Guide des bonnes pratiques

pour

la constitution, l'exploitation, la conservation et la diffusion des corpus oraux

Délégation générale à la langue française et aux langues de France

Version provisoire
Mai 2005

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
Culture
Communication

CNRS
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

www.bnf.fr www.cecoji.cnrs.fr www.ilf.cnrs.fr www.typologie.cnrs.fr

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.

Délégation générale à la langue française et aux langues de France
6, rue des Pyramides
75001 PARIS
http://www.dglflf.culture.gouv.fr

Guide des bonnes pratiques

pour

la constitution,
l'exploitation,
la conservation
et la diffusion des
corpus oraux

Délégation générale à la langue française et aux langues de France

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

Avec la collaboration de :
Michel **ALESSIO** *(DGLFLF)*, Alain **CAROU** *(BnF)*, Paul **CAPPEAU** *(Université de Poitiers)*, Ibrahim **COULIBALY** *(CDF-Université de Grenoble)*, Valérie **GAME** *(BnF)*, Fabrice **MOLLO** *(CNRS-CECOJI)*, Michel **RAYNAL** *(INA)*, Dominique **THERON** *(BnF)*, Luc **VERRIER** *(BnF)*.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ministère Culture Communication

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

1 Présentation

1.1 Les objectifs du "Guide des bonnes pratiques".

1.2 Les conditions d'élaboration de ce Guide

1.3 Les aspects juridiques

1.4 Les autres aspects de la collecte et de l'usage de données orales

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 scientifique, politique, juridique et institutionnel

2.1 Les sciences du langage et les corpus oraux

2.2 Cadres politiques de la diffusion de la recherche

2.3 Cadres juridiques de la constitution, de l'exploitation et de la diffusion des corpus oraux

3 La démarche (constitution, exploitation, conservation, diffusion)

3.1 Introduction

3.2 Eléments de la situation en jeu

3.3 Techniques d'enquête, recueil et production de données

3.4 Recueil de données et pratiques de terrain

3.5 Anonymisation

3.6 Transcription

4 Les corpus oraux, objets de patrimoine ?

4.1 Rappel de la situation des corpus oraux produits par des chercheurs au sein des institutions patrimoniales

4.2 La politique de l'Etat en matière de collecte et de conservation

4.3 Vers la reconnaissance d'un statut du patrimoine oral

5 Conclusion provisoire

6 Annexes

6.1 Fiches juridiques

6.2 Fiches techniques

6.3 Fiches formats et normes

6.4 Bibliographie

6.5 Glossaire

6.6 Index

6.7 Table des matières

1 Présentation

1.1 Les objectifs du "Guide des bonnes pratiques".

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).
Le Guide que nous vous proposons s'est fixé pour premier objectif de fournir les informations nécessaires à la constitution d'un 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, que 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'élaboration de ce Guide

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 terme de constitution des corpus, d'utilisation et de conservation ont conduit à adjoindre aux juristes des linguistes pratiquants de la 'linguistique de corpus' et travaillant sur des données orales, et des représentants des grandes institutions de conservation patrimoniale (INA, INSI, BNF).

Pour remplir cet objectif 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 de la collecte et de l'usage de données orales

Mais 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 méthode

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" proche 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 'memento 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 oeuvre. 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 connaître l'existence de ces droits et des conséquences 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 fréquentes

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.

Voici le genre de questions que nous entendons souvent:

1. *Quelles **autorisations** dois-je faire signer aux locuteurs que j'enregistre pour pouvoir ensuite exploiter ce corpus et pouvoir :*
   1.1. *le citer dans un travail universitaire;*
   1.2. *le citer dans un article publié dans une revue scientifique ;*
   1.3. *le citer dans un ouvrage à diffusion commerciale ;*
   1.4. *le mettre à disposition sur un site ;*
   1.5. *le diffuser sur CD*
   Ces différents types d'exploitation sont-ils soumis aux mêmes règles ?

2. *J'ai fait un enregistrement **de personnes que je connais bien**.*
   2.1. *A quelles conditions puis-je l'exploiter ? (exploiter est pris au sens de la question 1)*
   2.2. Si ces personnes m'ont autorisé(e) à exploiter cet enregistrement, *peuvent-elles revenir sur leur autorisation ?*

3. *Lorsque j'enregistre **des enfants**,*
   3.1. *qui peut donner son consentement ?*
   3.2. *lorsque l'enfant sera majeur peut-il revenir sur ce consentement ?*
   3.3. *si l'enregistrement a lieu dans le cadre scolaire, faut-il des autorisations particulières ?*

4. *Dans le cadre d'un travail au sein d'un **laboratoire**,*
   4.1. Qui est considéré comme l'auteur du corpus ?
   4.2. Quel(s) droit(s) ce travail donne-t-il au chercheur ?

5. *Pour moduler **l'accès aux corpus**, quelles formulations puis-je utiliser ?*

6. *Si je masque **les noms propres** de personnes, cela suffit-il pour que je puisse utiliser librement une transcription?*

7. *Sous quelles conditions puis-je **archiver mon corpus** sous la forme de fichiers informatiques?*

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 ?*

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.*
   9.1. *Puis-je me servir de ces documents ?*
   9.2. *Quelles précautions (quelles garanties) dois-je prendre ?*

10. *J'enregistre **une émission à la radio**.*
   10.1. *Puis-je utiliser librement la transcription ?*
   10.2. *Puis-je utiliser la version sonore ?*
   Du point de vue des autorisations, y a-t-il une différence entre émissions des radios publiques et des radios privées ?

11. *Dans les **émissions de radio ou télé**, 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.) ?*

12. Pour une **émission de télé**, *les droits d'exploitation sont-ils différents*
   12.1. *si j'achète une cassette, un DVD, un CD d'émission* ?
   12.2. *ou si j'enregistre moi-même l'émission lorsqu'elle est diffusée ?*

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 Les sciences du langage et les corpus oraux

Issu de la prise de conscience de linguistes soucieux d'assurer la pérennité 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 sont communes avec les documents produits dans le cadre d'autres disciplines. L'ethnologie, l'anthropologie, la sociologie, la psychologie, la démographie, l'histoire orale utilisent l'enquête orale, le témoignage, l'interview, le récit de vie. Cette première édition exclut volontairement les enregistrements musicaux.

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 les 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 "grapho-lectes", 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 fait beaucoup évoluer les théories parmi les spécialistes.
De nouveaux domaines, abordés dès les années 1970 en Grande-Bretagne (Sinclair & Coulthard, 1975 pour l'Ecole 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 Etats-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 aux 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 perdus, 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 1980-1990, 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 le *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.

Type de données 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 C.Lynch 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 5000 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'enregistrements 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 300h 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 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 évidemment des types de situations et de locuteurs choisis ; les principaux obstacles tiennent aux lieux bruyants, aux locuteurs trop nombreux, aux locuteurs affectés d'un défaut de parole, à un équipement technique insatisfaisant. 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).

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 durées assez limitées. 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 de la vitesse 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 comptent dix millions de mots transcrits. Malheureusement, les corpus actuels de français parlé ne dépassent pas 2 à 3 millions 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.

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 débats parlementaires, 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 connues sous le terme de TEI (*Text Encoding Initiative)* et, comme le signalait déjà E. Ochs en 1976, la transcription engage toujours une théorie. Certaines études imposent de choisir des transcriptions

phonétiques ou phonologiques (on peut se procurer des conventions comportant 95.000 caractères, dont ceux de *l'Alphabet Phonétique International* dans le système UNICODE: http://www.unicode.org). 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 2005).
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, PRAAT, Transcriber) facilitent les manipulations et permettent de réécouter facilement les portions d'enregistrement sous étude (les transcriptions automatiques restant pour l'instant de l'ordre de l'utopie, comme le montrent les travaux du LIMSI). La technique des *corpus alignés*

permet de lire sur écran des portions de texte écrit en même temps qu'on écoute les mêmes portions dans leur déroulement sonore (cf. *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, méta-données (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).

Exploitations et résultats

Les grands corpus actuels de langue parlée sont chers. Ils exigent plusieurs types d'engagements financiers : laboratoires universitaires, industriels, éditeurs, banques, puissance publique. Toutes ces dépenses ne seraient pas envisageables si, d'une façon ou d'une autre, ces corpus n'étaient "rentables". C'est évidemment dans l'ingénierie et dans les relations avec les industriels qu'ils le sont le plus : dialogue homme-machine, reconnaissance et synthèse de la parole, communications téléphoniques, Text-To-Speech systems, etc (des organismes comme ELRA/ ELDA se sont spécialisés dans la diffusion des corpus et des ressources disponibles dans ce domaine). Mais les disciplines universitaires y trouvent aussi leur compte.

Ces 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 de 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 bases 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 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), 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 (CORAL-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 1996). 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 et Moneglia 2005, Couper-Kuhlen & Selting, 1996). Dans les interactions, on a montré qu'intervenaient plusieurs niveaux d'organisations imbriqués les uns dans les autres (Turn-Constructional Units ou "Unités de Construction du Tour", cf. Selting, 1995, 1998, 2000, Auer et alii, 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 adoptée (Biber 1987). On peut montrer par exemple, pour le français, que les récits d'explications 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 de la langue 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 pré-construit (J. Sinclair, 1991).
Ceci 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 de langue maternelle et de langue 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 à tradition orale

Les problèmes rencontrés lors du recueil de données orales issues de communautés non-occidentales recoupent partiellement ceux rencontrés lors de l'établissement des grands corpus de langue orale. Les techniques de collectes sont variées (enregistrements sollicités –questionnaires lexicaux ou grammaticaux-, ou spontanés –enregistrements de textes lors d'activités de la communauté, en présence ou non du chercheur), les conditions d'enregistrement plus ou moins idéales, et le choix des locuteurs dépend du type d'enquête et de résultats attendus (locuteurs monolingues pour le recueil de textes spontanés, locuteurs bilingues pour les enquêtes dirigées, locuteurs choisis selon des critères géographiques ou sociaux pour les enquêtes sur les variétés, etc.). En raison des contraintes de transcription / traduction, les corpus sont généralement de taille réduite par rapport aux grands corpus de langues occidentales, ils restent ouverts et sont amenés à être enrichis au fur et à mesure des travaux de terrain du chercheur. On trouve parfois, face aux transcriptions de leur langue, des réactions de rejet de la part des locuteurs, et des jugements normatifs face à leurs propres enregistrements ou face à ceux de leurs compatriotes. Et dans la mesure où le travail s'effectue dans une relation de confiance entre le chercheur et ses informateurs, les mêmes contraintes de respect de la personne et de son intégrité doivent être appliquées dans la diffusion et l'exploitation des données recueillies (anonymisation, etc.).

La différence principale des enquêtes sur ce type de terrain tient à l'altérité du chercheur et à la difficulté pour lui de trouver sa place dans la communauté dans laquelle il souhaite travailler. De plus en plus de communautés revendiquent en effet un droit sur le travail du chercheur, dans ses méthodes et dans sa finalité, et exigent que celui-ci s'engage à restituer le produit de sa recherche (sous forme de corpus, et d'applications/implications dans des domaines divers –voir en ce sens les recommandations du rapport de l'UNESCO sur les langues en danger) ; par ailleurs, de nombreuses communautés ont vu se développer en leur sein une classe d'intellectuels, pas toujours formés aux méthodes des chercheurs, mais qui se revendiquent, en tant que locuteurs natifs, comme étant les seuls à même de mener légitimement des recherches sur leur langue, leur histoire, ou leur société.

C'est la raison pour laquelle, parallèlement à toutes les questions qui se posent pour l'établissement des corpus oraux et qui sont abordées dans ce guide, le travail dans certains contextes ne peut se faire sans une négociation préalable avec la communauté dans laquelle le chercheur souhaite travailler (avec ses autorités religieuses ou politiques, avec ses intellectuels, dans le respect du droit de cette communauté), qui permette à celui-ci de travailler dans les meilleures conditions possibles, et qui garantisse l'accès à la communauté pour les générations de chercheurs à venir.

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 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.

2.2 Cadres politiques de la diffusion de la recherche

- ***La diffusion des résultats 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"[1].

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'échange, et les opportunités offertes aux titulaires de droits pour faciliter la liberté d'accès dans la société 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 Etablissements 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 des politiques en charge 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, dans 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 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 droit. Les auteurs (ou l'institution) peuvent concéder un "droit gratuit, irrévocable et mondial

[1] Art 5 de la Loi no 82-610 du 15 juillet 1982 modifiée d'orientation et de programmation pour la recherche et le développement technologique de la France. JO du 16-07-1982, p. 2273 et ss.

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 numérisation 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 Etats 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 ne retire 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 de la constitution, de l'exploitation et de la diffusion des corpus oraux

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 des responsabilités 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 annexes……).
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 ?

- ***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 ambigüités 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 70 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 situation 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.

- ***Quand est-il question de droit d'auteur à propos des 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'oeuvre 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"[2]. 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.

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

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'Etat sur les créations de fonctionnaire. L'opportunité de 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 "Etat" quand la création de l'oeuvre 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'oeuvre pour les besoins de sa mission seraient cédés à son employeur Etat (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 à 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 droits (bénéficiaire d'une cession, héritiers...) d'autoriser ou interdire la reproduction ou la communication au public de l'oeuvre 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 (voir 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 annexe – citation).

- ***Le respect de la vie privée 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 annexe Consentement), respect des finalités annoncées…..[3] 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 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 annexe 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

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

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[4] - suivant le degré de sensibilité des données contenues dans le corpus).

- ***Quelles sont les responsabilités des personnes en charge de la diffusion des corpus sur Internet*** *?*

La diffusion des corpus oraux sur Internet peut être assimilé 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 annexe).

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

3 La démarche (constitution, exploitation, conservation, diffusion)

3.1 Introduction

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 de 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 (3.2), des pratiques de terrain (3.3/3.4), mais aussi des différentes étapes de leurs traitements (3.5/3.6), 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 Eléments 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 la rue 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 constituent le corpus et leurs supports (d'enregistrement, mais aussi de stockage pour exploitation, et de conservation) (3.2.1),
- les *différentes techniques* employées par les chercheurs pour récolter les données (3.3.),
- la définition des *participants* et de leur rôle (3.3.2),
- la *catégorisation des lieux* de la collecte (3.3.3).

Corpus et type de données

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. En opposition à ce modèle fonctionnel, les enregistrements modernes de la parole sont, eux, sous-tendus par un modèle physique. La parole y est vue comme une activité qui provoque des variations de la pression de l'air, variations qui vont être perçues de manière auditive par une oreille humaine. Enregistrer la parole revient donc à mesurer ces variations puis à en garder trace d'une manière ou d'une autre.
L'histoire moderne de l'enregistrement audio peut se voir à travers l'examen des modes d'enregistrement comme à travers celui des supports d'inscription utilisés.

3.2.1.1 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.

