Guide de production de ressources linguistiques : analyse des aspects techniques, juridiques et stratégiques

Franck Gandcher, Khalid Choukri, Olivier Hamon, Valérie Mapelli,
Nicolas Moreau, Niklas Paulsson, Djamel Mostefa

Référence :	ELDA/2008/GPROD
Titre :	Guide de production de ressources linguistiques : analyse des aspects techniques, juridiques et stratégiques
Auteurs :	Franck Gandcher, Khalid Choukri, Olivier Hamon, Valérie Mapelli, Nicolas Moreau, Niklas Paulsson, Djamel Mostefa
Date de publication :	15 septembre 2008
Format :	Version 4.0, 101 pages
Diffusion :	Rapport interne ELDA
Langue :	Français
Copyright :	© 2008. ELDA, 55-57, rue Brillat-Savarin, 75013 Paris, France. Tous droits réservés.

TABLE DES MATIERES

1. Introduction

1.1 Contexte

Les logiciels de veille informationnelle utilisent de plus en plus des ressources linguistiques sophistiquées demandant des efforts importants de développement. Ces ressources sont également de plus en plus multilingues car ni le français ni l'anglais ne sont plus suffisants pour exercer une veille sur Internet dans les domaines de l'information scientifique et technique.

Des ressources linguistiques de grande qualité, tant dans le domaine de la langue écrite que de la langue orale, sont produites en Europe aussi bien par les centres de recherche académique que par les industriels. De nombreux freins demeurent néanmoins avant de permettre la distribution de ces ressources et leur réutilisation par des tiers. Il est nécessaire de prévenir ces freins et de les prendre en compte à tous les niveaux de la mise à disposition d'une ressource (production, identification, distribution).

Ces freins sont de différents types :

- Techniques :
    - le format de codage et de stockage de données (par exemple SAM, Sphere, WAV, etc.) et les outils de conversion (UTF-8, ASCII, etc.),
    - l'utilisation de formats de description (metadata) de ressources incompatibles ou non conformes à l'état de l'art (descripteurs propriétaires non standards et/ou non normalisés), servant par exemple pour le catalogage et la recherche des ressources disponibles sur différents sites (IMDI, OLAC, catalogue ELDA, catalogue LDC, etc.),
    - l'utilisation de formats de production de ressources incompatibles ou non conformes à l'état de l'art (formats propriétaires non standards et/ou non normalisés).

- Juridiques :
    - l'absence de préoccupation juridique dans les centres académiques qui « oublient » de demander les autorisations préalables,
    - le recours à des modèles juridiques de diffusion fermée,
    - la non prise en compte des différentes « strates » de propriété intellectuelle (par exemple la production de ressources nouvelles intégrant des ressources déjà couvertes par la propriété intellectuelle),
    - multiplicité de modèles de contrats « maisons », souvent inspirés des modèles de distribution de logiciel (de type GNU, GPL, Creative Commons), généralement peu adaptés aux ressources linguistiques,
    - la diversité des modes de protection juridique en Europe et dans le monde, etc.

- Stratégiques :
    - le coût d'adaptation ou d'acquisition qui peut ne pas correspondre à un « marché » ou aux capacités d'utilisateurs potentiels,
    - l'inexistence d'une ressource spécifique correspondant à un besoin donné,
    - l'indisponibilité d'une ressource spécifique existante que le propriétaire/producteur ne souhaite pas mettre sur le marché pour des raisons de concurrence (technologique, stratégique, économique).

L'Agence pour l'évaluation et la distribution de ressources linguistiques, ELDA, considère ce type de questions depuis sa création en 1995 comme des questions primordiales faisant partie de ses missions originelles que sont l'identification, la collecte, la production, la validation et la distribution de ressources linguistiques. Par exemple, en ce qui concerne les questions juridiques, ELDA a plusieurs fois travaillé avec des juristes pour définir des contrats types et spécifiques à la distribution de ressources linguistiques. Dans ce contexte, nous avons produit différents types de licences servant à résoudre, d'une part, les relations entre le propriétaire des données et le distributeur, et, d'autre part, les relations entre le distributeur et l'utilisateur des données. Cela a permis à ELDA de consolider d'une façon stable et contrôlée l'échange et le partage des données, ce qui s'est concrétisé par la conclusion de plusieurs milliers de licences entre ELDA et les propriétaire/utilisateurs de ressources.

1.2 Objectifs

L'objectif du projet est de réaliser un guide indiquant les recommandations techniques, juridiques et stratégiques adressées aux producteurs de ressources linguistiques et permettant d'assurer la réutilisation des ressources.

Pour les différentes ressources écrites et orales sont décrits les formats, standards et bonnes pratiques correspondant à l'état de l'art. Sont également décrits les différents modèles juridiques susceptibles d'être utilisés pour une diffusion de ces ressources de la manière la plus large possible.

D'un point de vue stratégique, l'objectif est d'encourager la collaboration et le partage des ressources entre les différents acteurs du domaine, en proposant un plan de dissémination prenant en compte les différentes contraintes de ces acteurs.

Ce guide s'adresse notamment aux laboratoires académiques engagés dans la production de ressources afin de les aider dans leurs démarches, mais aussi pour tout acteur des technologies de la langue souhaitant suivre les meilleures pratiques du domaine. Les recommandations de ce guide porteront sur les différentes étapes de la vie d'une ressource linguistique que sont :

- les spécifications,
- la production,
- la validation,
- la distribution,
- la maintenance,

et cela selon les deux dimensions majeures mentionnées ci-dessus : technique et juridique.
Quelques aspects de tarification seront aussi abordés.

