniste, Cl. (2000) « Corpus de français parlé » dans Bilger ed. *Corpus Méthodologie et applications linguistiques*, rappelle quelques aspects du développement de ce corpus.

nombre de transcriptions il a fallu scanner les textes qui avaient été saisis avec des machines à écrire.

Tout cela a montré l'importance des tâches de gestion et de classement des archives. Une personne de l'équipe s'est spécialisée dans cette activité pour les nouveaux corpus afin de faire correspondre rapidement les divers documents qui doivent être reliés : enregistrement, fiche signalétique, transcription.

Pour le CRFP[87] (Corpus de Référence du Français Parlé) une grande énergie a été engagée dans les problèmes de gestion de ces divers documents et a montré la nécessité d'une organisation conséquente pour éviter (ou du moins limiter) la perte de certains documents (des fiches signalétiques), le travail inutile (corriger la version antérieure d'une transcription), etc.

Les enregistrements existent sous plusieurs formats (MD et fichiers numérisés) ce qui garantit, en partie, leur pérennité.

L'EXPLOITATION

Pour les corpus les plus récents, on procède ensuite à un alignement texte/son (pour lequel on utilise Transcriber). Ce traitement permet d'améliorer la qualité des transcriptions. D'autre part, les corpus peuvent être exploités à l'aide du logiciel *Contextes* réalisé par Jean Véronis. Dans sa dernière version, ce concordancier permet aussi d'écouter le passage sélectionné.

L'équipe, qui regroupe des linguistes et des informaticiens, développe des projets de description morphosysntaxique et d'analyse semi-automatique sur de gros corpus (écrits et oraux).

LA DIFFUSION

L'équipe possède actuellement plusieurs corpus dont on vient de rappeler qu'ils se rattachent à des projets distincts. Ils présentent des caractéristiques différentes (qualité des enregistrements, autorisation, etc.).

- CorpAix (le corpus du GARS) 1 700 000 mots
- CRFP 460 000 mots
- C-ORAL-ROM88 232 000 mots
- Corpus DELIC (en développement depuis 2000) 560 000 mots
- CRFP-2 (projet en cours) : enregistrement du français des médias.

Pour les plus anciens (CorpAix), la diffusion de passages longs est exclue, car ils ont été constitués en dehors d'un cadre juridique (pas d'autorisation). Le CRFP est consultable sous forme d'extraits sur le net (http://www.up.univ-mrs.fr/delic/). Un certain nombre d'autorisations manquent et bloquent sa diffusion. Le corpus C-ORAL-ROM est lui accessible par le biais de l'édition signalée.

Les résultats des analyses conduites sur les divers corpus mentionnés sont diffusés dans les publications des membres de l'équipe.

[87] Ce corpus est présenté dans *Recherche sur le Français Parlé* 18 (2004) Université de Provence.

[88] Ce projet européen qui porte sur 4 langues romanes est présenté dans Cresti, E. & Moneglia, M. *éd.* (2005) *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam, John Benjamins.

ESLO

LES ENQUETES SOCIO-LINGUISTIQUES A ORLEANS, 1968-2008

L'enquête ESLO (Enquête Socio-Linguistique à Orléans), conduite par des universitaires britanniques à des fins didactiques (enseignement du français langue étrangère dans le système public d'éducation anglais) en 1968, comprend environ 200 interviews, toutes référencées, et plus de 300 heures de parole incluant des enregistrements cachés, des conversations téléphoniques, des réunions publiques, des entretiens médico-pédagogiques, etc. Ce corpus constitue, par son ampleur et sa cohérence, le plus important témoignage sur le français parlé avant 1980.

Le premier objectif est de numériser les documents sonores à partir des enregistrements magnétiques et d'en proposer une indexation et un premier balisage afin de mettre les données en ligne sur Internet.

Parallèlement, une exploitation exhaustive d'un sous-ensemble est engagée. Partant de l'expérience acquise, le CORAL (Centre Orléanais de Recherche en Anthropologie et Linguistique) en partenariat avec d'autres laboratoires (CELITH-MODYCO) a mis en chantier une nouvelle enquête dénommée ESLO2. L'objectif est d'évaluer, à une quarantaine d'années de distance, la dynamique sociale du français (des usages de la langue comme des jugements sur son emploi). La prise en compte de la diversité des changements est rapportée aux paramètres sociaux, révélant l'inégalité des résistances ou des propensions à la transformation de la langue, mais aussi la typologie et la dynamique des évolutions.

Cette façon de procéder présente l'avantage de préfigurer la référence attendue dans un domaine qui en est encore à se structurer et dans lequel se manifeste de manière récurrente une demande de définition pour un format standardisé de *collecte*, de *conservation*, de *traitement* et d'*analyse* :

- la *collecte* sur le terrain est première, non seulement dans ses aspects techniques, aujourd'hui bien maîtrisés, mais aussi dans la définition du profil de l'échantillon représentatif et dans la problématisation des interactions entre les témoins et les enquêteurs ;
- la *conservation*, qui inclut la préservation des supports, l'indexation des contenus et l'accessibilité (c'est-à-dire la protection) des données, conditionne le partage des sources à des fins d'étude scientifique et d'expertise politique ;
- le *traitement*, en lien étroit avec le développement des matériels et des langages informatiques, suppose la maîtrise d'une chaîne d'opérations, depuis la conversion numérique des enregistrements jusqu'à une transcription balisée et ouverte à l'ensemble des interrogations pertinentes pour les demandes du linguiste, du sociologue ou des décideurs, des didacticiens voire du grand public ;
- l'*analyse* constitue l'épreuve des théories (et des logiciels) puisqu'elle compare les formalisations et les opérations et qu'elle valide ou infirme les hypothèses en prenant argument de leur compatibilité aux faits.

Avec la constitution et la comparaison de telles enquêtes, les politiques et les acteurs de la transmission linguistique ont à leur disposition un outil d'aide à la décision irremplaçable, qui permet d'appréhender, aussi objectivement que possible, le devenir du français parlé dans toutes ses dimensions (phonologique et prosodique, lexicale et syntaxique, sémantique et pragmatique). La définition d'un standard rigoureux et réaliste devrait orienter les descriptions du français parlé en France au service de la recherche, des applications et de l'expertise.