3.2.1.2 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 ont été longtemps utilisés durant le siècle dernier, et sont maintenant 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 20ème 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 aimentation des particules pourra suivant les modes d'enregistrement coder des informations sous forme binaire (comme dans les supports disques-dur, cassette DAT, disquettes informatiques, etc.) ou bien des informations sous forme analogiques (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 leurs formes 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. [Pour plus de détail cf. annexe : support d'enregistrement, supports d'archivage].

3.2.1.3 Les critères 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 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.). Enfin, pour la conservation, un support qui ne peut pas être effacé est aussi, peut être, une bonne garantie pour éviter des accidents malencontreux.
Le choix d'un format qui permet la reproduction à l'identique garantit une forme de pérennité aux données. En revanche la notion d'original se déplace alors du support aux données elles-mêmes.

Standardisation des annotations

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

3.2.1.4 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 p.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. Le terme de "donnée" de ce point de vue doit aussi être réinterrogé : les "données" ne sont jamais "offertes" ni" (re)cueillies" mais elles sont activement produites par les chercheurs. Du côté des données secondaires, l'importance de la dimension interprétative semble plus évidente : celle de la transcription a été soulignée dans la littérature, ainsi que celle du choix des phénomènes (segmentaux, suprasegmentaux, kinésiques, etc.) devant faire l'objet de transcription ou de description, avec des niveaux de granularité différents (on distinguera ainsi la description "(rire)" de la transcription "hh HHH h"), ainsi que d'annotations par des catégories théoriquement informées.

3.2.1.5 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 (p. ex. "verbe"). exprimés dans des types de structures particulières (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 texte. 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 (cf. fiche technique sur les formats).

3.2.1.6 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 texte, les noms des éléments de structure (balises, attributs...) doivent être connus et leurs définitions acceptées et

partagées, ainsi que l'ensemble des contraintes (enchainement des 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 d'une standardisation. C'est ce 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 de ces 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 (Pour en savoir plus, cf. fiche technique sur les formats).

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

- Cette technologie permet-elle de *coder de manière explicite* tous les commentaires?
- Cette technologie présente-t-elle un *caractère propriétaire ou une limite légale* qui empêcherait de partager les commentaires avec d'autres (formats propriétaires, techniques basée 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*?

3.3 Techniques d'enquête, recueil et production de données

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, 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 de conversation, psycholinguistique, linguistique de l'oral, traitement automatique de la parole, ethnolinguistique …). Cependant les recherches sur la méthodologie de l'enquête (Labov) ont conduit les chercheurs à considérer les données enregistrées comme étant le produit de la situation d'enquête en opposition à une conception de données préexistantes simplement (re)cueillies.

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.

3.3.1.1 Le questionnaire

Le questionnaire oral enregistré peut revêtir différentes formes, néanmoins 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 protégé). 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, il convient 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. 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.

3.3.1.2 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 (entretien directif, semi-directif), le rapprochant ainsi du questionnaire oral ou de l'interaction moins contrainte. 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 (protocole DELIC[5] par exemple), et pose donc le problème du choix du moment et du contenu des informations fournies aux interviewés.
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.

3.3.1.3 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

[5] http://www.up.univ-mrs.fr/delic/corpus/index.html

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 anodine 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.

3.3.1.4 Les récits 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. Ces types d'enregistrement représentent nécessairement une source importante de données personnelles concernant l'auteur du récit et des 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 leurs diffusions 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.

3.3.1.5 L'enregistrement en laboratoire selon un protocole expérimental.

Les enregistrements en laboratoire selon un protocole expérimental sont utilisés notamment dans les domaines de la psycholinguistique 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 cités en 3.3.1.1). La situation particulière de la personne enregistrée reste toutefois à rapprocher de tous les cas de recherches expérimentales sur personne humaine.

3.3.1.6 L'enregistrement d'activités provoquées dans des contextes sociaux, avec éventuellement des tâches/consignes proposées par le chercheur.

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…).

3.3.1.7 L'enregistrement d'activités dans leur contexte ordinaire et non provoquées par le chercheur.

Les recherches en sociolinguistique, analyse de conversation et analyse des usages des technologies (Computer Supported Cooperative Work ; Dialogue Homme Machine), s'intéressent au recueil de données en situation d'activités non orchestrées par le chercheur et non provoquées 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. Il y a donc de fortes probabilités pour que les données contiennent des informations sensibles au regard de la protection de la vie privée. De plus les conditions du recueil du consentement peuvent être particulièrement délicates puisqu'elles doivent être le plus souvent temporellement distinctes de l'activité de collecte.

3.3.1.8 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. 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.

3.3.1.9 La reprise d'enregistrements médiatiques

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 public.
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 absolument nécessaire 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*[6].

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, définissent autant de *situations* qui mettent en évidence des *participants* dont le rôle est le premier élément de catégorisation.

Rôles 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 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.

3.3.1.10 catégories 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 (cf. Cameron et alii, 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
- etc.

[6] art 122.5 du code propriété intellectuel

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 fondent 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.

a) 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 une communauté dont il fait partie.
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, 1976) - aussi bien qu'en termes de violence symbolique entre l'enquêté et l'enquêteur (Bourdieu, 19_&). 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, cf. Clifford & Marcus, 1986 ; Mondada, 1998).
Les enquêtes chez les "proches" du chercheur, lorsque celui-ci exploite ses propres réseaux pour son 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.
La recherche en sciences sociales et humaines a en outre souvent utilisé des "populations captives", dans le sens où l'enquêteur y a 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
- ....

L'usage est alors de doubler les autorisations pour les personnes par l'autorisation d'une responsabilité légale (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).

b) 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 visant à y intervenir par une démarche dite de "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 à 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.

Selon ces rôles, 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.

Lieux

L'information sur le lieu de la collecte conditionne des éléments de réponses juridiques particuliers de par ses propres caractéristiques et de par 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 perceptibles au sein des données) ou ne relever que d'une information éventuellement présente parmi les métadonnées.

3.4 Recueil de données et 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 établir une relation avec les personnes concernées (3.4.1) : ces modes d'approche sont étroitement liés non seulement aux méthodologies d'enquête (cf. supra 3.3/3.4.) mais aussi aux possibilités et limitations techniques du dispositif d'enregistrement choisi (3.4.2), dont dépendent les contraintes spécifiques pour les autorisations à effectuer un enregistrement (3.4.3.). Une fois terminée l'enquête et analysées les données (voir infra 3.6), 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 (3.4.4.).

Modes d'approche des personnes concernées par l'enquête

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 3.3.2.1). La mobilisation de ces personnes varie selon la méthode d'enquête choisie : nous insisterons ci-dessous sur la temporalité des différentes approches (3.4.1.1) auprès des personnes directement concernées ou de leur hiérarchie (3.4.1.2), et sur la question de savoir comment organiser le retour, le contre-don, éventuellement la rémunération de ces personnes (3.4.1.3).

3.4.1.1 Temporalité des modes d'approche et typologie des relations avec les informateurs

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 (Duranti, A. (1997), Hammersley, M., Atkinson, P. (1995), Moerman, M. (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' (voir section 3.4.4.) 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 recontactée 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.

3.4.1.2 Les personnes contactées

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.

3.4.1.3 Rémunération

Lors de l'approche des personnes concernées par l'enquête, des promesses peuvent être faites, de véritables contrats peuvent être proposés, des contre-parties, 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, 3.4.4.) 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 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 davantage 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.

Choix du dispositif d'enregistrement : modalités et contextes

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 (3.4.2.1) aux modalités de l'enregistrement (3.4.2.2).

3.4.1.4 Contexte de l'enregistrement

Par définition, il n'est pas possible de *tout* enregistrer et les chercheurs sont obligatoirement 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 (cf. 3.4.3).

- o 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).

3.4.1.5 Modalités 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 sur 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 -, quitte 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 (p. ex. 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, R.G.Jr. (1991), Mondada, L. (à paraître a), Welland, T., Pugsley, L. (eds.) (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 conséquences sur la naturalité des conduites enregistrées aussi bien que sur la gêne ou le dérangement occasionnés dans l'activité elle-même (selon les activités – comme opérer un patient, effectuer une consultation en thérapie, discuter d'un contrat délicat, être engagé dans un processus de création – ces interruptions peuvent être considérées comme à éviter absolument) ;

- le fait que l'enregistrement offre ou non des *angles morts* aux participants qui voudraient lui échapper un instant : par exemple, alors que le cadre et le champ délimités par une seule caméra permettent d'inférer des zones qu'elle ne couvre pas – de même que la puissance imaginée d'un micro -, le fait de recourir à plusieurs caméras sur la même scène peut 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 doive être effectué plus tard, ou par des tiers, peut donner l'impression d'une plus ou moins grande latitude à 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 (cf. 3.2.1.2).

Ces considérations 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.

Information aux enquêtés et demande d'autorisation (consentement éclairé)

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) est sensible 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" (3.4.3.1), en reprenant la question du moment auquel ces questions se posent (3.4.3.2), ainsi que la question des personnes que l'on informe et à qui on demande l'autorisation (3.4.3.3), des formes que prend cette information (3.4.3.4) et des objets à propos desquels on choisit d'informer (3.4.3.5), ainsi que sur les formes du consentement lui même (3.4.3.6).

3.4.1.6 Définition du "consentement éclairé"

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 doit privé n'a aucun sens?).

3.4.1.7 Moment de l'information et de la demande

La demande d'autorisation dépend du mode d'approche des personnes enregistrées (cf. supra 3.4.1). 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 (p.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 (p. 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) (cf. 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 ultérieurs.

3.4.1.8 Qui énonce l'information, la demande et qui y répond

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 (p.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.

3.4.1.9 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"? et informer "à propos de quoi"? à 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 éventuelles 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.

3.4.1.10 L'objet de la demande d'autorisation

Une fois l'enquêté informé, ayant donc satisfait à l'obligation de "l'éclairer", il s'agit de lui demander l'autorisation de collecter les données. 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 unes aux autres :

- les *actions* effectuées par les chercheurs dans le cadre du projet :
  - l'enregistrement,
  - la préparation du corpus (transcription, traduction, annotation, etc.),
  - 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).

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. 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).

3.4.1.11 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 l'exigence de demander la signature de l'enquêté n'a de sens que dans les cultures de l'écrit, de la *literacy* 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
- ou demande *orale*
- il est possible et utile de prévoir que l'autorisation orale soit elle-même *enregistrée*, sous forme audio ou vidéo. Cela 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 (p.ex. il peut accepter l'enregistrement audio mais refuser l'enregistrement vidéo) voire d'ajouts de contraintes (p.ex. il peut demander l'anonymisation de la vidéo et non pas seulement 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 des individus.

Prévoir l'après de l'enquête : retours, debriefings

On insiste souvent sur la préparation du terrain, mais il est également important de préparer 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 "pris" 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 (cf. aussi les questions de rémunération traitées supra, 3.4.1.3). 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 qu'elle soit (implication dans des projets éducationnels, de littéracie, etc.)[7].

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 comptes-rendus de résultats partiels par exemple. La distinction entre le "pendant" et l'"après" du terrain peut ainsi être relativisée.
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 *transcriptions* et d'analyse de transcriptions : les participants réagissent de manières très différentes (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 :
    - 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 (p. 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

7 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."

(ainsi pour les langues en danger on pourra[8] constituer un patrimoine légué 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 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.

3.5 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 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).

Définition

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 traitements des informations (tri, recoupement, requêtes croisées,…).
Les aspects touchés par ces considérations sont :

[8] C'est même un devoir selon les recommandations de l'UNESCO en la matière (Cf. annexe Unesco)

- 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 procédés d'"anonymisation" des données.

Données concernées par l'anonymisation

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 premières vidéo,
- les données premières audio,
- les données premières 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[9].

Moments auxquelles peut intervenir l'anonymisation

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 *tardivement* 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 premières originales de référence car elle pourrait endommager les données elles-mêmes ; par

[9] art 122.5 du code propriété intellectuel

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 des données en dehors de lui) sur des données non anonymisées et garantir en revanche une anonymisation de tout extrait figurant dans un écrit ou une présentation orale.
- 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).

Modes d'anonymisation

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.

3.5.1.1 Formes ou éléments des données pouvant être concernés 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, numéros 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 rendent capable le chercheur d'imaginer les usages, les connaissances et les associations qui pourraient permettre l'identification d'une personne sur la base d'une forme donnée.

3.5.1.2 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.

a) **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).

a. *remplacement par un "blanc"* : c'est la solution la moins informative et surtout la moins visible.
b. *remplacement par un hyperonyme*, 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).
c. *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 :
    i. 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". Cet exemple montre que le pseudonyme tentera 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, p. ex. s'il est à la base de plaisanteries ou de jeux de mots ; le nombre de syllabes et certaines caractéristiques phonétiques et prosodiques, si elles sont exploitées dans l'interaction.

ii. 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).
iii. 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 – p. ex. "Monsieur Gros").
iv. 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 [toujours] 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 b.

b) Audio

a. *remplacement par du silence*. Cette solution a comme désavantage le fait que le remplacement peut être confondu avec une pause.
b. *remplacement par un bip ou un autre bruit* qui ne se confond avec aucun signal pouvant intervenir dans l'enregistrement.
c. *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).

c) 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.

a. 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.
b. *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.
c. placement d'un bandeau noir sur les yeux de la personne

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 et affirmées face au public 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 (3.5.5.1.), le second concernant les contraintes que l'anonymisation fait peser sur les objets mêmes de la recherche (3.5.5.2.).

3.5.1.3 Limitations issues des contextes de production et de circulation des données

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 (3.4.4.1) ; 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.
- 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 leur collègue 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 de linguistique d'une université. 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).

3.5.1.4 Limitations issues du travail 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 voir 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 peut donc affecter de manière cruciale cette information.
Cela peut être le cas de l'anonymisation des visages sur une bande vidéo qui rende 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.

3.6 Transcription

La transcription est une pratique qui, loin de se limiter à un exercice technique de reproduction, intègre de nombreux enjeux théoriques et interprétatifs (cf. 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 aux identités 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 et la stigmatisation des locuteurs et de leurs façons de parler. Nous nous limiterons ici à considérer 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.

Introduction de la transcription : description ethnographique

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 et 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 rendant l'appartenance à telle catégorie ou a 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.

L'identification des locuteurs dans la transcription

La transcription doit intégrer les résultats de l'anonymisation (cf. supra 3.5). 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 (cf P.Roulon-Doko chez les Gbaya –Centrafrique-), alors que dans d'autres, les mentionner est une atteinte à la vie privée (cf O.Lescure chez les Kali'na –Amazonie-)).

Enjeux liés aux choix effectués dans le corps du texte

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.

3.6.1.1 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 (Cf. 2.1.3.). Les transcriptions phonétiques (API ou autres) ne sont lisibles que par les spécialistes et seulement pour des dimensions limitées. Ainsi, pour lire de grands corpus, tous les 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.
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.

3.6.1.2 La représentation 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 chacune 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.

3.6.1.3 Enjeux du multimodal et du détail 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 (cf. Jefferson, 1985) peuvent nuire à la représentation de conduites non-standards (p.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 à la préservation et à l'accès aux corpus oraux ?

4.1 Rappel de la situation des corpus oraux produits par des chercheurs au sein des institutions patrimoniales

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'Etat. Les Archives de la Parole, conservatoire des langues et dialectes de France (Université de Paris) naissent 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 B.n.F. Les grandes collectes ethnographiques menées par le Musée des ATP, également Centre d'ethnologie de la France, concernent par exemple la Bretagne en 1939. 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 (Daniel Berteaux)[10] introduisent l'histoire de vie dans leur méthode. Cette piste ouvre la voie à des recherches pluridisciplinaires.

Mais la France est un pays de tradition écrite et n'attribue pas aisément de valeur culturelle spécifique à cette forme d'expression. L'université n'a donc pas développé de méthodologie critique spécifique et adaptée à sa problématique. Les historiens ont éprouvé pendant fort longtemps des réticences à considérer le témoignage oral comme une source fiable et digne de considération. Philippe Joutard, promoteur de l'histoire orale, reste très isolé en France alors que la Grande-Bretagne, l'Italie, l'Espagne, l'Argentine entres autres connaissent un développement dynamique et foisonnant au sein même de l'Université. De nombreuses revues attestent de cette vitalité (voir bibliographie).

L'excellente enquête[11] 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 sans reconnaissance officielle ni lieu pour les accueillir, ces documents sont restés dans les laboratoires et surtout n'ont pas été décrits ni documentés et les autorisations des témoins, lorsqu'elles existent, sont limitées à l'usage des chercheurs.

---

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

[11] CRIBIER, Françoise ; FELLER, Elise. *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. Doc.dactylogr.,avril 2003. 2 vol. Une autre enquête très rapide a été réalisée par Claude Dubar à la demande du CNRS (voir bibliographie)

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.
L'intérêt nouveau pour l'oral, donnée sensible et contenu souvent unique, trouve une sorte de résurrection grâce aux techniques de numérisation.

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 l'écrit. L'enregistrement analogique jusque dans les années 80/ 90, devait être écouté dans la durée. Prendre connaissance de l'intérêt d'un enregistrement demandait alors du temps et rebutait plus d'un chercheur. 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 collections. 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 ?

L'oral en forme, l'oral en mots[12]

Corpus oraux, collectes orales, archives orales, sources orales, documents oraux, patrimoine oral... de quoi parle-t on ?

En l'absence d'outils méthodologiques spécifiques, les sources orales sont qualifiées différemment selon les disciplines qui les produisent, parfois contradictoirement.

Une certaine confusion règne entre la typologie de formes d'enregistrement, *récits de vie, témoignages, entretiens*, produits de situations d'enquête ou de collecte, voire d'émissions de radio, et l'analyse de ces formes. C'est ainsi que, par exemple, les témoignages sont classés entre témoignages *provoqués* et *a posteriori*. Ce formalisme essentiellement fondé sur la notion de temporalité (témoignages " *ultérieurs* ", " *rétrospectifs* " " *récits de vie a posteriori* "), utile pour les besoins d'analyse du chercheur, n'est pas opérant pour la gestion de ces collections au sein d'une institution de conservation. Ces critères font certes partie de la description objective 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.

Mais les collections de corpus oraux constituent-elles, pour autant, une catégorie du Patrimoine ?

[12] 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, n010, septembre 1998.

La voix parlée est une succession de sons c'est-à-dire un ébranlement de l'air provoqué par une vibration mécanique. Fixé sur un support qui lui donne sa matérialité, l'immatériel devient objet à conserver. Au regard de la gestion patrimoniale, les documents oraux s'intègrent à une catégorie plus large et complexe : les collections orales, organisées par les institutions en autant de sources orales pour la recherche.
Pourtant ces dernières ne figurent pas dans le Code de la Propriété intellectuelle 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 en général. Les autres collections orales enregistrées très nombreuses mais dispersées et ressortissant à des domaines très variés, 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[13] 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 annexe). Plus modestement la Mission du Patrimoine ethnologique créée dans les années 80 au sein du Ministère de la culture et de la communication, avait dans ses débuts l'ambition de les réhabiliter !

4.2 La politique de l'Etat en matière de collecte et de conservation

Quelle place, les textes de lois qui régissent les différentes institutions patrimoniales réservent-ils aux collections de corpus oraux ?

Le Code du Patrimoine, qui a intégré en 2004 les différents textes, n'en fait pas une catégorie particulière.

4.2.1.1 Les textes de la Bibliothèque nationale de France. Pratiques et usages.

Dans son article 2, le " décret n° 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[14] ou relatif à la civilisation française ".

A ce titre,

> 1 - d'une part, elle exerce, en vertu de l'article 5, alinéa 2, de la loi du 20 juin 199215 […] 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'Etat, dans les

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

[14] Nous soulignons

[15] Remplacée depuis par les articles L131-1 à L133-1 relatifs au dépôt légal du Code du patrimoine (*Journal officiel* du 24 février 2004)

> 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 ".
> 2 - d'autre part, " elle rassemble, au nom et pour le compte de l'Etat, 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 ".

Outre la collecte du dépôt légal, la Bibliothèque nationale de France a donc vocation et mission d'enrichir ses collections par acquisitions, donations, dons, legs, dations, etc. Ce second volet d'accroissement des collections a été formalisé dans une " charte d'enrichissement des collections ", établie au niveau de l'ensemble de la Bibliothèque, et qui en détermine les grands axes.

C'est le département de l'Audiovisuel de la Bibliothèque nationale de France qui**,** dans le cadre de cette " charte d'enrichissement des collections ", parallèlement au dépôt légal des documents sonores, vidéographiques, multimédia et informatiques dont il a la charge, a vocation et mission à enrichir ses collections par tout autre mode d'entrée.

Les grandes lignes de la politique documentaire du département 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'immensité du champ possible aux contenus multiples (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 documentaire en la matière.

Le critère 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) illustrent ce propos.

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 à priori aucun lieu de conservation et / ou de consultation déterminée. C'est le cas, par exemple, de certaines archives personnelles ou de fonds en déshérence dans certains laboratoires, faute de structures appropriées.

L'accessibilité 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 sera fourni en même temps que l'archive sous forme de métadonnées ; soit, éventuellement, que toutes les informations seront fournies à la BnF sous une forme ou une autre pour lui en permettre le traitement documentaire.

L'accessibilité 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 :

- d'être le propriétaire des supports physiques sur lesquels ont été réalisés les enregistrements, et que ces enregistrements soient susceptibles d'être donnés à la Bibliothèque ;
- d'être titulaire ou de pouvoir garantir
    - o les droits d'auteur sur les œuvres réalisées ;
    - o 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 nécessaires 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'être consulté 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 inédits reçus en don par le département 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)

La Bibliothèque nationale de France et la coopération 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 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 leurs 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.

4.2.1.2 Les textes des Archives. Pratiques et usages.

Les Archives de France
Les Archives de France constituent une direction du Ministère de la Culture et de la communication (arrêté du 23 octobre 1979 modifié par l'arrêté du 19 février 1988) qui assure la mise en œuvre et le contrôle de la loi 79-18 du 3 janvier 1979 sur les archives. Elles coordonnent toutes les attributions confiées par la loi du 3 janvier 1979 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 les Archives de France ne gèrent plus directement les archives départementales, placées désormais sous l'autorité du Conseil général.
La direction des Archives de France est constituée :

- d'une Inspection générale,
- d'une sous-direction des services administratifs,
- d'un service technique,
- d'une délégation aux célébrations nationales.

Qu'est-ce qu'une archive au sens de la loi 79-18 du 3 janvier 1979 ?
Dans le Titre 1er, les archives sont définies à l'article premier 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 :

Dans le Titre II, article 3, **les archives publiques** sont :

" 1er Les documents qui procèdent de l'activité de l'Etat, des collectivités locales, des établissements et entreprises publics ;
2e 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 ;
3è Les minutes et répertoires des officiers publics ou ministériels.
Les archives publiques, quel qu'en soit le possesseur, sont imprescriptibles.
Les conditions de leur conservation sont déterminées par le décret en Conseil d'Etat prévu à l'article 32 de la présente loi. "
Les archives publiques font l'objet de procédures de tri et, selon certaines règles précises, d'élimination.
" Article 4. - A l'expiration de leur période d'utilisation courante par les services, établissements et organismes qui les ont produits ou reçus, les documents visés à l'article 3 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. "

Dans le Titre III, la loi définit **les archives privées :**

" Art. – 9 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 3 "

C'est le mode de production, et non pas le type de support ni le sujet, qui définit l'appartenance à l'une ou l'autre catégorie. Par exemple : l'enregistrement d'une séance du Conseil général est une archive publique, tandis que l'enregistrement de l'interview d'un personnage politique à la radio est une archive privée.

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 7 ci-dessous.
Article 7. – Le délai au-delà duquel les documents d'archives publiques peuvent être librement consultés est porté à :

1° Cent-cinquante ans à compter de la date de naissance pour les documents comportant des renseignements individuels de caractère médical
2° Cent-vingt ans à compter de la date de naissance pour les dossiers de personnel ;
3° 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 ;
> 4° Cent ans à compter de la date du recensement ou de l'enquête, pour les documents contenant des renseignements 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 ;
> 5° 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'Etat ou la défense nationale, et dont la liste est fixée par décret en Conseil d'Etat. "

Pour **les archives privées :**

> Titre III, Article 10 . – Lorsque l'Etat et les collectivités locales reçoivent des archives privées à titre de don, de legs, de cession, de dépôt révocable ou de dation au sens de la loi n° 68-1251 du 31 décembre 1968 tendant à favoriser la conservation du patrimoine artistique national, 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 versement.

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 (Titre III, article11). L'article 14 spécifie que " les archives historiques sont imprescriptibles " mais que, article 12, " le classement de documents comme archives historiques n'emporte pas transfert à l'Etat de la propriété des documents classés ".
Enfin, " toute destruction d'archives classées est interdite ", Article 15.
Ni les documents conservés par la BnF, ni les archives publiques et privées n'ont de valeur probatoire. Par contre, la date d'entrée du document dans les collections de ces institutions peut constituer une preuve d'antériorité parmi d'autres en cas de nécessité.

Place des archives orales au sein des Archives nationales et des Services d'archives départementaux 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 Archives nationales

Placées sous la responsabilité de la direction des Archives de France, elles regroupent cinq centres, dont trois seulement reçoivent des documents sonores et audiovisuels :

le **Centre Historique des Archives Nationales** (CHAN) à Fontainebleau. C'est là, par exemple, que sont versées les 400 heures d'enregistrements 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*...) De même les enregistrements vidéo des archives

judiciaires (procès de Klaus Barbie, de Paul Touvier ou du sang contaminé) et les Archives de la Présidence (discours et conférences de presse des présidents de la République).
le **Centre des Archives contemporaines**. C'est au sein de la section XXe siècle 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 également 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 ".[16]
le **Centre des Archives du monde du Travail** à Roubaix qui collecte tout type d'archives sur son domaine dont des enregistrements collectés…
Les deux autres centres sont celui d'Esperran qui ne conserve que des microfilms et celui des Archives de la France d'Outre-mer qui conserve un fonds imprimé clos.

Les Services **d'archives départementales**. Décentralisés bien avant les autres, ces services collectent souvent, à l'initiative de leur directeur, des copies d'émissions de radio, des films d'amateurs, des documentaires, et conduisent des programmes d'enquêtes orales, seuls ou avec des concours associatifs et universitaires. Tout comme certains services **d'archives municipales** qui, 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 " (par exemple à Martigues et à Lille).

4.2.1.3 Les textes des Musées de France. Pratiques et usages

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'Etat, à une autre personne morale de droit public ou à une personne morale de droit privé à but non lucratif ".
> Il 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.

---

[16] Agnès Callu, *Aux Archives Nationales, une politique raisonnée en faveur des témoignages oraux* in : Colonnes : archives d'architecture du XXe s., n° 20, décembre 2002.
[Actes, *Les archives orales dans l'histoire de l'architecture* décembre 2000], pp.31-33

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 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'une forme de musée de France dans les deux cas.

Place des corpus oraux dans les musées
Le musée, dans son acception la plus large, collecte et conserve des objets. Sa mission première est de les étudier et de les présenter au public dans le cadre d'une muséographie attractive.
Les enregistrements oraux par définition constituent, pour le musée, de l'immatériel. 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 très récemment sur la dimension immatérielle du patrimoine intangible. Le malaise ressenti par les musées occidentaux classiques, 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 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.

4.2.1.4 Les textes de l'INA / Inathèque site. Pratiques et usages.

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.

- Le secteur des archives professionnelles assure l'archivage et l'exploitation commerciale des fonds issus des chaînes publiques de radio et de télévision.

- L'Inathèque de France assure depuis 1995, selon la loi du 20 juin 1992 sur le dépôt légal, « la conservation patrimoniale et la communication des œuvres et des documents, diffusés, de la radio et de la télévision françaises, à des fins de recherche »

l'Inathèque de France[17]

[17] Renseignements pratiques :

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

Rapide historique du Dépôt légal de la radio télévision :

- **1992.** La loi du 20 juin étend le Dépôt Légal à la radio-télévision, dans une perspective patrimoniale et de recherche ;
- **1993.** Les décrets d'application fixent au 1er janvier 1995 le démarrage de la conservation ;
- **1995.** Naissance de l'Inathèque de France. L'INA reçoit en dépôt les émissions des diffuseurs hertziens et capte les chaînes de Radio France. Préfiguration d'une activité de consultation ;
- **1998.** Ouverture, en octobre, du Centre de consultation sur le site de la Bibliothèque François-Mitterrand ;
- **2002.** Le Dépôt Légal est étendu aux chaînes du Câble et du Satellite. Lancement des premiers travaux sur le Dépôt Légal du Web ;
- **2003.** Nouvelle phase d'extension aux chaînes du Câble et du Satellite ;
- **2005.** Collecte et conservation des chaînes gratuites de la Télévision Numérique Terrestre (TNT)

Les secteurs d'activités

De la captation des programmes à la consultation, en passant par l'enrichissement documentaire, une chaîne ininterrompue de compétences et d'outils.

Collecte et conservation

Ce sont 45 chaînes de télévision et 17 chaînes de radio qui sont suivies 365 jours par an, et ce seront à terme plus de 70 chaînes collectées. Pour suivre chaque année plus de 380 000 heures de programmes de télévision et 150 000 heures de radio, correspondants de chaînes et magasiniers alimentent au quotidien les bases informatiques en données d'identification et de catalogage.

Captation Télévision

Service de la consultation, Quai François Mauriac, 75706 Paris Cedex 13, Téléphone : +33 (0)1 53 79 48 30, Télécopie : +33 (0)1 53 79 48 40, consultation.inatheque@ina.fr. Pour consulter la catalogue national du dépôt légal de la radio et de la télévision françaises. Pour s'informer sur les manifestations et les publications de l'inathèque de France : www.ina.fr/inatheque/.