INVENTAIRE DES CORPUS

INVENTAIRE DE LA DGLFLF

L'aventure du *Guide des bonnes pratiques* a mis en lumière la nécessité de disposer d'une meilleure vision des corpus de langue française qui existent en France et à l'étranger. La DGLFLF a donc piloté un inventaire qui fournit diverses indications dont :

- le nom du corpus ;
- le responsable ;
- la taille, le contenu, l'état de ces données (supports utilisé, etc.) ;
- le type d'accès possible (accès libre, partiel, limité à une équipe, etc.).

La présentation de l'inventaire reprend et développe ces différents paramètres.

Cet inventaire (qui peut être téléchargé sur le site de la DGLFLF) pourrait faciliter les contacts et les échanges entre équipes, permettre d'identifier les manques les plus flagrants dans le domaine des données orales constituées et aider les futurs projets de constitution de grandes banques de données à mieux cerner les forces disponibles et les besoins.

En l'état, le lecteur dispose d'un état des lieux (partiel à cause des oublis) qui peut être complété en fournissant toute information utile à :

Paul.Cappeau@univ-poitiers.fr.

www.dglflf.culture.gouv.fr

BIBLIOGRAPHIE

BIBLIOGRAPHIE

BIBLIOGRAPHIE GENERALE

ABOU-HAIDIR, L., dir. (2002) « Transcription de la parole normale et pathologique », *Revue Parole* 22/23/24.

ACHARD, P. (1991) « Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie », *Langage et société* 55 : 5-40.

ADLER, P.A. (1987) *Membership Rules in Field Research*, Sage, Newbury Park.

AIJMER, K. & ALTENBERG, B. eds (1992) *English Corpus Linguistics. Studies in honour of Jan Svartvik,* London/New-York, Longman.

ATKINSON, J.M. & HERITAGE, J. eds (1984) *Structures of Social Action*, Cambridge, CUP.

AUER, P. et al. (1999) *Language in Time. The Rhythm and Tempo of Spoken Interaction*, Oxford, OUP.

BANGE, P. (1983) « Points de vue sur l'analyse conversationnelle », *DRLAV* 29 : 1-28.

BARRAS, C., ADDA, G., ADDA-DECKER, M., HABERT, B., BOULA DE MAREÜIL, P. & PAROUBEK, P. (2004) « Automatic audio and manual transcripts alignment, time-code transfer and selection of exact transcripts », *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne : 877-880.

BAUDE, O. (2004) « Les corpus oraux entre science et patrimoine. L'expérience de l'observatoire des pratiques linguistiques », *Actes du Colloque international du GRESEC « La publicisation de la science »* (Grenoble) : 7-11.

BEAUD, S. & WEBER, F. (1997) *Guide de l'enquête de terrain : produire et analyser des données ethnographiques*, Paris, La Découverte.

BECKER, H. S. & GEER, B. (1960) « Participant observation : the analysis of qualitative field data », ADAMS & PREISS eds : 267-289.

BERGOUNIOUX, G. dir. (1992) « Enquêtes, Corpus et Témoins », *Langue Française* 93.

BIBER, D. (1985) *Variations across spoken and written language*, Cambridge, CUP.

BIBER, D. (1999) *Longman Grammar of Spoken and Written English*, Londres, Longman.

BILGER, M. dir. (2000) « Linguistique sur corpus, études et réflexions », *Cahiers de l'université de Perpignan*, Perpignan, Presses universitaires.

BILGER, M. ed. (2000) *Corpus, Méthodologie et applications linguistiques*, Paris, Champion.

BLANCHE-BENVENISTE, Cl. & JEANJEAN, C. (1987) *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

BLANCHE-BENVENISTE, Cl. (1997) « Transcription et technologie », *Recherches sur le Français Parlé* 14 : 87-100.

BLANCHE-BENVENISTE, Cl. BILGER, M., ROUGET, C. & VAN DEN EYNDE, K. (1999) *Le Français Parlé : Études grammaticales*, Paris, CNRS-Editions.

BLANCHE-BENVENISTE, Cl., ROUGET, C. & SABIO, F. (2001) *Choix de textes de français parlé : trente-six extraits*, Paris, Champion.

BOURDIEU, P. (1982) *Ce que parler veut dire. L'économie des échanges linguistiques*, Paris, Fayard.

BOURDIEU, P. (1993) *La misère du monde*, Paris, Seuil.

BÜRKI, Y. & DE STEFANI, E *ed.* (à paraître), *Transcryptio*, Berne, Peter Lang.

CAMERON, D., FRAZER, E., HARVEY, P., RAMPTON, M. & RICHARDSON, K. (1991) *Researching Language : Issues of Power and Method*, London, Routledge.

CLIFFORD, J. & MARCUS, G. E. eds (1986) *Writing Culture. The Poetics and Politics of Ethnography,* Berkeley, University of California Press.

CONDAMINE, A. *ed.* (2006) *Sémantique et corpus*, Paris, Hermes.

COUPER-KUHLEN, E. & SELTING, M. *ed.* (1996) *Prosody in Conversation : Interactional Studies*, Cambridge, CUP.

CRESTI, E. & MONEGLIA, M. ed. (2005) *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphie, Benjamins.

CRIBIER, F. & FELLER, E. (2003) *Projet de conservation des données qualitatives des sciences sociales recueillies en France auprès de la « société civile »* rapport présenté à Madame la Ministre déléguée à la Recherche et aux nouvelles technologies,

dactylogr. 2 vol.

et http://www.iresco.fr/labos/lasmas/rapport/Rapdonneesqualita.pdf

DEPPERMANN, A. (2000) « Ethnographische Gesprächsanalyse : zur Nutzen und Notwendigkeit von Ethnographie für die Konversationsanalyse », *Gesprächsforschung* 1 : 96-124.

DURANTI, A. (1997) *Linguistic Anthropology,* Cambridge, CUP.

ENCREVE, P., & FORNEL de, M. (1983) « Le sens en pratique », *ARSS* 46, L'usage de la parole.

GADET, F. (2003) *La variation sociale en français*, Paris, Ophrys.

GUILHAUMOU, J., MESINI, B. & PELEN, J.-N. (1997) Récits de vie. « Dynamiques et autonomies des récits de vie dans le champ de l'"exclusion" ». *Cahiers de littérature orale* 41 : 91-126.