En moins de deux ans, le secteur de la captation télévision a réalisé sa révolution numérique. Aujourd'hui, une fibre optique ou une liaison satellite permet à la régie technique de l'Inathèque d'enregistrer 24h/24 les programmes. Les signaux sont compressés en Mpeg 2 et stockés sur un support de conservation informatique. Parallèlement, une deuxième compression est réalisée et stockée sur DVD pour les supports de consultation.

Captation Radio

Le secteur technique radio avait fait figure de précurseur en numérisant le son gravé sur CD-Rom dès 1994. Depuis cette date, les programmes de radio sont disponibles sous une forme numérique. La modularité de l'installation a permis d'assurer sans difficulté l'extension du Dépôt Légal à l'ensemble des réseaux nationaux, publics et privés, généralistes ou thématiques.

La documentation écrite

Avants-programmes, conducteurs d'antenne, scripts, scénarios, conducteurs d'émission... : soixante mille documents écrits, sous forme papier ou électronique sont également versés chaque année au titre du Dépôt Légal. A ces documents s'ajoutent la presse audiovisuelle, près de 900 titres suivis depuis 1929, les revues de presse de l'ORTF ou des sociétés qui lui ont succédé, une collection de plus de 9000 livres, ainsi que l'ensemble des travaux scientifiques sur les médias audiovisuels et des fonds d'archives écrites provenant de donations des personnalités du monde des médias.

Indexation et enrichissement documentaire

Chaque année, près de 70 000 émissions de télévision et de radio sont analysées et indexées par les documentalistes. Ils réalisent une description des contenus des programmes retranscrits sous la forme de mots-clés, de résumés, complétée par tout autre élément d'information nécessaire à l'exploitation de ces documents par les chercheurs.

Consultation

Depuis octobre 1998, 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 73 places dont 56 sont équipées d'un poste de consultation multimédia (S.L.A.V Station de Lecture AudioVisuelle) qui permet à la fois la consultation des bases de données de l'Ina, l'écoute ou le visionnage des émissions, leur analyse à l'aide d'outils dédiés et la gestion de corpus de travail.

Plus d'un million d'heures de télévision et de radio sont consultables à travers une soixantaine de SLAV (station de lecture audiovisuelle). 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.

Depuis 1995, 10 000 personnes ont été accréditées à l'Inathèque de France dont 20% d'usagers de province et 8% de chercheurs étrangers.
1800 directeurs de recherche issus de plus de 400 universités ou centre de recherche mènent ou encadrent des travaux qui considèrent les sources de radio-télévision.

Accès à l'inathèque de France

L'Inathèque de France accueille les usagers qui justifient d'un objet de recherche, qu'il soit d'ordre universitaire, professionnel ou personnel, nécessitant la consultation de documents du dépôt légal de la radio et de la télévision françaises, de certaines émissions d'archives, d'ouvrages ou d'archives écrites.

Une accréditation préalable doit être obtenue auprès de l'Inathèque de France, à «l'Orientation des lecteurs Est», en haut-de-jardin de la Bibliothèque nationale de France, site François MITTERRAND.
La réservation est obligatoire. Pour cela la carte d'accès est indispensable.

Un lieu de réflexion sur les médias
Les Prix de l'Inathèque de France : le Prix de la recherche, d'une valeur de 4000 €, récompense chaque année, depuis 1997, une thèse de doctorat ayant pour objet l'étude de la radio-télévision. D'une valeur de 2000 €, le Prix d'encouragement distingue pour sa part un mémoire de maîtrise, de DEA ou de DESS.

Les publications :
À travers une revue (MédiaMorphoses), et des collections d'ouvrages ("Médias Recherches", Mémoires de télévision"), l'Inathèque met à disposition des professionnels, enseignants, chercheurs, publics spécialisés et de tous ceux qui s'intéressent aux médias, des éléments d'information et de compréhension sur le monde de la communication audiovisuelle.

Les collections de corpus oraux. Pratiques et usages des institutions patrimoniales.

Les corpus oraux ne constituant pas une catégorie particulière et étant produits de façon unique par des producteurs individuels ou non, le législateur n'a pas prévu de dispositif particulier pour collecter et conserver un ensemble riche et foisonnant dont l'Université s'est désintéressée. 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 **l'initiative 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.

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'objet ou leur difficulté à témoigner de la dimension humaine à l'intérieur d'une collectivité. A Fécamp où l'enregistrement des ouvrières des anciennes pêcheries révèle une forme d'organisation sociale de la cité au début 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-Monceau-les-mines (Saône-et-Loire).

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 fonction du support de la collection. Rien de semblable dans les musées. A l'exception du MNATP [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.

Des corpus oraux contextualisés

Mais 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 au 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.

Des usages diversifiés qui évoluent avec le temps.
C'est une des vertus du Plan de numérisation des documents sonores mis en place fin 1999 par le Ministère de la Culture et de la communication que d'avoir 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 l'internet.
Mais 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] .

---

[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

[19] Françoise Cribier , Elise Feller op. cité p..

Elles doivent pouvoir être analysées, au cours des temps, par différents chercheurs avec leurs grilles d'analyse personnelles.

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 ? Les institutions s'engagent a priori à assurer leur pérennité physique et à organiser leur consultation dans le respect des droits de ceux qui ont participé à la création des corpus oraux, et que les cessionnaires doivent leur céder, au moins en ce qui concerne les autorisations de consultation.

Précisons que les collections entrent (le support ou les données numériques) alors de façon définitive et imprescriptible dans les fonds de l'institution, à l'exception des dépôts qui, par nature, sont toujours révocables. Cette cession, nous venons de le rappeler, n'emporte pas, sauf accord spécifique, cession des droits d'exploitation.

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 inimaginables. Pour y parvenir en ce qui concerne les documents du passé, le problème se pose en termes de conservation et d'identification. Il se chiffre en moyens financiers et en personnel.

Les initiatives privées

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[20] intégrés à une institution (Comité d'histoire du Ministère de la Culture et de la Communication, Comité d'histoire de la BnF), à l'exception de quelques-uns uns dont l'**AHICF,** Association pour l'histoire des chemins de fer en France, qui 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.

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 (voir BnF, Pôles associés). Ces centres ne disposent que très rarement des droits complets des corpus qu'ils conservent.

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

L'accès aux collections

Les conditions de consultation sont définies par contrat mais 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 malade, histoires de vie, témoignages délicats, propos 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 l'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 ? Acceptent-ils de les céder , dans quelles conditions et pour quel usage ? Pour quelle durée ? de façon immédiate ? différée ?
Le collecteur-chercheur dont l'enregistrement des corpus constitue un moment dans une recherche approfondie devrait pouvoir être protégé également en tant qu'auteur. Il est dans la plupart des cas appelé collecteur. Pour reconnaitre un droit d'auteur à l'intervieweur, il faudrait prouver qu'il y a eu forme originale.

Les institutions 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.

Il subsiste 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'Etat. Ce problème sur les droits des salariés « auteurs » butte dans la fonction publique sur des questions financières non résolues.

Questions :
Que dire sur les droits d'étudiants bien peu aguerris 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 60 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 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 CNRS / BN de 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.

Un réseau de gestion, de protection des collections de corpus oraux 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, peuvent avoir la capacité et la volonté de créer un grand réseau des sciences humaines et sociales, à travers lequel les corpus mis à disposition des autres chercheurs pourraient être protégés et rendus accessibles.

Qualidata (Essex) pourrait servir d'exemples. Le service est très sélectif et les droits de gestion ne sont pas négligeables. La conservation est prise en charge à l'extérieur.

Le SIDOS, Service suisse d'information et d'archivage des données pour les sciences sociales, créé en 1992, constitue lui aussi une sorte d'agence de gestion des données qualitatives ou quantitatives produites par les chercheurs[21].

La mise en place de tels réseaux aurait un incontestable intérêt pour la recherche. Nous ne sommes pas persuadés **que le statut patrimonial et la pérennité de la conservation 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 accessible des documents compréhensifs et cohérents. Ce travail devrait toujours être réalisé par le chercheur. Quand en prendra-t il le temps ?

[21] voir enquête réalisée par Françoise CRIBIER et Elise FELLER , op. citée.

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.

4.3 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 Conclusion provisoire

Cette version du Guide est une version provisoire, la conclusion le sera également.

Ce n'est en effet qu'au terme d'une réflexion commune entre chercheurs, juristes, conservateurs et "décideurs" qu'une politique scientifique et une politique culturelle permettront de favoriser la constitution, l'exploitation, la conservation et la diffusion des corpus oraux. D'autres chercheurs appartenant à d'autres discipline devraient être sollicités.

La journée d'étude, tout comme l'espace de discussion dédiés à cette version de travail ont justement pour objectifs d'apporter les éléments qui concluront la version 2005 du Guide.

A VOUS DE NOUS FAIRE DES SUGGESTIONS!

6 Annexes

6.1 Fiches juridiques

Données 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 soit pour mieux connaitre 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, de par 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.

Première partie : les données

Données à caractère 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'information utilisé, 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 les progrès de la génétique comme 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** à une personne pour qu'une donnée acquière un caractère personnel indirect. Les analyses ne sont pas toujours explicites mais il n'est possible de négliger cet argument[22]. 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

[22] Lamy Droit de l'informatique et des Réseaux 508 et suivants.

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. Un informaticien qui gère un système traitant des données génétiques, lorsqu'il accède à ces données, celles-ci sont-elles pour lui des données à caractère personnel ?

La notion de traitement des données à caractère 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**[23]

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 plus haut : la difficulté qu'éprouve le législateur à éviter la systématisation trop grande des notions qu'il veut défendre. Une donnée n'acquiert pas forcément le 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[24].

Deuxième partie : 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 "*dépersonalisation*" 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.

Après avoir rapidement passé en revue quelques principes clés des lois encadrant l'informatique, la difficulté soulevée par la question de l'anonymisation apparait plus clairement. 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 nous fournit quelques pistes de réflexion.

La condition première est celle 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.

---

[23] Considérant 27, de la directive 95/46

[24] FRAYSINNET, J. (sous la dir.), *Droit de l'Informatique et de l'Internet*, PUF, Paris, 2001, §127, pp85-86.

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 constituées. Des bases primaires contiennent toutes les données mais elles ne sont pas connectées au réseau, elles servent de sécurité et disposent de tables de concordances 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 une 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, ainsi que des résultats que l'on peut en tirer. 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 des 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 ne pas obtenir l'accord des autorités compétentes.

LE DROIT DE CITATION

Citer des œuvres, lors de la construction d'un travail, semble un acte banal qui a priori ne soulève aucune question. Pourtant, 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. Les principes généraux seront tout d'abord exposés en s'attachant au support de l'écrit, puis les particularités des autres médias complèteront cette analyse.

Première partie : définition du droit de citation.

Le droit de citation est une exception au droit de reproduction. En effet, on ne peut reproduire tout ou partie d'une œuvre originale par quelque moyen que ce soit, sans autorisation de l'auteur (*voir fiche œuvres protégées*). Des cas particuliers sont prévus afin de favoriser l'information ou les débats d'opinion. Ainsi l'article L122-5 3° permet les analyses et courtes citations. Pour s'en prévaloir, trois conditions doivent être observées :

1) la citation doit être **justifiée** : il faut un but (critique, scientifique..) **et** elle doit s'incorporer à un développement lié à ce but (démonstration, exposé). Sinon on tombe dans le cas du recueil qui constitue une œuvre en lui-même ; sauf pour ce qui appartient au domaine public.
2) la citation doit être **courte** : les extraits ne peuvent reprendre l'essentiel de l'œuvre dont ils sont issus. L'œuvre qui utilise ces citations doit pouvoir survivre sans elles. Cela signifie qu'il doit y avoir un véritable apport personnel de la part de l'auteur et que si l'on retirait les citations, le texte constituerait en lui-même une œuvre.
3) 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 son intégrité, tant dans la forme que dans l'esprit. Enfin, si une œuvre n'est plus divulguée, ou qu'elle ne l'a jamais été car son auteur s'y refuse, alors la citation est interdite.

Toutes les règles qui viennent d'être énoncées conviennent parfaitement au manuscrit papier. Il faut à présent s'intéresser aux autres supports qui existent aujourd'hui, et voir dans quelles conditions le code de la propriété intellectuelle prévoit des aménagements.

Deuxième partie : les cas particuliers

1) Les œuvres graphiques ou appartenant aux arts plastiques

Cette catégorie pose le problème suivant : la citation étant une reproduction partielle de l'œuvre, une partie du dessin, tableau ou photo (etc.) seulement sera donc présentée. Il y a donc atteinte à l'intégrité de l'œuvre, puisqu'elle ne peut être identifiée que lorsqu'elle est complète. La citation de ce type d'œuvre s'avère donc impossible.

2) Les œuvres musicales

A priori, rien n'interdit l'insertion d'un court extrait d'un morceau ou d'une chanson. Pourtant, cela se révèle très difficile car la jurisprudence est on ne peut plus stricte. L'affaire Jacques Brel de 1996 est exemplaire. Des étudiants d'une grande école parisienne furent condamnés pour avoir mis en ligne de courts extraits (inférieurs à 30 secondes) de chansons sans avoir obtenu au préalable l'autorisation des ayant-droits. Mieux vaut donc s'abstenir sauf si l'on dispose des accords adéquats. Il en va de même pour les enregistrements sonores (émissions de radios.)

3) 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. Il faut s'assurer que l'objet cité n'est pas illicite ou que 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.

4) 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 des prérogatives particulières. Depuis l'arrêt Microfor de 1988, 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 exigences de justification et de courte citation disparaissent.

CONSENTEMENT

Avant tout recueil de données, il est impératif d'obtenir le consentement de la personne qui va faire l'objet de l'enquête et de son exploitation. Le but poursuivi est de s'assurer que les données fournies le sont, en toute connaissance de cause. Une fois le consentement donné, le responsable du traitement doit garantir que la vie privée de l'individu ne sera pas menacée.

Il est vrai que l'écrit tient une place importante dans le formalisme du consentement. Mais 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, elles sont évoquées à la fin de cette fiche.

Première partie : les principes

Il s'agit en premier lieu de bien cerner la nature des données qui vont être recueillies. Si ce ne sont pas des données personnelles (cf fiche *Anonymisation et données personnelles*), aucune formalité ne s'impose. Dans le cas contraire, une autre question survient : ces informations entrent-t-elles dans la catégorie des données sensibles ?

I) les données personnelles

La définition du terme "donnée personnelle" ou "à caractère personnel" se trouve dans la fiche *Anonymisation*. La collecte de données sera réalisée sous la direction d'une personne responsable. Le responsable du traitement doit s'assurer que la collecte respecte certains principes (cf. *Responsable du traitement*).

Il lui faut aussi procéder ou faire procéder à une **information préalable** de la personne avant de recueillir des données. L'article 32 de la nouvelle loi de 1978 énonce clairement les informations qui doivent être fournies :

- l'identité du responsable du traitement et, le cas échéant, de celle de son représentant ;
- la finalité poursuivie par le traitement auquel les données sont destinées ;
- le caractère obligatoire ou facultatif des réponses ;
- les conséquences éventuelles, à son 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 ;
- des transferts de données à caractère personnel envisagés à destination d'un Etat non membre de la Communauté européenne.

Lorsqu'il s'agit d'une personne déclarée **incapable** (qu'elle soit majeure ou mineure), l'information doit parvenir au **représentant légal**.

La **forme** que doit prendre le consentement n'est **pas précisée** dans les textes. Même s'il ressort de ses avis que la CNIL a une nette préférence pour l'écrit, rien ne s'oppose à ce que d'autres formes soient choisies. L'important est de pouvoir faire la preuve de la volonté de la personne concernée. Toute collecte s'effectue pour un traitement précis qui doit avoir une **finalité** clairement établie. Un **détournement** de finalité constitue un acte illégal, sanctionné par une **condamnation pénale**.

II) les données sensibles

Situées dans la catégorie des données personnelles, les données sensibles constituent un ensemble encore plus protégé. Tout ce qui touche à la santé, aux mœurs, ainsi qu'aux opinions philosophiques ou religieuses, se trouve **a priori exclu** de tout traitement. (article 8 de la loi nouvelle). L'article 9 y ajoute les condamnations pénales.

Les principes qui viennent d'être présentés sont les héritiers d'une rigueur qui disparait peu à peu pour laisser la place à une réflexion plus pragmatique. On assiste à une multiplication des dérogations. La toute puissance du consentement écrit se voit contestée.

Deuxième partie : les alternatives

La loi de 1978 entendait lutter contre la menace de l'ordinateur. Cette dernière n'a pas disparu, bien au contraire, mais l'outil informatique fait aujourd'hui partie intégrante de notre quotidien ; si bien qu'il ne s'agit plus de stopper les traitements pour protéger les individus, mais plutôt d'encadrer leur réalisation afin d'éviter les effets néfastes.

I) les dérogations législatives

Le consentement de la personne doit être recherché, mais lorsque cela s'avère impossible ou trop complexe, la loi prévoit des exceptions. Ainsi, l'article 7 tout en posant le principe du consentement admet un certain nombre de cas particuliers. Le 5e alinéa admet

> "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."

Bien que la rédaction soit quelque peu **imprécise**, cela ouvre la voie à d'autres solutions lorsque le consentement n'est pas possible à obtenir. Attention toutefois à ne pas abuser de cette ouverture. Il faut toujours veiller à présenter des **garanties suffisantes** pour préserver la vie privée des personnes.

Dans le même esprit, l'obligation de finalité unique avait déjà été aménagée par la CNIL avec le principe d'**extension de finalité**. La loi le reprend intégralement en autorisant notamment les traitements réalisés à des fins de recherche, avec les mêmes obligations de garanties.

Concernant les données sensibles, le traitement, normalement interdit, se voit autorisé si les données sont **anonymisées à bref délai**. Il faut encore insister sur les précautions à prendre avant d'utiliser cette possibilité. Mais encore une fois, le consentement n'apparait plus comme la seule possibilité.

II) la recherche de garanties

Toutes ces avancées viennent de la réflexion engagée dans le domaine de la recherche médicale. Beaucoup des exceptions dégagées ne s'appliquent pas aux besoins des corpus oraux, mais il faut insister sur deux points :

- l'*anonymisation* : (cf fiche). Il s'agit surtout de rendre les données accessibles ou non selon la qualité des personnes qui les consultent.
- Le développement de *règles et d'usages professionnels* : la CNIL sera plus encline à accepter des alternatives au principe du consentement si une profession s'engage sur des garde-fous et des garanties pour la protection des personnes.

L'important est la réalisation d'une collecte loyale en vertu de fondements légitimes qui concilient les intérêts de tous.

LES BASES DE DONNÉES

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. Il est peu d'activités qui n'utilisent pas de bases de données. L'utilisation des nouvelles technologies n'a fait qu'amplifier le phénomène ; les bases de données constituent d'ailleurs un des principes fondamentaux d'Internet.

Pour en constituer une, il faut des données mais aussi une structure pour les ordonner. Lorsque l'on se pose la question de savoir à qui appartient la base, ainsi que les données qui la composent, le droit d'auteur n'apporte qu'une réponse partielle puisqu'il ne prend pas en compte les investissements en temps et en argent qui ont été nécessaires. C'est pourquoi il a fallu adopter un régime spécifique, mais qui coexiste malgré tout avec celui du droit d'auteur.

Première partie : un cas à part du droit d'auteur

I) La titularité des droits

C'est l'alinéa 2 de l'article L112-3 du Code de Propriété Intellectuelle (CPI) qui 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*".

La définition est suffisamment ouverte pour pouvoir s'appliquer à toute forme de communication.

La majorité des articles du CPI s'articulent autour de la notion d'auteur, mais dans ce cas particulier la définition clé est celle de **producteur** Selon l'article L341-1 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. On privilégie avant tout l'investisseur et non pas l'auteur.

D'ailleurs l'article L341 CPI précise qu'il faut un **investissement** financier, matériel ou humain **substantiel**. 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, car il y a juste eu extraction de données. L'élément déterminant c'est la prise de risque. Il ne suffit pas de démontrer l'existence de charges, il faut aussi mettre en évidence le volume de travail, ainsi que la volonté de promotion.

II) La définition des devoirs et des droits

Le producteur doit 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).

Le droit principal concerne l'interdiction d'extraction et de réutilisation. Tout dépend du volume de données concerné. Les articles 341-1 et 341-2 CPI parlent de quantité "qualitativement ou quantitativement substantielle". Le terme substantiel s'apprécie au cas par cas. Ainsi, 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 une obligation. Le producteur peut céder ses droits, ou plutôt autoriser – moyennant contrepartie – ces extractions. La protection court sur 15 ans. Si à la fin de la période, il y a un nouvel investissement substantiel, la protection se voit renouvelée.

En cas d'infractions les peines sont de 2 ans de prison et 150 000 euros d'amende. Pour les personnes morales l'emprisonnement se transforme en interdiction d'exercice.

Deuxième partie : la coexistence avec les autres droits et les limites

I) Le droit d'auteur

L'auteur et le producteur bénéficient de droits différents. Mais la principale difficulté vient de ce que ces régimes ne s'excluent pas. Ainsi, une structure d'organisation originale mais n'ayant pas bénéficié d'investissement substantiel pourra se trouver protégée par le droit d'auteur et exclue de celui des

bases de données. Toutes les situations sont possibles. Il faut donc juger au cas par cas pour étudier quel mode de protection correspond le mieux au but recherché.
Les droits du producteur ne doivent pas non plus porter atteinte au jeu normal de la concurrence.

II) Les limites

L'extraction ne peut être interdite si elle porte sur une partie non substantielle et qu'elle est effectuée par un utilisateur qui y a légalement accès. Il en va de même pour une base de données, non électronique, si l'usage en reste privé. Il faut noter que la directive européenne de 1996 prévoyait 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.
Le producteur ne bénéficiera plus de son droit d'interdiction dès la vente de la première copie matérielle de la base, sur tout le territoire de l'Union européenne. Cela ne s'applique pas aux transmissions électroniques de cette même base.

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 l'utilisation des données recueillies. La loi lui fixe donc des obligations. La directive européenne 95/46/CE ("Flux de données transfrontières") dans son article 2d, repris pour la refonte de loi "informatique et libertés" donne la définition suivante :

> "***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.***"

Première Partie : les principes généraux

Le responsable du traitement se doit donc de veiller à :

1. la qualité des données ;
2. à leur(s) finalité(s) ;
3. au recueil du consentement.

I) la qualité des données

Pour pouvoir être traitées les données doivent avoir été recueillies selon un ensemble de principes qui garantissent la protection des personnes. La directive ne modifie pas ces conditions déjà présentes dans la loi de 1978.

- **Loyauté :** Toute donnée collectée doit avoir été recueillie loyalement. Ceci suppose une information préalable, une demande écrite de 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 surtout au contexte dans lequel s'est effectuée la collecte. La mécanique législative mise en place suppose que la finalité soit invariable. Or une base de données peut avoir de multiples usages. Comment faire alors pour éviter de se retrouver sanctionné pour détournement de finalité ? La CNIL cherche les indices d'une éventuelle déloyauté si la personne concernée n'a pas les moyens de s'opposer ou si on lui cache volontairement la finalité du traitement. On peut tout de même rapprocher l'article 11-1 qui prévoit que si la personne n'a pas été avertie, le responsable du traitement peut le faire avant la première communication à un tiers.
- **Caractère adéquat, pertinent et non excessif**. 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 ou obligé de revoir ses questionnaires car les données collectées étaient jugées trop nombreuses ou inutiles par rapport à la finalité annoncée. Plus l'on acquiert de données sur un même individu, plus le risque est grand de voir ce traitement surveillé étroitement, voire refusé par la CNIL.
- **Exactitude.** Les données doivent être exactes et mises à jour. Cela renvoie au droit d'accès, d'opposition et de rectification ouvert à chaque personne concernée par le traitement.
- **Confidentialité et Sécurité.** 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 le type de données ou de traitement effectué. Les données peuvent être conservées sur une longue durée en respectant les prescriptions de la loi sur les archives, si elles le sont à des fins de recherche (art 36 de la loi de 78 consolidée). L'article 13 organise divers régimes dérogatoires surtout liés aux intérêts vitaux ou à la politique des Etats-Membres.

II) la ou les finalités

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 traitement. 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. Pas question de fournir une justification **après** les travaux car cela constituerait un **détournement de finalité** passible de sanctions pénales.
La constitution de corpus oraux requérant parfois un grand nombre de données, avant de recueillir tel ou tel type de données, il est important de s'assurer qu'il sera **véritablement utile** à la recherche poursuivie.

III) le consentement

C'est l'un des principes **fondamentaux** de tous les textes régissant les traitements de données. La règle générale veut que le consentement soit "*éclairé*" et suppose une **information préalable**. S'il n'a pas été possible de l'obtenir avant le recueil de données, le responsable du traitement doit demander le consentement avant la première communication au chercheur ou avant le premier traitement. Toute personne ayant consenti au recueil de données dispose d'un **droit de rétractation**. Veuillez vous reporter à la fiche qui lui est consacré pour plus d'informations. **toute donnée recueillie illégalement peut entraîner des poursuites pénales pour le responsable du traitement ; mais aussi empêcher la valorisation du travail de recherche**

DEUXIEME PARTIE : LES PROCEDURES DE DECLARATION

> "Tout traitement de données personnelles à caractère direct ou indirect doit être déclaré à la CNIL".

La distinction entre secteur public et secteur privé se trouve supprimée. Tout dépend du type de traitement, de la finalité, des destinataires, du type de données et de leur conservation.
Si en fonction de ces critères, il n'y a pas d'atteinte à la vie privée et aux libertés, la CNIL accorde une dispense (art 24-II). S'il s'agit de traitements courants (*voir le site de la CNIL*), une déclaration simplifiée suffira. Pour d'autres types de traitement la règle, c'est la déclaration. **sauf** :

- utilisation de données sensibles (autorisation),
- utilisation de données relatives à des infractions, des condamnations,
- Utilisation de données comportant des appréciations sur les difficultés des personnes.

La législation relative aux traitements informatiques encadre plus spécifiquement certaines données ou certains types de traitement. Ainsi, l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques, l'appartenance syndicale et les données génétiques sont a priori exclues de toute collecte. Il est cependant possible de faire pour cela une demande particulière auprès de la CNIL.

Le Patrimoine immatériel et l'UNESCO

Les corpus oraux ou corpus de langue parlée sont "toujours des enregistrements de données sonores, éventuellement accompagnés de données visuelles (prise en vidéo, ou à la télévision), presque toujours accompagnés de transcription et traitements informatisés[25]". Ces corpus sont constitués d'une grande variété de données sonores allant de la parole simple au chant en passant par les contes et les extraits de musique recueillis dans des situations tout aussi variées (protocole expérimental mis en place par le chercheur ou parole recueillie en situation normale).

Diverses recherches, dans diverses disciplines, peuvent porter sur ces corpus et viser des objectifs différents. Dans le cadre de la recherche linguistique, la constitution de grands corpus oraux

> "servent en premier lieu de documentation générale sur la langue nationale…Une documentation sur la langue parlée est parfois le point de départ pour lancer des activités nouvelles : des corpus de langue ont servi de bases pour diffuser des langues peu (ou pas du tout) écrites, comme on l'a fait pour la langue maori[26]".

Ces corpus peuvent déceler de nouveaux problèmes linguistiques et favoriser l'émergence de nouvelles disciplines (recherche en pragmatique par exemple).

La langue, qui sert de base à la constitution des corpus oraux, fait l'objet d'une appréhension au regard de plusieurs textes de l'Unesco. La caractéristique essentielle de ces textes est de cerner la langue d'un point de vue collectif en ceci qu'elle fait partie du patrimoine culturel d'une communauté. Ainsi, le point A de la recommandation sur la sauvegarde de la culture traditionnelle et populaire[27] définit la culture traditionnelle et populaire comme "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".** La déclaration universelle de l'Unesco sur la diversité culturelle[28] considère quant à elle 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 5e 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".** Dernière en date, la convention pour la sauvegarde du patrimoine culturel immatériel[29]. Par patrimoine culturel immatériel il faut entendre selon l'article 2 de ladite convention :

> "les pratiques, représentations, expressions, connaissances et savoir-faire - 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-faire liés à l'artisanat traditionnel.

Les composantes sus-citées du patrimoine culturel (y compris la langue pour ce qui nous concerne) 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

---

[25] Claire Blanche Benveniste, Les sciences du langage et les corpus oraux

[26] Idem

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

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

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

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. Lesquelles ont été publiées et commentées[30] par l'Unesco en collaboration avec l'OMPI[31]. 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[32], pour définir le folklore, se réfère entièrement à la convention de 1989 sur la culture traditionnelle et populaire. D'ailleurs, selon les dispositions types de législation nationale[33], les œuvres du folklore sont considérées comme faisant partie intégrante du patrimoine culturel. Ses éléments comprenant les contes populaires, les chansons, la musique, les danses, les dessins, etc.[34] Par expressions du folklore, en effet, 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
- 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.

La conséquence première inhérente à cette qualification de la langue comme élément du patrimoine d'une communauté culturelle est de la placer sous un régime de sauvegarde. Ainsi, la recommandation de 1989 porte t-elle sauvegarde de la culture traditionnelle et populaire. Quid de la convention de 2003 sur le patrimoine culturel immatériel ?

Diverses mesures sont prévues au titre de cette sauvegarde. Au sens de la convention de 2003 sur le patrimoine culturel immatériel, il s'agit de l'identification, la documentation, la recherche, la préservation, la protection, la promotion, la mise en valeur, la transmission, etc.[35] Si l'on a des doutes sur le sens du mot recherche cité ici, celui-ci peut-être éclairé au regard de l'article 13c portant autres mesures de sauvegarde qui parle des études scientifiques. Donc la recherche scientifique et partant la recherche en ethnolinguistique peut constituer un moyen de sauvegarde du patrimoine culturel.