GUMPERZ, J. J., & HYMES, D. eds (1972) *Directions in Sociolinguistics : The Ethnography of Communication,* New-York, Hold, Rinehart & Winston.

HABERT, B., NAZARENKO, A. & SALEM, A. (1997) *Les linguistiques de corpus*, Paris, A. Colin.

HAMMERSLEY, M. & ATKINSON, P. (1995) *Ethnography : Principles in Practice*, Londres, Routledge.

HEATH, C. (1997) « Analysing work activities in face to face interaction using video », Silverman *ed.*

HOUTKOOP-STEENSTRA, H. (2000) *Interaction and the Standardized Survey Interview*, Cambridge, CUP.

JACOBSON, M. (2004) « Corpus oraux en linguistique de terrain », *Traitement Automatique des Langues*, 45/2 : 63-88.

JACOBSON, M. (2004) « Les archives sonores au LACITO », *Bulletin de liaison de l'AFAS* 26 (http://afas.mmsh.univ-aix.fr/bulletin/Bulletin AFAS 26.pdf).

JEFFERSON, G. (1973) « A Case of Precision Timing in Ordinary Conversation : Overlapped Tag-Positioned Address Terms in Closing Sequences », *Semiotica* 9 : 47-96.

JEFFERSON, G. *ed.* (1983) « Issues in the transcription of naturally occurring talk : caricature versus capturing pronunciational particulars, Tilburg Papers », *Language and Literature* 34.

JEFFERSON, G. (1985) « An Exercise in the Transcription and Analysis of Laughter », T. van Dijk *ed.* : 25-34.

JEFFERSON, G. (1996) « A case of transcriptional stereotyping », *Journal of Pragmatics* 26/2 : 159-170.

JORDAN, B. & HENDERSON, A. (1995) « Interaction analysis : Foundations and practice », *The Journal of the Learning Sciences* 4/1 : 39-103.

KALLMEYER, W. & SCHÜTZE, F. (1976) « Konversationsanalyse », *Studium Linguistik* 1 : 1-28.

KENNEDY, G. (1998) *An introduction to Corpus Linguistics*, Londres, Longman.

KNOBLAUCH, H., RAAB, J., SOEFFNER, H.-G. & SCHNETTLER, B. *ed.* (2006) *Video analysis*, Berne, Peter Lang.

LABOV, W. (1972) *Sociolinguistic Patterns*, Philadelphie, University of Pennsylvania Press.

LEECH, G. (1992) « The state of the art in corpus linguistics », Aijmer & Altenberg eds : 8-29

MARTIN, Ph. (1987) « Prosodic and Rhythmic Structures in French », *Linguistics* : 925-949.

MAYNARD, D. W., HOUTKOOP-STEENSTRA, H., SCHAEFFER, N. C. & ZOUWEN, J. V. D. *eds.* (2002) *Standardization and Tacit Knowledge. Interaction and Practice in the Survey Interview*, New York, John Wiley.

MITCHELL, R. G. Jr (1991) « Secrecy and disclosure in fieldwork », Shaffir, W.B., Stebbins, R.A. eds : 207-222.

MOERMAN, M. (1988) *Talking Culture : Ethnography and Conversation Analysis*, Philadelphie, University of Pennsylvania Press.

MONDADA, L. (1998) « Technologies et interactions sur le terrain du linguiste. Le travail du chercheur sur le terrain. Questionner les pratiques, les méthodes, les techniques de l'enquête ». Actes du Colloque de Lausanne 13-14.12.1998, *Cahiers de l'ILSL* 10 : 39-68.

MONDADA, L. (2000) « Les effets théoriques des pratiques de transcription », *Linx* 42 : 131-150.

MONDADA, L. (2001) « Pour une linguistique interactionnelle », *Marges Linguistiques* 1, http://www.marges-linguistiques.com.

MONDADA, L. (2002) « Pratiques de transcription et effets de catégorisation », *Cahiers de Praxématique* 39 : 45-75.

MONDADA, L. (2003) « Observer les activités de la classe dans leur diversité : choix méthodologiques et enjeux théoriques », Perera, Nussbaum, Milian eds : 49-70.

MONDADA, L. (2006) « Video recording as the reflexive preservation-configuration of phenomenal features for analysis », Knoblauch, H., Raab, J., H.-G. Soeffner, Schnettler, B. eds.

MONDADA, L. (2006) « L'analyse de corpus dans la perspective de la linguistique interactionnelle : des analyses de cas singuliers aux analyses de collections », Condamine *ed.*

MONDADA, L. (à paraitre) « La demande d'autorisation comme moment structurant pour l'enregistrement et l'analyse des pratiques bilingues », *Tranel*, Université de Neuchâtel.

MONDADA, L. (à paraitre), « La pertinenza del dettaglio : registrazione e trascrizione di dati video per la linguistica internazionale », Bürki, E. de Stefani (à paraître).

NØLKE, H & ANDERSEN, H.L. *ed.* (2002) « Macro-syntaxe et macro-sémantique », *Actes du Colloque International d'Aarhus, mai 2001*, Berne, Peter Lang.

OCHS, E. (1979) « Transcription as theory », OCHS, E. & SCHIEFELLIN, B.B. (1979) : 43-72.

OCHS, E., & SCHIEFELLIN, B.B. *eds* (1979) *Developmental Pragmatics*, New-York, Academic Press.

OCHS, E., SCHEGLOFF, E. & THOMPSON, S. *eds.* (1996) *Interaction and Grammar*, Cambridge, CUP.

ONG, W. (1988) *Orality and Literacy*, Londres, Routledge.

PERERA, J., NUSSBAUM, L. & MILIAN, M. *ed.* (2003) *L'educacio linguistica en situacions multiculturals i multilingues,* Barcelone, ICE Universitat de Barcelona.

PLATT, J. (1983) « The development of the "participant observation" method in sociology : origin, myth, and history », *Journal of the History of the Behavioral Sciences* 19 : 379-393.

QUERE, L. *et al. ed.* (1984) *Arguments ethnométhodologiques*, Paris, Centre d'Étude des Mouvements Sociaux, EHESS.

Recherches sur le Français Parlé 5 (1984) « Pourquoi le français parlé est-il si peu étudié ? ».