Avec les autres mesures prévues pour la sauvegarde, la recherche entretient un rapport particulier. Si la recherche est un moyen de sauvegarde au même titre que la conservation, la préservation ou la promotion, ces différentes mesures peuvent trouver leur consécration avec et à travers la recherche. A cet égard, la recommandation de 1989 sur la culture traditionnelle et populaire est assez probante. En effet, la recherche[36] intervient dans les différentes mesures de sauvegarde. Ainsi, la préservation exige des recherches appropriées pour créer des systèmes d'identification et d'enregistrement (collecte, indexation, transcription)[37]. On peut tout aussi bien imaginer la nécessité de recherches pour conserver le patrimoine culturel.

Une seconde conséquence non incompatible avec la première consiste à voir dans les éléments du patrimoine culturel l'expression, la manifestation de la créativité intellectuelle individuelle ou collective. Ce qui justifie une protection devant s'inspirer de celle qui est accordée aux productions intellectuelles[38]. C'est évoquer ici la délicate question des droits de propriété intellectuelle.

---

[30] Dispositions types de législation nationale sur la protection du folklore contre leur exploitation illicite et autres actions dommageables, document officiel publié par l'Unesco et par l'OMPI, 1985, 88p

[31] Organisation mondiale de la propriété intellectuelle

[32] Shyllon Folarin, Conservation, préservation et protection juridique du folklore en Afrique, in Bulletin du droit d'auteur, volume XXXII, No 4, p.41

[33] Précité. On peut citer aussi des références nationales proprement dites. Selon l'article 2 de la loi angolaise sur le droit d'auteur, le folklore s'entend de toutes œuvres littéraires, artistiques et scientifiques transmises de génération en génération et constituant l'un des éléments fondamentaux du patrimoine culturel traditionnel. L'article 6 de la loi gabonaise de 1987 sur la protection du droit d'auteur et des droits voisins prévoit la même chose.

[34] Dispositions types précité, p.33

[35] Cf. Article 2.3

[36] 'Recherches appropriées' pour reprendre les termes du texte

[37] Point Bc de la recommandation

[38] Point F recommandation 1989 et aussi les dispositions types p.37

Présentation de quelques lois africaines sur la protection du patrimoine culturel

Relais aux textes d'ordre international, si on s'intéresse maintenant aux lois qui ont été adoptées au sein des différents Etats (les Etats africains notamment), on constate que celles-ci ont très tôt suivi le modèle de protection conféré par le droit d'auteur[39] corroboré dans le contexte typiquement africain par la Loi Type de Tunis sur le droit d'auteur à l'usage des pays en développement[40]. Cette tentative de protection par le droit d'auteur a été souvent jugée peu efficace, voire peu rationnelle. C'est d'ailleurs ce qui a justifié la recherche d'un système adéquat pour les aspects propriété intellectuelle de la protection des expressions du folklore dont ont voulu faire figure les Dispositions types de législation nationale de 1985. Une autre critique fut adressée aux législateurs nationaux au motif que ceux-ci, "pressés de souligner que le folklore est un aspect important de l'identité culturelle,…ont relégué à l'arrière plan la conservation et la préservation[41]". Ces dernières, avec l'identification et la diffusion, prennent place dans le contexte plus général de la protection juridique du patrimoine culturel. On dénombre de nombreuses législations nationales qui protègent le patrimoine culturel national dont l'Unesco a constitué une base de données[42]. Si certaines de ces lois visent expressément la protection du patrimoine culturel (Burundi, 1983 ; Burkina Faso, 1985 ; Mali, 1985 ; Côte d'Ivoire, 1987 ; Comores, 1994, etc.), d'autres portent plus spécifiquement la protection des antiquités (Egypte, Ethiopie, etc.) ou les monuments historiques et les sites (Maroc, Cameroun notamment). A l'analyse de ces textes un constat général s'impose : le caractère lacunaire des dispositions concernant expressément le patrimoine immatériel. C'est la protection du patrimoine matériel qui semble attirer l'attention du législateur. La plupart des articles premiers définissant l'objet de la loi visent les biens meubles et immeubles. La seule exception stricte étant la loi ivoirienne qui, dans son article 1er, cerne "l'ensemble des biens mobiliers et immobiliers, des arts et traditions populaires, des styles, des formes, etc." L'alinéa de l'article 2 renvoie aux œuvres du folklore, compris dans le patrimoine national, et faisant par ailleurs lieu d'une protection au titre de la loi sur la protection des œuvres de l'esprit. La seule disposition de la loi applicable aux traditions populaires est consignée dans l'article 4 qui prévoit l'établissement annuel d'un inventaire des arts et des traditions populaires, le reste du texte visant le patrimoine matériel. On relève une disposition sur la recherche scientifique (article 60 sur le classement et l'exportation des biens culturels mobiliers). Il s'agit du droit reconnu à l'Etat de photographier toute pièce de stocks des antiquaires aux fins de la documentation iconographique et de la recherche scientifique. On peut trouver des dispositions se rapportant au patrimoine immatériel dans la loi malienne. Non pas à l'article 2 qui, lui, vise les biens meubles et les biens immeubles mais aux articles 36 et 37 sur la promotion du patrimoine culturel. Selon ces articles, l'Etat reconnait à tout citoyen un droit d'accès aux valeurs du patrimoine en assurant la fixation par l'image et le son des traditions culturelles de la nation. On retrouvera les dispositions qui s'appliquent à la recherche scientifique dans le décret 90-332 portant fonctionnement du musée national. Selon l'article 5 de ce décret, le musée est doté d'une section documentation et recherche qui est chargée, entre autres, de participer à la recherche en ethnologie, en archéologie et en histoire. Mais les modalités de cette participation ne sont pas précisées. On peut essayer de trouver des indices dans l'ordonnance 90-43 portant création du musée national. L'article 2 dispose que le musée a pour mission, entre autre, la collecte, la conservation et la diffusion du patrimoine culturel national archéologique, traditionnel et contemporain sous forme de biens culturels, de documents iconographiques et visuels. La participation pourra consister, dans le cadre de la recherche, à garantir l'accès aux éléments du patrimoine.

La loi égyptienne ne concerne pas le patrimoine immatériel. Des dispositions existent concernant la recherche scientifique mais il s'agit de celles visant à la mise au jour des antiquités (article 5). Comme la loi égyptienne, la loi éthiopienne vise à protéger les antiquités. C'est à ce titre que des recherches peuvent être autorisées (article 16). Les articles 23 (publication of report and results of studies) et 24 (ownership over results of studies) sont relatifs au droit exclusif du chercheur quant à la publication des rapports et des résultats de sa recherche et au droit de propriété sur les résultats de la recherche.

---

[39] Le § 5 des observations liminaires sur les Dispositions types donne une liste de pays à laquelle on pourra se reporter. Trois exemples : Tunisie, 1967 ; Bolivie, 1968 ; Côte d'Ivoire, 1978.

[40] Elaboré conjointement par l'Unesco et l'OMPI en 1976

[41] Shyllon Folarin, article précité, p. 42

[42] Base de données des législations nationales sur la protection du patrimoine culturel, http://portal.unesco.org/culture/fr/file

Ainsi donc, au-delà du caractère lacunaire des dispositions sur le patrimoine immatériel, au niveau international comme au niveau interne, la langue fait l'objet d'une double appréhension. Ce sont ces deux qualifications juridiques de l'oralité comme élément du patrimoine culturel tout d'abord, et comme manifestation de la créativité intellectuelle ensuite qui vont se conjuguer pour nous permettre d'entrevoir les principes posés pour la recherche. L'analyse des textes laisse poindre diverses mesures en faveur de la recherche, dont celle mettant à la charge des Etats la promotion de la recherche. Cela ne va pas sans contraintes pour les chercheurs qui sont appelés à se doter d'une éthique dans la conduite de leurs travaux. C'est pour nous l'occasion d'envisager maintenant les droits et devoirs du chercheur inhérents à la recherche sur le patrimoine oral.

6.2 Fiches techniques

Eléments de méthode pour la prise de son et l'enregistrement de terrain

Luc Verrier, Alain Carou, BnF, avril 2005.

1. 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'émissions 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 :

- facile d'utilisation (pour éviter les erreurs de manipulation en cours d'enregistrement) ;
- autonome (batterie suffisante) ;
- robuste ;
- portable ;
- 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 (cf. fiche "Supports de collecte, supports d'archivage) ;
- 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 ;
- 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.

2. Comparaison des types d'enregistreurs disponibles

➢ ***Enregistreur dédié sur mémoire 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 2005) ;
- 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, Fire Wire, SCSI).
Ces appareils possèdent généralement des entrées et sorties audio numériques (SPDIF, AES) permettant un raccordement et un transfert des données sans passer dans 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 marché du précédent) :***

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 à 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" 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 décharger 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.

	Enregistreur dédié Flash, MicroDrive, disque dur	Baladeur MP3	Ordinateur portable	Nagra analogique	DAT	Cassette audio	MiniDisc
Technologie actuelle ou en fin de vie	actuelle	actuelle	actuelle	fin de vie	fin de vie	fin de vie	fin de vie
Prix	1 000-10 000 €	150-450 €	ordinateur + 200 € d'équipement spécifique	1 000 € (occasion)	1 500 €	700 € (si appareil neuf)	150 €
Facilité d'utilisation	semi-pro et pro	grand public	grand public ou semi-pro	pro	semi-pro	grand public	grand public
Ergonomie	+	-	+	+	+	+	-
Capacité	selon disque dur	selon disque dur	selon disque dur	30 min	2 h	1-2 h	80 min
Formats d'enregistrement	wave, BWF, MP2, MP3	wave, MP2, MP3…	potentiellement tous	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	-

(essai de synthèse comparative)

3. Conseils pratiques

➢ **Médias 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). Eloigner des sources de magnétisme et de chaleur.

Eviter de stocker beaucoup plus que sa consommation.

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 (100000 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 placés proches 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: chère mais très pratique (spectacle, cinéma, TV …)
- bonnette (anti-vent, plop voix…)
- suspension élastiques pour micro (isolation mécanique des vibrations)
- filtre adaptateur coupe bas (vent, plop)

➢ Réglages

Faire des essais avant de lancer l'enregistrement. Bien se préparer (longueur de câbles, batterie chargées, bloc secteur, cassettes vierges de durée suffisante, …)
En numérique, vérifier que l'appareil est bien paramétré : le bon format et la 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 part dans le 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 enregistrement" (i.e. possibilité de contrôler ce qui est effectivement enregistré 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 deux entrées micros : 1) interviewé 2) les questions (on peut utiliser une petite mixette pour plus de sources).
Faire attention aux bruits de manipulations 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 (souplesse au niveau de 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. (Cf. sur ce sujet la fiche "Support d'enregistrement, support d'archivage".)

➢ Transfert et édition

Cette étape est cruciale même si elle parait simple à 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) sur du son : 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 et 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 d'enregistrement, supports d'archivage : Le son

Luc Verrier, Alain Carou, BnF, mars 2005.

I/ Assurer la conservation à long terme du son numérique

1. 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 de choix.

Un support d'archivage numérique doit quant à lui réunir de tout autres qualités :

1. Garantie de pouvoir trouver du matériel de lecture à moyen terme : large diffusion de la technologie, fabrication par plusieurs constructeurs différents.
2. 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.
3. 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 recopies d'information en temps utile.
4. Robustesse (capacité à 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 :
5. Simplicité des opérations de recopie.
6. 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).

> 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 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 (recopies ponctuelles et recopies en masse) en fonction des nécessités.

Alors que dans le monde analogique, chaque génération de recopie est source de perte qualitative, le nombre de recopies est indifférente dans le monde numérique, du moment qu'elles sont engagées à temps.

2. Archivage sur CD enregistrable

N.B. 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 abaisser 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 de 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 les 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 très recommandée.

Les limites du CD-R deviennent manifestes pour de grandes masses de données :
le contrôle sur analyseur et la recopie 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.

3. Archivage de 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 18 mois à 2 ans). Le renouvellement du parc de lecteurs-enregistreurs et la migration des données sur une nouvelle génération de supports sont à prévoir tous les 4 à 6 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 numérique est générateur de couts relativement importants tout au long de la vie de l'information à conserver.

Supports d'enregistrement, supports d'archivage : La vidéo

Alain Carou, Dominique Théron, BnF, avril 2005.

I) Numériser la vidéo pour la sauvegarder

1. Vie et mort des technologies vidéo 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 degrés 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 pour pouvoir se lancer dans l'action.

2. 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.

L'hypothèse Beta numérique
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 se préparer à 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.

Métadonnées
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 dans l'avenir d'engager des actions de pérennisation (par exemple 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.

3. Organisation de la chaine de numérisation

La numérisation se décompose en plusieurs étapes mais a pour règle de base la meilleure relecture possible du document d'origine.

Préparation documentaire et physique des éléments
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 de type Excel est 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.

Numérisation, compression
Arrivent 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).

Contrôle qualité
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.
La restauration ne doit en tout cas pas être considérée comme un substitut à une lecture de qualité, mais comme une opération complémentaire.

Versement dans l'archive numérique
Il n'existe pas de technologie de stockage numérique pérenne. Le stockage numérique est donc affaire de migrations (recopies) en masse périodiques et contrôlées. Rien à voir avec la lourdeur de la recopie d'analogique en numérique : la recopie 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 recopie sans manipulation humaine. Ces technologies sont soumises à un cycle d'obsolescence rapide, qui contraindra à des migrations de masse tous les 5-6 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

6.3 Fiches formats et normes

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 fichiers. Ces derniers sont eux-mêmes stockés sur des supports ayant leur propre organisation physico-logique.

1. 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 détient seule sa description qu'il garde secrète. 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 Etat (normes nationales : par exemple l'AFNOR) ou entre des Etats (normes internationales : par ex l'ISO). On privilégiera les normes internationales aux autres dans la mesure ou elles représentent une 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 définit 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.

2. Les données 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 passées. 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 ni 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 le nombre canaux[43].

43 Pour ces caractéristiques nous conseillons d'adopter celles des CD-Audio (bon compromis qualité/quantité), c'est-à-dire : 44100 Hz, 16 bits, mono ou stéréo (en fonction des conditions de l'enregistrement).

Il est aussi d'usage de parler aussi de codages pour les algorithmes de compression que l'on peut appliquer aux données. Ces algorithmes proposés dans des programmes appelés *codec* sont généralement avec perte d'information (MACE, MPEG, u_law, etc.), c'est-à-dire que le résultat de la décompression des données ne donnera pas la même chose que l'original. En général, une bonne compression pour de la parole ou pour de la musique propose de supprimer en priorité les informations que la physiologie de l'oreille humaine ne permet pas d'entendre (cf. codages MPEG). Ces algorithmes ont pour but de diminuer la taille des fichiers ou du débit des transferts. Pour de la conservation de document il est bien évident que l'on ne se tournera pas vers de telles solutions. C'est pour cette même raison que l'on évitera l'utilisation d'outils comme les enregistreurs miniDisc qui utilisent dès la source un algorithme de compression.

3. 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 plutôt qu'un autre 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 liées à des techniques soumises à des brevets que vous pouvez peut-être acquitter mais que d'autres que vous ne pourront peut-être pas. 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 logiciels ne sont pas tenus de maintenir leur logiciels).

4. Les annotations linguistiques

Les codages

Les annotations linguistiques sont composées de définitions d'objets linguistiques (les mots, les morphèmes, les tours de parole, etc.) ainsi que 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[44]) est synchronisé sur la norme ISO-10646 qui a le même objectif. Il est déjà largement utilisé et a été adopté notamment par le web. Il permet donc de coder des documents multilingues mélangeant des écritures aux caractères et aux propriétés différentes, et ceci de manière indépendante de la plate-forme informatique utilisée, ce qui facilite l'échange et le partage de documents. Dans la mesure où il n'y a pas d'autres propositions de codage concurrentes, 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. Il existe d'autres initiatives comme le CES (Corpus Encoding Standard). Ces propositions de codage ne sont pas forcément adéquates à toutes les analyses possibles, mais il est bon 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 des métadonnées documentaires.

5. Les formats

44 Site web du Consortium Unicode (http://www.unicode.org)

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 texte. 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 traitements de texte).

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 maintenant 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 logiciels 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 rendue 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[45].

6. Les metadonnées

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 propres 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.

7. Les 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[46] normalisé ISO-15836 en 2003 propose un jeu de 15 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 supports informatiques. Il existe aussi des communautés qui ont proposé des recommandations comme par exemple OLAC[47] (basé sur du Dublin-Core enrichi et spécifié pour l'adapter aux ressources linguistiques), ou IMDI[48].

8. Les 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

45 Document Type Definition
46 Site web du Dublin Core Metadata Initiative (http://dublincore.org)
47 Open Language Archives Initiative (http://www.language-archives.org)
48 EAGLES/ISLE Metadata Initiative : (http://www.mpi.nl/IMDI/)

une base de données ou toute autre solution. Le choix d'une solution plutôt qu'une autre repose sur les critères que nous avons énoncés précédemment.

9. Les protocoles d'échange

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 est une organisation plus récente qui définit entre [?]un protocole relativement simple pour la récolte de métadonnées qui 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 performances lorsque certains nœuds du réseau ralentissent ou bloquent la poursuite.

6.4 Bibliographie

Bibliographie générale

ABOU-HAIDAR, L. (dir.), *2002, Transcription de la parole normale et pathologique.* Revue Parole, n° 22/23/24.

AUER, P. (1993), Ueber, *Zeitschrift für Literaturwissenschaft und Linguistik (Lili)*, 90/91, 104-138.

AUER, P. et alii ,1999, *Language in Time. The Rhythm and Tempo of Spoken Interaction.* Oxford : Oxford UP.

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.

BERGOUNIOUX, G. (dir), *1992, Enquêtes, Corpus et Témoin, in Langue Française n°93,* Larousse, Paris.

BERGMANN, J. R., 1985, *Flüchtigkeit und methodische Fixierung sozialer Wirklichkeit,* In W. Bonss & H.

BANGE, P., 1983, *Points de vue sur l'analyse conversationnelle,* DRLAV 29, 1-28.

BIBER D., 1985, *Variations across spoken and written language,* Cambridge : Cambridge U. P.

BIBER, D. and alii, 1999, *Longman Grammar of Spoken and Written English.* London : Longman.

BILGER M. (dir), 2000, *Linguistique sur corpus, études et réflexions*, Cahiers de l'université de Perpignan, Presses universitaires de perpignan.

BILGER M. (ed) 2000, *Corpus, Méthodologie et applications linguistiques*, Edition Champion, Paris.

BLANCHE-BENVENISTE, C. et JEANJEAN, C., 1987, *Le français parlé : transcription et édition.* Paris : Didier-Erudition.

BLANCHE-BENVENISTE, C, BILGER, M., ROUGET, C. Et van den EYNDE, K., 1999, *Le Français Parlé: Etudes grammaticales.* Paris : CNRS-Editions.

BLANCHE-BENVENISTE, C., ROUGET, C. et SABIO, Frédéric, 2001*, Choix de textes de français parlé : trente-six extraits*, Paris : Champion.

BOURDIEU, P., & alii, e. 1993. *La misère du monde*. Paris: Seuil.

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.

COUPER-KUHLEN, ELIZABETH & SELTING, MARGRET, (eds.), 1996, *Prosody in Conversation: Interactional Studies*, Cambridge, CUP.

CRESTI, E. and MONEGLIA, M. (eds.), 2005, *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam / Philadelphia : Benjamins.

DELAIS-ROUSSARIE E., DURAND J., (eds), 2003, *Corpus et variation en phonologie du français*, Presses universitaires du Mirail, Toulouse.

DURANTI, A. (1997). *Linguistic anthropology.* Cambridge: Cambridge University Press.

DREW, P.& HERITAGE, J. (eds.), 1992. *Talk at work. Cambridge* : CUP.

EDWARDS, J. ET LAMPERT, M.D. (eds.), 1991, *Transcription and Coding Methods for Language Research*, Hillsdale : Erlbaum.

ENCREVE, P., FORNEL de, M., 1983, *Le sens en pratique, IN ARSS n°46, L'usage de la parole*, Le Seuil, Paris.

FORNEL de, M., 1994, Le cadre interactionnel de l'échange visiophonique, Réseaux n°64, CNET

HABERT,B., NAZARENKO,A. et SALEM, A ., 1997, *Les linguistiques de corpus*. Paris :Colin.

HAMMERSLEY, M., ATKINSON, P.,1995. *Ethnography: Principles in Practice.* London: Routledge (2d edition).

JACOBSON, M., *Corpus oraux en linguistique de terrain. Traitement automatique des langues. 45/2, 2004,* p. 63-88.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., 1983, *Issues in the transcription of naturally occurring talk: caricature versus capturing pronunciational particulars*, Tilburg Papers in Language and Literature, 34.

JEFFERSON, G., 1985, *An Exercise in the Transcription and Analysis of Laughter.* in T. van Dijk (ed.), Handbook of Discourse Analysis Volume 3, New York : Academic Press, 25-34.

KALLMEYER, WERNER ET SCHÜTZE, FRITZ. (1976) *Konversationsanalyse, Studium Linguistik 1,1-28.*

KENNEDY, G., 1998, *An introduction to Corpus Linguistics*. London : Longman.

LAHIRE, B. , 1996. *Variations autour des effets de légitimité dans les enquêtes sociologiques*. Critiques sociales, 8-9, 93-101

MARTIN, P.,1987, *Prosodic and Rhythmic Structures in French, Linguistics 925-949*, et logiciel

MELLET, S. (dir), 2002, *Corpus et recherches linguistiques*, Université de Nice, Nice.

BIBLIOGRAPHIE

MITCHELL, R.G.JR., 1991. *Secrecy and disclosure in fieldwork*. In : W.B. Shaffir, R.A. Stebbins (eds.), Experiencing Fieldwork : An inside View of Qulitative Research. London : Sage.

MOERMAN, M., 1988. *Talking Culture: Ethnography and Conversation Analysis*. Philadelphia: 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 (http://www.marges-linguistiques.com), 1(mai).

MONDADA, L., 2002, *Pratiques de transcription et effets de catégorisation*, Cahiers de Praxématique, 39, 45-75.

MONDADA, L. (à paraître a). *La demande d'autorisation comme moment structurant pour l'enregistrement et l'analyse des pratiques bilingues*. Tranel (Université de Neuchâtel).

MONDADA, L. à paraître, *La pertinenza del dettaglio : registrazione e trascrizione di dati video per la linguistica interazionale*. In Y. Bürki, E. de Stefani, eds. Transcriptio. Bern : Lang.

MONDADA, L. (à paraître). *L'analyse de corpus dans la perspective de la linguistique interactionnelle : des analyses de cas singuliers aux analyses de collections*. In. A. Condamine (éd.), *Sémantique et corpus*, Paris : Hermès.

OCHS, E., 1979, *Transcription as theory*, in E. OCHS and B.B. SCHIEFELLIN (eds.), Developmental Pragmatics. New York : Academic Press, 43-72.

OCHS, E., SCHEGLOFF,EMANUELA..& THOMPSON, SANDRA (eds.), *Interaction and Grammar*, Cambridge, CUP, 1996.

ONG, W., 1988, *Orality and Literacy*. London : Routledge

QUÉRÉ, L. et al. (eds.), 1984, *Arguments ethnométhodologiques*, Paris, Centre d'Etude des Mouvements Sociaux, EHESS.

Recherches Sur le Français Parlé, 1984, n° 5. *Pourquoi le français parlé est-il si peu étudié ?*
Revue Française de Linguistique Appliquée ,numéro spécial, 1996 (1-2) et 1999, IV-1.

ROULET, E., AUCHLIN, A., MOESCHLER, J. RUBATTEL, C., SCHELLING, M., 1985. *L'articulation du discours en Français contemporain*, Bern, Lang.

SACKS, H., 1972a, *An initial investigation of the usability of conversational materials for doing sociology*. In D. Sudnow (Ed.), Studies in Social Interaction (pp. 31-74). New York: Free Press.

SACKS, H., 1972b, *On the Analyzability of Stories by Children*. In J. J. Gumperz & D. Hymes (Eds.), Directions in Sociolinguistics: The Ethnography of Communication (pp. 325-345). New York: Holt, Rinehart and Winston.

SACKS, H., 1984, *Notes on methodology*. In J. M. Atkinson & J. Heritage (Eds.), Structures of Social Action (pp. 21-27).

SACKS, HARVEY, SCHEGLOFF, EMANUEL A. ET JEFFERSON, GAIL, 1974*, A simplest systematics for the organization of turn-taking for conversation*, Language 50, 696-735.

SACKS, H., 1984, *Notes on methodology*. In J. M. Atkinson & J. Heritage (Eds.), Structures of Social Action (pp. 21-27), Cambridge: Cambridge University Press. (Edited by Gail Jefferson from various lectures).

SACKS, H., 1992, *Lectures on Conversation* [1964-72] (2 Vols.). Oxford: Basil Blackwell.

SCHEGLOFF, E.A., 1967, *The First Five Seconds: The Order of Conversational Opening*. PhD, University of California/Berkeley.

SCHEGLOFF, E.A., 2000. *Overlapping talk and the organization of turn-taking for conversation*. Language in Society, 29, 1, 1-63.

SELTING, M., 1995, *Der "mögliche Satz" als interaktiv relevante syntaktische Kategori*e, 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.

SINCLAIR, J. 1991, *Corpus, Concordance, Collocation*. Oxford University Press.1991

SINCLAIR, J. 1996, *Preliminary recommendations on corpus Typology*, Technical Report, EAGLES.

SINCLAIR, J. ET COULTHARD, R. M., 1975. *Towards an Analysis of Discourse*, London, OUP.

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 de enregistrements en vidéo*, CALAP 18, 161-177.

WELLAND, T., PUGSLEY, L. (eds.), 2002, Ethical Dilemmas in Qualitative Research, Hants : Ashgate.

Corpus

- PFC, Phonologie du Français Contemporain, (http://www.projet-pfc.net/),
- British National Corpus, British National Corpus, (http://www.info.ox.ac.uk/bnc),
- Corpus de portugais parlé du CLUL (fbacelar.nascimento@clul.cc.fc.ul.pt)
- futur corpus de néerlandais élaboré à Anvers (dijkstra@nwo.nl) comptent 10 millions de mots.
- corpus de français parlé collecté à l'Université de Provence compte 2 millions de mots : www.up.univ-mrs.fr/delic/
- échantillon d'environ 80.000 mots en a été publié dans BLANCHE-BENVENISTE, C., ROUGET, C. et SABIO, Frédéric, 2001, Choix de textes de français parlé : trente-six extraits. Paris : Champion.
- CLAPI compte 300h, parmi lesquelles 1 million de mots alignés avec de l'audio et en partie avec de la vidéo (http://icar.univ-lyon2.fr/).
- Corpus de français collectés en Belgique :
-VALIBEL http://juppiter.fltr.ucl.ac.be/FLTR/ROM/VALIBEL/valibel.html, et ELICOP, et bach.arts.kuleuven.ac.be/pmertens/papers/elicop/pdf, sont parmi les plus importants. La banque de données –

ELRA / ELDA diffusion des corpus et des ressources disponibles dans ce domaine: htttp://www.elra.info

EAGLES (Expert Advisatory Group on Language Engineering Standards), à l'initiative de laCommission Européenne
CHILDES
TEI (Text Encoding Initiative).
UNICODE rassemblant plus de 95.000 caractères : http://www.unicode.org

Aides à la transcription

ANVIL (http://www.dfki.uni-sb.de/~kipp/anvil/)
CLAN (http://childes.psy.cmu.edu/clan/)
ELAN (www.mpi.nl/tools/elan.html).)
PRAAT (www.fon.hum.uva.nl/praat/)
TRANSCRIBER(www.ldc.upenn.edu/mirror/Transcriber/),
WINPITCH (http://www.winnitch.com)

Langues en danger : liens utiles

Les organismes et les institutions finançant la recherche sur les langues en danger mènent des réflexions similaires à celle proposée dans ce guide. Nous donnerons ci-dessous à titre informatif quelques adresses de sites web :

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 Plank 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 Plank 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*]

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*. Paris : Belin, 7 vol., 1 CD-Rom, 1 DVD-Rom

DESCAMPS, F., *L'historien, l'archiviste et le magnétophone, Comité pour l'histoire économique et financière de la France,* Paris 2001, 864 pages.

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*", Culture et Recherche, n° 91-92, 2002).

DURAND, C., 1999-2000, *Folklore et droit d'auteur*, mémoire de DESS, Propriété intellectuelle et communication. Université Montesquieu- Bordeaux IV.

JOUTARD, 1979, P., *Historiens, à vos micros. Le document oral, une nouvelle source pour l'histoire*, L'Histoire, n°12, , p. 106-113.

JOUTARD, P., 1983, *Ces voix qui nous parlent du passé*, Paris, Hachette.

NORA(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, Chantal de, 1990, *Le témoignage oral aux archives...* , Paris, Archives nationales, 1990

VOLDMAN, D., 1992, (dir.) *La Bouche de la vérité ? La Recherche historique et les sources orales*, Les Cahiers de l'IHTP, n° 21, novembre.

MICHEL V., 2002, *Ethnographie de la France : histoire et enjeux contemporains des approches du patrimoine ethnologique*, Armand Colin, collection Cursus, Paris.

Revues et périodiques

Bulletin de l'IHTP, n°1, juin 1980, " Problèmes de méthode en histoire orale", table ronde de l'Institut d'histoire du Temps présent.

Bulletin de l'IHTP, n°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, 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 :

BIBLIOGRAPHIE

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

Equipement 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

6.5 Glossaire

Annonymisation :

Annotation :

Archives :

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". Dictionnaire comparé du droit d'auteur et du copyright.