Revue Française de Linguistique Appliquée (1996) 1-2, (1999) IV-1.

SACKS, H. (1972a) « An initial investigation of the usability of conversational materials for doing sociology », Sudnow *ed.* : 31-74.

SACKS, H. (1972b) « On the Analyzability of Stories by Children », Gumperz & Hymes *eds.* : 325-345.

SACKS, H. (1984) « Notes on methodology », J. M. Atkinson & J. Heritage *ed.* : 21-27.

SACKS, H. SCHEGLOFF, E.A., & JEFFERSON, G. (1974) « A simplest systematics for the organization of turn-taking for conversation », *Language* 50 : 696-735.

SACKS, H. (1992) *Lectures on Conversation* [1964-72] (2 Vol.) Oxford, Basil Blackwell.

SANKOFF, D., SANKOFF, G., LABERGE, S. & TOPHAM, M. (1976) « Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale », *Cahiers de Linguistique* 6 : 85-125.

SCARANO, A. ed. (2003) *Macro-syntaxe et pragmatique. L'analyse linguistique de l'oral*, Rome, Bulzoni editore.

SELTING, M. (1995) « Der "mögliche Satz" als interaktiv relevante syntaktische Kategorie », *Linguistische Berichte* 158 : 298-325.

SELTING, M. (1996) « On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation », *Pragmatics* 6 (3) : 371-389.

SELTING, M. (2000) « The construction of units in conversational talk », *Language in Society* 29 : 477-517.

SHAFFIR, W.B. & STEBBINS, R. A. eds. (1991) *Experiencing Fieldwork : An inside View of Qualitative Research*, Londres, Sage.

SILVERMAN, D. ed. (1997) *Qualitative Research. Theory Method and Practice*, Londres, Sage.

SINCLAIR, J. (1991) *Corpus, Concordance, Collocation*, Londres, OUP.

SINCLAIR, J. (1996) *Preliminary recommendations on corpus Typology*, Technical Report, EAGLES.

SINCLAIR, J. & COULTHARD, R. M. (1975) *Towards an Analysis of Discourse*, Londres, OUP.

« SPEECH ANNOTATION AND CORPUS TOOLS », A special issue of *Speech Communication* 33, 1-2 (2001) Steven Bird and Jonathan Harrington.

SPRADLEY, J. P. (1980) *Participant Observation*, New-York, Hold, Rinehart & Winston.

SUDNOW, D. *ed.* (1972) *Studies in Social Interaction,* New York, Free Press.

TEUBERT, W. (1999) « Corpus Linguistics. A Partisan View », *TELRI-Newsletter (Trans-European Language Ressources Infrastructures)* 8 : 4-19.

TIOUKA, A. (2005) « La question du droit autochtone sera-t-elle résolue en France ? » *Ethnies* 31-32.

TRAVERSO, V. (2002) « Transcription et traduction des interactions en langue étrangère », *Cahiers de Praxématique* 39 : 77-99.

VAN DER STRATEN (1998) « Remarques sur la transcription des enregistrements en vidéo », *CALAP* 18 : 161-177.

VAN DIJK, T. *ed.*, *Handbook of discourse Analysis,* Volume 3, New-York, Academic Press.

WELLAND, T. & PUGSLEY, L. *eds.* (2002), *Ethical Dilemmas in Qualitative Research*, Aldershot, Ashgate.

Bibliographie Patrimoine de l'oral et conservation

ARON-SCHNAPPER, D., HANET, D., DEWARTE, S. & PASQUIER, D. (1980) *Histoire orale ou archives orales ? Rapport d'activité sur la constitution d'archives orales pour l'histoire de la sécurité sociale*, Paris, Association pour l'étude de l'histoire de la Sécurité sociale.

CALLU, A. & LEMOINE, H. (2004) *Patrimoine sonore et audiovisuel français : entre archive et témoignage : guide de recherche en sciences sociales*, 7 vol., 1 CD-Rom, 1 DVD-Rom, Paris, Belin.

DESCHAMPS, F. (2001) *L'historien, l'archiviste et le magnétophone*, Paris, Comité pour l'histoire économique et financière de la France.

DOURNON, G. (1996) *Guide pour la collecte des musiques et instruments traditionnels*, Edition augmentée, Paris, UNESCO.

« Musique et son : les enjeux de l'ère numérique. Création musicale, recherche, archivage, transmission », (2002), *Culture et Recherche* 91-92.

DURAND, C. (1999-2000) *Folklore et droit d'auteur*, mémoire de DESS, Propriété intellectuelle et communication, Université Montesquieu-Bordeaux IV.

JOUTARD, P. (1979) « Historiens, à vos micros. Le document oral, une nouvelle source pour l'histoire », *L'Histoire* 12 : 106-113.

JOUTARD, P. (1983) *Ces voix qui nous parlent du passé*, Paris, Hachette.

NORA, P. dir. (1983) *Les lieux de mémoire*, Paris, Gallimard.

PROST, A. (1996) *Douze leçons sur l'histoire*, Paris, Seuil.

RICOEUR, P. (2000) *La Mémoire, l'histoire, l'oubli*, Paris, Seuil.

TOURTIER-BONAZZI (de), C. (1990) *Le témoignage oral aux archives...*, Paris, Archives nationales.

VOLDMAN, D. dir. (1992) « La Bouche de la vérité ? La recherche historique et les sources orales », *Les Cahiers de l'IHTP* 21.

VALLIERE M. (2002) *Ethnographie de la France : histoire et enjeux contemporains des approches du patrimoine ethnologique*, collection Cursus, Paris, Armand Colin.

Revues et periodiques

Bulletin de l'IHTP 1 (juin 1980) « Problèmes de méthode en histoire orale », table ronde de l'Institut d'Histoire du Temps Présent.

Bulletin de l'IHTP 75, (juin 2000), Danièle Voldman, « Le témoignage dans l'histoire du temps présent », Les Cahiers de l'IHTP (Institut d'Histoire du Temps Présent)

Sonorités, bulletin de l'AFAS, Association française des détenteurs de documents audiovisuels et sonores.

International Journal of Oral History

Aspects techniques

BONNEMASON, B., GINOUVES, V. & PERENNOU, V. (2001) *Guide d'analyse documentaire du son inédit pour la mise en place de banques de données*, Parthenay, Modal-AFAS.