Balisage :

Corpus alignés :

Corpus de référence :

Corpus oraux :

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 un œuvre composite / samplage) ; Share alike (obligation de rediffuser selon la même licence). Symbole général : cc
Le mouvement Creative Commons propose des contrats-type d'offre de mise à disposition d'œuvres en ligne. Inspirées par les licences de logiciels libres et le mouvement open source, ces textes facilitent l'utilisation et la réutilisation d'œuvres (textes, photos, musique, sites web...). 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…)". Dictionnaire comparé du droit d'auteur et du copyright

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.

Données primaires :

Données secondaires :

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". Dictionnaire comparé du droit d'auteur et du copyright.

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 confère 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". Dictionnaire comparé du droit d'auteur et du copyright.

Droits patrimoniaux : "Droit d'exploitation qui confèrent 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 le 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". Dictionnaire comparé du droit d'auteur et du copyright.

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". Dictionnaire comparé du droit d'auteur et du copyright.

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". Dictionnaire comparé du droit d'auteur et du copyright.

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 la 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)". Dictionnaire comparé du droit d'auteur et du copyright

Droit au respect de l'œuvre : "V. 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". Dictionnaire comparé du droit d'auteur et du copyright.

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 ". Dictionnaire comparé du droit d'auteur et du copyright.

Droit de citation : 3V. 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". Dictionnaire comparé du droit d'auteur et du copyright.

Droit à l'oubli :

Etiquetage :

Floutage :

Langue à tradition orale :

Langues en danger :

Métadonnées:

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". Dictionnaire comparé du droit d'auteur et du copyright.

Pixélisation :

Récits de vie :

Supports d'enregistrements :

Valeur probatoire :

6.6 Index

A

anonymisation · 8, 20, 27, 31, 50, 52, 53, 54, 55, 56, 57, 58, 59, 60, - 82 -, - 83 -, - 84 -, - 87 -, - 123 -, - 124 -
Anonymisation · 5, 52, - 83 -, - 86 -, - 123 -
archives · 22, 26, 28, 42, 51, 52, 63, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 78, - 90 -, - 102 -, - 104 -, - 105 -, - 110 -, - 111 -, - 115 -, - 116 -
Archives de France · 67, 69, - 116 -
Archives nationales · 69, 75, - 115 -
auteur · 25
autorisation · 9, 15, 17, 26, 27, 36, 38, 39, 41, 45, 46, 47, 48, 49, 50, 52, 66, 77, - 84 -, - 85 -, - 91 -, - 113 -, - 117 -, - 123 -

B

B.n.F · 62
base de données · 7, 19, 25, - 85 -, - 88 -, - 89 -, - 90 -, - 94 -, - 111 -
British National Corpus · 15, 18, - 114 -

C

CHILDES · 17
CLAPI · 14, - 114 -
CNIL · 27, - 82 -, - 83 -, - 84 -, - 86 -, - 87 -, - 90 -, - 91 -
cobayes · 13, 37
codage · 30, 31, 33, 59, - 84 -, - 108 -, - 109 -
Code de la Propriété intellectuelle · 64, 73, 77
Code de Propriété Intellectuelle · 37, - 88 -
Computer Supported Cooperative Work · 36
consentement · 9, 10, 15, 27, 36, 39, 43, 45, 46, 47, 77, - 84 -, - 86 -, - 87 -, - 90 -, - 91 -, - 123 -
Consentement · 27
consentement éclairé · 36, 39, 45, 46, 47, - 123 -
CORAL-ROM · 19
corpus alignés · 17
Corpus d'Orléans · 12, 13
corpus de référence · 7, 18
corpus ouverts · 15
creatives commons · 23

D

Déclaration de Berlin · 22
DELIC · 34
dépôt légal · 65, 71, 72, 74
Dépôt légal · 72
Dépôt Légal · 72, 73
Dialogue Homme Machine · 36
diffusion · 5, 7, 10, 18, 20, 22, 23, 24, 27, 28, 29, 31, 35, 36, 37, 42, 48, 49, 52, 53, 65, 71, 72, 80, - 83 -, - 92 -, - 94 -, - 101 -, - 106 -, - 114 -, - 122 -
domaine public · 23, 24, 25, 27, - 85 -
données personnelles · 24, 27, 34, 35, 37, 53, 54, - 82 -, - 84 -, - 86 -, - 91 -
données premières · 53
données primaires · 31
données secondaires · 31
droit à l'oubli · 27
droit à la copie privée · 27
droit d'auteur · 9, 23, 24, 25, 26, 27, 76, 77, - 88 -, - 93 -, - 94 -, - 115 -, - 117 -, - 118 -, - 119 -
droit de citation · 7, 27, - 85 -
droit de divulgation · 26, - 117 -, - 118 -
droit de la propriété intellectuell · 24
droit moral · 24, 25, 26, 35, 77, - 85 -, - 117 -, - 118 -
droits de propriété intellectuelle · 23, - 93 -
droits patrimoniaux · 24, 25, 26, - 118 -

E

EAGLES · 21
ELRA/ ELDA · 18
empowerment · 51
enquêtés · 9, 37, 38, 39, 45, 46, 47, 48, 49, 51, 56, - 123 -
entretien · 34, 43, - 96 -, - 97 -, - 122 -
EuroSpeech 2003 · 16
extension de finalité · 27, - 87 -

F

fieldwork · 38, 41, - 113 -
finalités de l'enquête · 47, 49
formats · 5, 32, 33, 49, 53, 54, 61, - 100 -, - 101 -, - 102 -, - 104 -, - 106 -, - 108 -, - 109 -, - 110 -, - 124 -
Français Fondamental · 12, 13

G

GAT · 17
grapho-lectes · 12
groupe de travail · 8

I

ICOR · 17
INA · 3, 8, 71, 72, - 123 -
Inathèque · 71, 72, 73, 74, - 123 -
informateurs · 20, 34, 37, 40, 41, 42, 43, 45, 46, 50, 51, - 123 -

L

Language Resources · 13
langues à tradition orale · 20, 35
langues en danger · 20
le droit à la paternité · 26
le droit au respect de l'œuvre · 26
le droit de divulgation · 26
le droit de repentir et de retrait · 26
lieux publics · 39, 43
locuteurs · 13, 14, 15, 16, 18, 19, 20, 24, 37, 58, 59, 60, 61, - 96 -, - 123 -
loi Informatique et libertés · 27
Longman Grammar of Spoken and Written English · 18

M

magicien d'Oz · 39
métadonnées · 14, 32, 39, 53, 57, 58, 63, 66, 75, - 100 -, - 105 -, - 106 -, - 109 -, - 110 -, - 111 -
modes d'enregistrement · 30
musées de France · 70

O

œuvres · 26
original · 25, 31, 55, 56, 61, 63, - 109 -
originale · 25

P

paradoxe de l'observateur · 14, 38
Patrimoine · 64, 70, 76, - 92 -, - 115 -, - 123 -, - 124 -
patrimoine immatériel · 21, 79, - 94 -, - 95 -
PFC · 13, - 114 -
Phonothèque Nationale · 62
populations captives · 38
Praat · 17

Q

questionnaire · 29, 34, 77, - 122 -

R

récits de vie · 14, 35, 63, 67, 77, - 122 -
Rémunération · 42, - 123 -
responsable du traitement · - 82 -, - 86 -, - 87 -, - 90 -, - 91 -, - 117 -
rétractation · 44, 45, 46, 48, - 91 -

S

SpeechDat Exchange · 13
SpeechDat Exchange Format · 18
Standardisation · 31, 32, - 122 -
Supports optiques · 30

T

TEI · 16, 17, 33, - 109 -, - 110 -, - 114 -
témoins · 8, 37, 38, 41, 63, 78
Transcriber · 17
transcription · 13, 16, 17, 20, 30, 32, 33, 48, 52, 53, 55, 58, 59, 60, 61, - 92 -, - 93 -, - 112 -, - 113 -, - 114 -, - 123 -, - 124 -
transcriptions · 13, 14, 16, 17, 19, 20, 31, 42, 49, 51, 53, 56, 57, 58, 60, 61, - 109 -

U

UNESCO · 20, 24, 50, 51, 64, 71, - 92 -, - 114 -, - 115 -, - 124 -

V

valeur probatoire · 69, 76
vidéo · 13, 14, 17, 31, 32, 33, 39, 43, 44, 49, 50, 53, 55, 56, 58, 69, 70, - 92 -, - 104 -, - 105 -, - 106 -, - 107 -, - 113 -, - 114 -, - 124 -
vie privée · 27, 34, 36, 43, 52, 60, 66, 69, 77, - 82 -, - 86 -, - 87 -, - 91 -

X

XML · 32, - 105 -, - 110 -, - 111 -

6.7 Table des matières

1 Présentation 5

1.1 Les objectifs du "Guide des bonnes pratiques". 5

1.2 Les conditions d'élaboration de ce Guide 5

1.3 Les aspects juridiques 5

1.4 Les autres aspects de la collecte et de l'usage de données orales 6

1.5 La méthode 7

1.6 Le cadre juridique français 7

1.7 Un "guide des bonnes pratiques"? 8

1.8 Quelques questions fréquentes 8

2 Le contexte scientifique, politique, juridique et institutionnel ... 10

2.1 Les sciences du langage et les corpus oraux 10

Type de données et de locuteur 11
Dimensions 13
Transcriptions. 14
Exploitations et résultats 16

2.2 Cadres politiques de la diffusion de la recherche 20

2.3 Cadres juridiques de la constitution, de l'exploitation et de la diffusion des corpus oraux 21

3 La démarche (constitution, exploitation, conservation, diffusion) 27

3.1 Introduction 27

3.2 Eléments de la situation en jeu 27

Corpus et type de données 27
3.2.1.1 Les modes d'enregistrement 28
3.2.1.2 Les supports d'enregistrement 28
3.2.1.3 Les critères de choix 29
Standardisation des annotations 29
3.2.1.4 Données primaires vs. données secondaires 29
3.2.1.5 Explicitation de la structure des données 30
3.2.1.6 Standardisation / Normalisation 30

3.3 Techniques d'enquête, recueil et production de données 31

3.3.1.1 Le questionnaire 32
3.3.1.2 L'entretien 32
3.3.1.3 Le recueil de contes, chants,... 32
3.3.1.4 Les récits de vie 33
3.3.1.5 L'enregistrement en laboratoire selon un protocole expérimental. 33
3.3.1.6 L'enregistrement d'activités provoquées dans des contextes sociaux, avec éventuellement des tâches/consignes proposées par le chercheur. 33
3.3.1.7 L'enregistrement d'activités dans leur contexte ordinaire et non provoquées par le chercheur. 34
3.3.1.8 La reprise d'enregistrements 34
3.3.1.9 La reprise d'enregistrements médiatiques 34

Rôles des participants 35
3.3.1.10 catégories de participants 35
Lieux 37

3.4 Recueil de données et pratiques de terrain 38
Modes d'approche des personnes concernées par l'enquête 38
3.4.1.1 Temporalité des modes d'approche et typologie des relations avec les informateurs 38
3.4.1.2 Les personnes contactées 39
3.4.1.3 Rémunération 40
Choix du dispositif d'enregistrement : modalités et contextes 41
3.4.1.4 Contexte de l'enregistrement 41
3.4.1.5 Modalités d'enregistrement 42
Information aux enquêtés et demande d'autorisation (consentement éclairé) 43
3.4.1.6 Définition du "consentement éclairé" 44
3.4.1.7 Moment de l'information et de la demande 44
3.4.1.8 Qui énonce l'information, la demande et qui y répond 45
3.4.1.9 Qu'est-ce qu' "informer"? 45
3.4.1.10 L'objet de la demande d'autorisation 46
3.4.1.11 Les formes de l'autorisation 47
Prévoir l'après de l'enquête : retours, debriefings 48

3.5 Anonymisation 50
Définition 50
Données concernées par l'anonymisation 51
Moments auxquelles peut intervenir l'anonymisation 51
Modes d'anonymisation 52
3.5.1.1 Formes ou éléments des données pouvant être concernés par l'anonymisation .. 52
3.5.1.2 Formes de remplacement 53
Les limites de l'anonymisation 54
3.5.1.3 Limitations issues des contextes de production et de circulation des données ... 55
3.5.1.4 Limitations issues du travail d'analyse 56

3.6 Transcription 56
Introduction de la transcription : description ethnographique 56
L'identification des locuteurs dans la transcription 57
Enjeux liés aux choix effectués dans le corps du texte 58
3.6.1.1 Enjeux (ortho)graphiques 58
3.6.1.2 La représentation du parler exolingue 58
3.6.1.3 Enjeux du multimodal et du détail de la transcription 59

4 Les corpus oraux, objets de patrimoine ? 61

4.1 Rappel de la situation des corpus oraux produits par des chercheurs au sein des institutions patrimoniales 61
L'oral en forme, l'oral en mots 62
Mais les collections de corpus oraux constituent-elles, pour autant, une catégorie du Patrimoine ? 62

4.2 La politique de l'Etat en matière de collecte et de conservation 63
Quelle place, les textes de lois qui régissent les différentes institutions patrimoniales réservent-ils aux collections de corpus oraux ? 63
4.2.1.1 Les textes de la Bibliothèque nationale de France. Pratiques et usages. 63
4.2.1.2 Les textes des Archives. Pratiques et usages. 66
4.2.1.3 Les textes des Musées de France. Pratiques et usages 69
4.2.1.4 Les textes de l'INA / Inathèque site. Pratiques et usages 70
Les collections de corpus oraux. Pratiques et usages des institutions patrimoniales. 73
Les initiatives privées 75
L'accès aux collections 76

4.3 Vers la reconnaissance d'un statut du patrimoine oral 78

5 Conclusion provisoire 79

6 Annexes 80

6.1 Fiches juridiques ..... - 81 -
Données personnelles et anonymisation ..... - 81 -
LE DROIT DE CITATION ..... - 84 -
CONSENTEMENT ..... - 85 -
LES BASES DE DONNÉES ..... - 87 -
RESPONSABLE DU TRAITEMENT ..... - 89 -
Le Patrimoine immatériel et l'UNESCO ..... - 91 -
Présentation de quelques lois africaines sur la protection du patrimoine culturel ..... - 93 -

6.2 Fiches techniques ..... - 95 -
Eléments de méthode pour la prise de son et l'enregistrement de terrain ..... - 95 -
Supports d'enregistrement, supports d'archivage : Le son ..... - 100 -
Supports d'enregistrement, supports d'archivage : La vidéo ..... - 103 -

6.3 Fiches formats et normes ..... - 106 -
codages et formats pour les ressources enregistrées, leurs annotations linguistiques et documentaires ..... - 106 -

6.4 Bibliographie ..... - 110 -
Corpus ..... - 112 -
Aides à la transcription ..... - 112 -

6.5 Glossaire ..... - 115 -

6.6 Index ..... - 118 -

6.7 Table des matières ..... - 120 -

Imprimerie Ministère de la Culture et de la Communication
Mai 2005