CALAS, M.-F. & FONTAINE, J.-M. (1996), *La Conservation des documents sonores*, Paris, CNRS Editions.

GENDRE, C. (1999) *Enregistrement et conservation des documents sonores*, Paris, Eyrolles.

Pour la conservation des données numériques, voir les sites suivants :

Association française des détenteurs de documents audiovisuels et sonores (AFAS) :

http://afas.mmhs.univ-aix.fr/

Le compte rendu et les principales interventions du séminaire commun AFAS / BnF des 7 et 8 octobre 2004 portant sur : « La numérisation des archives sonores au service de la conservation : principes généraux et recommandations pratiques » sont consultables en ligne sur le site de l'Association.

Bibliothèque nationale de France :

http://bibnum.bnf.fr/conservation/infopreservation_fr.pdf

International Association of Sound and Audiovisual Archives :

http://www.iasa-web.org/

Voir notamment :

Bradley, K. dir., Guidelines on the production and preservation of digital objects. International Association of Sound and Audiovisual Archives. ISBN 8799030918 (voir sur le site Internet de l'Association).

Ministère de la Culture et de la Communication :

http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_04.htm

Références techniques sur la conservation :

Pickett et Lemcoe, Preservation and storage of sound recordings, Wahington, 1959.

Gilles Saint-Laurent, Care and handling of sound recordings :

http://palimpsest.stanford.edu/byauth/st-laurent/carefr.html

Cylinder, Disc and Tape Care in a Nutshell :

http://www.loc.gov/preserv/care/record.html

Équipement pour l'enregistrement de terrain :

http://www.vermontfolklifecenter.org/res_audioequip.htm.

Sur les techniques de prise de son et les matériels :

voir collections spécialisées chez Eyrolles et Dunod

Conseils sur le site de l'ASPPAC : www.asppac.com

Recommandations des Archives de France pour la gravure sur CD-R : http://www.archivesdefrance.culture.gouv.fr/fr/circAD/DITN.2005.004.recommandations.pdf

D'autres informations pratiques sur le CD (surtout pour qui n'a pas un puissant analyseur) : http://www.mrichter.com/cdr/primer/primer.htm

LANGUES EN DANGER : LIENS UTILES

Les organismes et les institutions finançant la recherche sur les langues en danger mènent des réflexions similaires à celle qui est proposée dans ce guide. Nous donnons ci-dessous à titre informatif quelques adresses de sites Internet :

http://www.unesco.org/culture/heritage/intangible/meetings/paris_march2003.shtml

[Site de l'UNESCO et page du colloque intitulé « *Safeguarding endangered Languages* »]

http://www.mpi.nl/DOBES/INFOpages/applicants/legal-ethics-issues.html

[Site du Max-Planck Institute, et du programme DOBES pour la description des langues en danger – recommandations légales]

http://www.eva.mpg.de/lingua/files/ethics.html

[Recommandations du Département de Linguistique du Max-Planck Institute for Evolutionary Anthropology].

http://sapir.ling.yale.edu/~elf/ethics.html

[Rapport du SALSA Special Colloquium sur *Archiving Language Materials in. Web-Accessible Databases: Ethical Challenges*, 22 avril 2001. By D. H. Whalen, President, Endangered Language Fund].

http://www.hrelp.org/

[Programme de financement de recherches sur les langues en danger de la SOAS (School of Oriental and African Studies), University of London].

http://www.ogmios.org/home.htm

[Site de la Fundation for Endangered Languages, dont la dernière conférence (octobre 2004) a eu pour thème : Endangered Languages and Linguistics Rights]

GLOSSAIRE JURIDIQUE

Sauf mention contraire, les citations sont conformes au Dictionnaire comparé du droit d'auteur et du copyright

Anonymisation :

Opération par laquelle se trouve supprimé d'un ensemble de données recueillies auprès d'un individu ou d'un groupe tout lien permettant l'identification de ces derniers (voir fiche *Données personnelles et anonymisation*)

Auteur :

« Personne physique qui crée l'œuvre. Investie à titre originaire des droits d'auteur quel que soit son statut (indépendant, salarié, etc.) et les circonstances dans lesquelles elle réalise l'œuvre. Seule titulaire du droit moral de son vivant ».

Creative commons :

Le « Creative Commons » est une organisation dévouée à l'expansion des œuvres qui sont libres à la réutilisation et/ou la distribution. C'est dans ce but qu'elle a créé la licence Creative Commons. Cette licence autorise certains usages librement définis par les auteurs, parmi onze possibilités combinées autour de quatre pôles : Attribution (signature de l'auteur initial) ; Commercial (possibilité de tirer profit commercial de l'œuvre) ; No derivative works (possibilité d'intégrer tout ou partie dans une œuvre composite/ samplage) ; Share alike (obligation de rediffuser selon la même licence). Symbole général : cc.

Le mouvement Creative Commons propose des contrats-types d'offre de mise à disposition d'œuvres en ligne. Inspirées par les licences de logiciel libre et le mouvement open source, ces textes facilitent l'utilisation et la réutilisation d'œuvres (textes, photos, musique, sites Internet...). Au lieu de soumettre toute exploitation des œuvres à l'autorisation préalable des titulaires de droits, les licences Creative Commons permettent à l'auteur d'autoriser à l'avance certaines utilisations selon des conditions exprimées par lui, et d'en informer le public.

L'objectif recherché est d'encourager de manière simple et licite la circulation des œuvres, l'échange et la créativité.

Domaine public :

« Sphère d'exploitation libre et gratuite des œuvres de l'esprit qui échappent au monopole de l'auteur lorsque le monopole d'exploitation est expiré. Comprend aussi les éléments de libre parcours qui ne donnent pas prise au droit d'auteur (idées, hypothèses scientifiques…) ».

Données personnelles :

(Loi du 6 aout 2004) Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne.

Droit d'auteur :

« Droit de propriété incorporelle exclusif et opposable à tous, qui comprend l'ensemble des prérogatives morales (*droit de divulgation, droit à la paternité, droit à l'intégrité de l'œuvre, droit de repentir ou de retrait)* et patrimoniales (*droit de reproduction, droit de représentation* et *droit de suite*) dont jouit l'auteur sur son œuvre du seul fait de sa création. Dans la pratique, désigne également la rémunération perçue par l'auteur à l'occasion de l'exploitation de son œuvre ».

Droits de propriété intellectuelle :

(V. vocabulaire Cornu) « Terme générique englobant la propriété industrielle et la propriété littéraire et artistique ».

Droit moral :

« Ensemble des prérogatives extrapatrimoniales qui confèrent à l'auteur sur son œuvre, à l'artiste interprète sur sa prestation, un pouvoir de contrôle, indépendamment de la cession des droits patrimoniaux et de l'extinction du monopole. Comporte plusieurs attributs : pour l'auteur, droit de divulgation, droit à la paternité, droit à l'intégrité, droit de repentir ou de retrait ; pour l'artiste interprète, les seuls droits à l'intégrité, à la paternité. Indisponible, perpétuel, il se transmet à cause de mort aux héritiers du titulaire initial ou aux personnes désignées par lui ».

Droits patrimoniaux :

« Droit d'exploitation qui confère à l'auteur ou ses ayants- droit le pouvoir exclusif d'autoriser ou d'interdire, durant une période limitée, tout mode d'exploitation consistant en la représentation ou la reproduction d'une œuvre de l'esprit. Jouissent également d'un monopole d'exploitation : l'artiste interprète, sur sa prestation, le producteur de phonogrammes ou de vidéogrammes sur son enregistrement, l'entreprise de communication audiovisuelle sur son programme ».

Droit de divulgation :

« Attribut du droit moral de l'auteur d'une œuvre de l'esprit en vertu duquel l'auteur (ou, à sa mort, ses représentants) peut, seul, décider de porter sa création à la connaissance du public, au moment et selon les modalités qu'il détermine librement, ou, au contraire, s'y refuser. L'exercice de ce droit est le préalable nécessaire à l'exploitation patrimoniale de l'œuvre ».

Droit de repentir et de retrait :

« Attribut du droit moral permettant à un auteur, qui regrette sa décision de divulgation d'une œuvre, de remettre en cause l'exécution à venir d'un contrat d'exploitation pourtant régulièrement passé par lui. Il permet à l'auteur : soit de retirer entièrement l'œuvre du commerce (« retrait »), c'est-à-dire faire cesser l'exploitation ; soit de remanier l'œuvre (« repentir »), c'est-à-dire de changer l'objet du contrat, et cela bien que la transformation modifie pour l'exploitant les conditions et l'intérêt du contrat ».

Droit à la paternité :

« Attribut du droit moral qui permet, d'une part, à l'auteur de proclamer le lien qui l'unit à sa création et, d'autre part, à l'artiste interprète d'affirmer le lien qui l'unit à sa prestation. Positivement, droit pour le bénéficiaire d'apposer ses nom et qualités sur l'œuvre ou la prestation, de choisir l'anonymat ou la pseudonymie. Négativement, droit de s'opposer à ce qu'un tiers appose son propre nom sur l'œuvre. Parfois étendu par la jurisprudence à l'usurpation du nom (faux artistique) ».

Droit au respect de l'œuvre :

« Droit à l'intégrité. Attribut du droit moral permettant à un auteur ou un artiste interprète d'imposer à toutes personnes un devoir de respect de son œuvre ou de sa prestation, qu'il s'agisse de tiers (vandales, iconoclastes…) ou de personnes qui ont acquis des droits sur l'œuvre (cocontractant des bénéficiaires, propriétaire du support matériel de l'œuvre). Comporte d'une part le droit au respect de la forme de l'œuvre ou de la prestation qui fait échec à toute suppression, adjonction, destruction ou modification. Inclut d'autre part le droit au respect de l'esprit de l'œuvre ou de la prestation, qui permet de s'opposer à toute altération du sens ou de la destination ».

Droit à la copie privée :

« Reproduction totale ou partielle d'une œuvre de l'esprit strictement réservée à l'usage privé du copiste et non destinée à une utilisation collective. Exception légale au droit de reproduction ».

Droit de citation :

« Exception de citation. Liberté de procéder à de courts emprunts d'une œuvre de l'esprit à des fins critique, polémique, pédagogique, scientifique ou d'information, lorsque l'œuvre est divulguée et à condition d'en respecter l'intégrité, la paternité et la source. »

Droit à l'oubli :

Principe qui limite la conservation des données à caractère personnel à une durée qui n'excède pas celle nécessaire aux finalités pour lesquelles ces données ont été collectées et traitées. Souffre des exceptions quand la conservation a pour finalité des traitements à des fins historiques, statistiques ou scientifiques dans les conditions prévues pour les archives publiques.

Droit pour toute personne physique d'exiger du responsable du traitement des données que celles-ci soient effacées quand la durée de conservation est expirée. (voir fiche *Données personnelles et anonymisation*).

Original :

« Œuvre à partir de laquelle peuvent être réalisées des copies. Dans le domaine des arts graphiques et plastiques, objet matériel dans lequel est incorporée l'œuvre de l'esprit qui, émanant de la main de l'artiste ou réalisée grâce à ses instructions et sous son contrôle donne naissance à un droit de suite. Il peut s'agir d'un objet unique ou d'exemplaires effectués en tirage limité dont le nombre est fixé en fonction de la technique de reproduction et conformément aux usages de la profession ».

Valeur probatoire :

Ce qui mesure la valeur d'un mode de preuve (écrit, témoignage) comme élément de conviction. Détermine la confiance qu'il faut accorder à ce mode de preuve dans la hiérarchie des modes de preuve.

INDEX

A

Alphabet Phonétique International (API) · 30
annotations · 31, 45, 46, 47, 86, 153, 154, **155**, 156, 172
anonymisation · 21, 42, 45, 65, **67**, 68, 69, 70, 71, 72, 73, 74, 107, 108, 109
archivage de masse · 146
archives · 36, 41, 42, 56, 66, 67, **82**, 83, 90, 91, 92, 145, 147, 148, 149, 156, 157, 190, 191
Archives de France · 191
archives de la Parole · 15, **79**, 84, 160
Archives nationales · 83, **190**
archives publiques · 87
artiste-interprète · 106
auteur · 34, 37, **39**, 40, 41, 48, 49, 51, 56, 67, 68, 107, 125, 127, 132, 193, 194, 195
autorisation · 22, 23, 24, 28, 31, 34, 41, 42, 50, 53, 54, 55, **60**, 61, 63, 64, 65, 67, 86, 91, 92, 100, 101, 105, 109, 110, 111, 113, 115, 116, 122, 132, 171, 173, 178, 188, 193

B

balisage · 46, **155**, 172
base de données · 20, 33, 39, **157**
BnF · 7, 20, 79, 83, 85, 86, 90, 91, 147, **159**, 161, 172, 191
British National Corpus · **28**, 32

C

chaine de numérisation · 149
chants · 49
CHILDES · 30
CLAPI · 27, **173**, 174
CNIL · 41, **107**, 108, 109, 110
cobayes · 26, 52
codage · 44, 45, 47, 74, 109, **153**, 154, 155
Code de la Propriété Intellectuelle · 51, 81, **91**
Compression · 148
Computer Supported Cooperative Work · 50
consentement éclairé · 34, 50, 54, **60**, 62
conservation · 19, 20, 23, 26, 34, 37, 41, 42, 43, 44, 45, 46, 47, 62, 63, 68, 108, 129, **143**, 151, 153, 154, 171, 177, 190, 191, 195
contes · **34**, 49, 125, 127
contrat de travail · 40
C-ORAL-ROM · 32
corpus · 19, **29**, 30, 32, 33, 34, 35, 75, 123
corpus alignés · 31
Corpus d'Orléans · 25, 26, **179**
corpus de référence · 19, **32**
corpus ouverts · 28
creatives commons · 37
cryptage · 22

D

DAT · 44, 136, 137, 139, **144**
Déclaration de Berlin · 36
DELIC · 48, **177**, 178
déontologie · 91, 123
département de l'Audiovisuel · 79, 84, 85, 86, **159**, 160, 161
dépôt · 22, 41, 63, 68, 80
dépôt légal ·
Dialogue Homme Machine · 50
diffusion · 20, 23, 32, **36**, 37, 38, 41, 42, 43, 45, 49, 50, 51, 56, 63, 67, 68, 108, 143, 150, 167
DOBES · **35**, 192
domaine public · **38**, 39, 41
données personnelles · 38, 41, 42, 48, 49, 51, 68, 69, **107**, 109, 113, 119, 193
données primaires · 45
données secondaires · 45
droit à l'image · 20, **110**, 132
droit à l'oubli · 42
droit à la copie privée · 41
droit à la paternité · 41
droit au respect de l'œuvre · 41

droit d'auteur · 21, 38, 39, 41, 90, 92, **190**, 193
droit de citation · 20, 23, 41, 101, 105, **111**
droit de divulgation · 41, 194
droit de la propriété intellectuelle · 37, 39
droit de repentir · 41, 194
droit de représentation · 86, 105, **194**
droit de reproduction · 86, 100, 101, 105, 111, **194**, 195
droit moral · 39, 40, 41, 49, 91, 193, **194**, 195
droits du producteur · 120
droits patrimoniaux · 39, 40, **194**
droits voisins · 86, 101, 103, **120**
Dublin-Core · **156**, 157, 172

E

EAGLES · 35
ELRA/ ELDA · 32
émissions de Radio et de Télévision · 101
empowerment · 66
enregistrements médiatiques · 24, **51**
enregistreurs · **136**, 137, 146, 154
entretien · **48**, 49, 58, 135, 137
éthique · 19, 21, 35, 43, 60, 123, **130**, 132, 133
EuroSpeech 2003 · 29
extension de finalité · 42

F

fieldwork · **52**, 55, 187
finalités · 22, 30, 41, 50, 55, **60**, 62, 64, 68, 109, 121, 132, 195
floutage · **21**, 22, 71
folklore · **123**, 125, 126, 127, 130
fonds commun · 38
fonds sonores · 87
formats · 46, 47, 63, 68, 69, 76, 141, 144, 147, 151, **153**, 154, 155, 157
Français Fondamental · 25, 26
français parlé · 25
FRANTEXT · 26

G

GAT · 30
grapho-lectes · 25
gravure · **135**, 136, 137, 141, 144, 145, 160, 191
groupe de travail · 20

I

ICAR · 7, 9, 27, 30, 115, **173**, 174
ICOR · 30, **173**
identification · 43, 67, 69, 72, 74, 84, 107, 109, 116, **127**, 128, 141, 145, 148, 149, 156, 157, 173, 193
INA · **7**, 20, 101
informateurs · 48, 52, **54**, 55, 56, 57, 58, 60, 65, 66
INSEE · 121
INTERMARC · 160
interopérabilité · 19, 22, 37, 135, 157, 173, **176**
intervieweur · **75**, 92
inventaire des corpus · 181
ISO-10646 · **30**, 47, 155

J

juxtalinéaires · 30

L

language resources · 26
langues à tradition orale · **34**, 49
langues sans traditions écrites · 25
libre accès · **36**, 39, 42
licences · **37**, 193
lieux publics · **54**, 58
LIMSI · 31
littérisme · 65
locuteurs · **26**, 27, 28, 29, 31, 33, 38, 52, 73, 74, 75, 76, 77, 135
loi Informatique et libertés · 41
Longman Grammar of Spoken and Written English · 32

M

macro-syntaxe · 33
magicien d'Oz · 53
Max-Planck Institute · 35, **192**
métadonnées · 27, 46, 54, 68, 72, 73, 81, 83, 141, **148**, 149, 150, 151, 155, 156, 157
micro · 27, 46, 50, 55, 59, 136, **137**, 139, 140, 141, 177
MiniDisc · **137**, 139, 140, 143
modes d'enregistrement · 44
MP3 · **137**, 138, 153, 154
MPEG · **148**, 150, 153, 154
Musée de la Parole et du Geste · 84

N

natifs · 76
normalisation · 47, **153**, 163
numérisation · 35, 37, 41, 47, **147**, 148, 149, 150, 154, 172, 191

O

OAI · **157**, 172
OAPI · 126
observateur participant · 53
œuvre collective · 40, **104**
œuvre de collaboration · 104
œuvre orale · 99
œuvres · 40
OLAC · **156**, 157, 172
original · 39, 45, **70**, 71, 76, 81, 154
orthographe standard · 30, **75**

P

paradoxe de l'observateur · **22**, 27, 52
parole · 19, 26, 27, 28, 29, **31**, 32, 33, 35, 43, 44, 47, 49, 73, 74, 76, 154, 155, 185
parole privée · **27**, 28
parole publique · **27**, 28
patrimoine · **81**, 90, 123, 167, 190
patrimoine immatériel · **35**, 95
PCM · **139**, 144, 153, 154
PFC · 26, **175**, 176
phonétique · **25**, 26, 28, 43, 47, 49, 75, 172
phonologie · **25**, 28
Phonothèque Nationale · 79
politiques linguistiques · 32
populations captives · 53
Praat · 31
prise de son · **135**, 154, 191
protocole Z39.50 · 157

Q

Qualidata · 93
questionnaire · **43**, 48, 92

R

radio · **24**, 27, 29, 32, 50, 87, 88, 89, 101, 105, 117, 164, 169
RAID · 146
Recherches Sur le Français Parlé · **25**, 188
récits de vie · 27, 49, 91
reconnaissance automatique · **31**, 174
rémunération · 56
responsabilité pénale · 21
responsable du traitement · **107**, 193
rétractation · **58**, 59, 60, 63
Revue Française de Linguistique Appliquée · **25**, 26, 188

S

SIDOS · 94
signal sonore · **31**, 73
sociolinguistique · **27**, 47, 50, 176
SpeechDat Exchange · 26
SpeechDat Exchange Format · 31
standardisation · 31, 35, 45, 46, 47, 73, **153**, 157
stockage de masse · **146**, 147
support numérique · 37, **143**, 160
supports optiques · 44

T

TEI · 30, 47, **155**, 156, 157
témoins · 20, **52**, 55, 80, 92
Text-to-Speech data · 26
titularité des droits · 119
traçabilité · 25, 37, 42, 64
traitement automatique de la parole · **31**, 175
Transcriber · 31
transcription · 26, 29, 30, 31, 44, 46, 47, 63, 67, 68, 70, **73**, 74, 75, 76, 77, 185, 187, 189
transcription automatique · 31

U

UNESCO · 38, 65, 66, 81, 88, **123**, 190, 192
Unicode · 30, 47, **155**, 172

V

valeur probatoire · 84
VALIBEL · 175
valorisation · 9, 19, 23, 36, 85, **160**
vie privée · 41, 49, 50, 58, 67, 91, 107

X

XML · 46, **148**, 155, 156, 157

TABLE DES MATIERES

1 Présentation....19
1.1 Les objectifs....19
1.2 Les conditions d'élaboration....19
1.3 Les aspects juridiques....20
1.4 Les autres aspects....21
1.5 La méthode....21
1.6 Le cadre juridique français....22
1.7 Un « guide des bonnes pratiques » ?....22
1.8 Quelques questions fréquentes....23

2 Le contexte....25
2.1 La linguistique et les corpus oraux....25
2.1.1 Type de données et de locuteur....26
2.1.2 Dimensions....28
2.1.3 Transcriptions....29
2.1.4 Traitement automatique de la parole....31
2.1.5 Exploitations et résultats....32
2.2 Cadres politiques de la diffusion de la recherche....36
2.3 Cadres juridiques....37
2.3.1 Le domaine public et le droit d'auteur....38
2.3.2 Le respect de la vie privée....41

3 La démarche....43
3.1 Expliciter la démarche....43
3.2 Éléments de la situation en jeu....43
3.2.1 Corpus et type de données....43
3.2.2 Techniques d'enquête....47
3.2.3 Rôle des participants....51
3.2.4 Lieux....54
3.3 Pratiques de terrain....54
3.3.1 Modes d'approche....54
3.3.2 Dispositif d'enregistrement....57
3.3.3 Demande d'autorisation et consentement éclairé....60
3.3.4 Après l'enquête : retours, debriefings....65
3.4 Anonymisation....67
3.4.1 Définition....67
3.4.2 Données concernées....68
3.4.3 Quand anonymiser ?....68
3.4.4 Comment anonymiser ?....69

3.4.5 Les limites de l'anonymisation ... 71
3.5 Transcription ... 73
3.5.1 Les descriptions ethnographiques ... 73
3.5.2 L'identification des locuteurs ... 74
3.5.3 Enjeux ... 75

4 Les corpus oraux, objets de patrimoine ? ... 79
4.1 Rappel de la situation ... 79
4.1.1 Les collections de corpus oraux ... 82
4.1.2 La Bibliothèque nationale de France ... 84
4.1.3 Les Archives de France ... 87
4.1.4 Place des corpus oraux dans les musées ... 88
4.1.5 les « Corpus oraux » à l'Ina ... 89
4.2 Les initiatives privées ... 90
4.3 L'accès aux collections ... 91
4.3.1 Quel réseau pour demain ? ... 93
4.3.2 Vers la reconnaissance d'un statut du patrimoine oral ... 95

5 Annexes ... 97

Fiches juridiques
L'Œuvre orale ... 99
Les œuvres protégées ... 103
Données personnelles et anonymisation ... 107
Le droit de citation ... 111
Le consentement ... 113
Exemples d'autorisations ... 115
Bases de données, objet d'un droit « sui generis » ... 119
Responsable du traitement ... 121
Le patrimoine immatériel et l'UNESCO ... 123

Fiches techniques
Prise de son et enregistrement sur le terrain ... 135
Supports pour enregistrer et archiver le son ... 143
Supports pour enregistrer et archiver la vidéo ... 147
Codages et formats ... 153

Institutions
Bibliothèque nationale de France 159
Les Archives : législation 163
Musées de France : législation 167
Inathèque de France 169

Travaux
Programme « ARCHIVAGE » du LACITO 171
CLAPI 173
PFC 175
DELIC 177
ESLO 179
Inventaire des corpus 181

Bibliographie 183
Bibliographie générale 185
Bibliographie Patrimoine de l'oral et conservation 190
Revues et périodiques 190
Aspects techniques 190
Langues en danger : liens utiles 192

Glossaire juridique 193

Index 197

Table des matières 201

WWW.BNF.FR WWW.CECOJI.CNRS.FR WWW.ILF.CNRS.FR WWW.TYPOLOGIE.CNRS.FR

Mise en page Pascale Racaud, Presses Universitaires d'Orléans.