La diffusion de ce guide se fera via le site Technolangue.net (www.technolangue.net). Ce portail internet, maintenu par ELDA, est consacré au domaine des technologies du traitement de la langue écrite et parlée. Il est destiné à valoriser le secteur, à informer sur ses acteurs, à rendre compréhensibles ses enjeux et à alerter sur les principales évolutions scientifiques, technologiques, industrielles et normatives.

La mise en ligne du guide s'effectuera après la recette du guide par le ministère (Bureau de l'information scientifique et technique).

2. Aspects techniques

2.1 Standards et bonnes pratiques pour la production de ressources linguistiques

Au cours de ses travaux d'inventaire de ressources linguistiques, ELDA a pu mettre en évidence plusieurs types de « Ressources Linguistiques ». Par souci de clarté, ce rapport se concentrera sur quatre types de ressources linguistiques : les corpus écrits, les lexiques, les corpus de parole et les ressources multimodales/multimédia.

Ainsi, pour chaque type de ressource, nous présenterons quatre critères de première importance pour la production de ressources linguistiques :

- Les spécifications
- Les formats de codage
- Les formats de stockage
- Les standards de validation

Un historique sera également proposé pour chaque type de ressources.

2.1.1 Corpus

2.1.1.1 Historique

Parmi les corpus précurseurs, le corpus Brown[1] (*The Brown Corpus of Standard American English*) est le premier corpus informatisé à avoir été compilé en 1967 par la *Brown University* de Providence, aux Etats-Unis. La taille du corpus est de 1 million de mots, provenant de textes écrits en 1961. Encore utilisé de nos jours, il est composé de 500 documents de 2 000 mots chacun, repartis en 15 catégories (presse, religion, humour, etc.).
Peu de temps après, en 1968, un nouveau corpus de 1 million de mots (200 documents de 5 000 mots chacun) est constitué par la *University College of London*. Son contenu est représentatif de l'anglais britannique écrit, mais également parlé, puisqu'il contient autant des documents écrits que des transcriptions de documents parlés.
Le corpus Brown a ainsi inspiré bon nombre de corpus, tels que les corpus LOB, Kolhapur, SSE (*Survey of Spoken English*), Wellington, Australian, etc. Ces corpus concernent chaque fois une langue anglaise régionale (Nouvelle-Zélande, Australie, Inde, etc.) et sont basés sur le même modèle de structure.

De la même manière, de nombreux et larges corpus « nationaux », auxquels s'ajoutent la plupart du temps des annotations, ont vu le jour à partir de la fin des années 1980, à commencer par le *British National Corpus* (100 millions de mots), puis le *American National Corpus* (100 millions de mots). Sur cette base, on peut citer entre autres les *Croatian National corpus*, *Czech National Corpus*, *Hellenic National Corpus*, *Hungarian National corpus*, etc.

L'utilisation de corpus pour la recherche ou des applications commerciales apporte de nombreuses possibilités pour le traitement automatique des langues. Lors de développement d'applications, l'approche à base de corpus permet de profondes améliorations, comme notamment en étiquetage morphosyntaxique ou en traduction automatique. Cela nécessite l'annotation des corpus, souvent coûteuse mais nécessaire. Il est utile ici de citer les travaux de (Francis & Kucera, 1982) qui ont été les premiers a étiqueter un corpus, le corpus Brown, et à le rendre disponible à l'ensemble de la communauté, permettant ainsi des progrès conséquents en recherche.

Les corpus sont également utilisés avec succès lors de campagnes d'évaluation, comme la plupart d'entres elles peuvent le démontrer (MUC, TREC, CLEF, campagnes EVALDA, etc.). Dans un même temps, ces campagnes permettent la production de nouvelles ressources, favorisant la recherche et le développement d'applications, lorsque ces ressources sont réutilisées.

En 1992, le corpus SUSANNE[2] (*Surface and Underlying Structural ANalysis of Natural English*) est développé par la *University of Sussex* (Sampson, 1995) et a pour but de représenter une taxinomie des unités grammaticales sur l'anglais. Le corpus est étiqueté syntaxiquement et est un sous-ensemble du corpus Brown. Il contient 64 fichiers de 2 000 mots annotés syntaxiquement.

[1] http://www.ldc.upenn.edu/cgi-bin/ldc/textcorpus?doc=yes&corpus=BROWN

[2] http://www.csc.fi/english/research/software/susanne

En 1993, EAGLES[1] (*Expert Advisory Group on Language Engineering Standards*) définit des normes pour les corpus textuels à propos de la typologie, l'encodage, l'annotation (syntaxique et morphosyntaxique) et autres recommandations sur ce type de corpus. Un des résultats concrets se trouve dans la définition du standard d'encodage de corpus (CES[2] – *Corpus Encoding Standard*), étant une application SGML des standards de la TEI[3] (Text Encoding Initiative). Ces standards se sont développés et ont été maintenus depuis 1994, ils ont pour objectif de définir un guide pour l'encodage et la représentation des textes sous forme digitale.

En 1994-1995, le projet européen MLAP-93/20 a développé et aligné un corpus parallèle (c'est-à-dire contenant des traductions réciproques) anglais, français et espagnol (CRATER[4] – *Corpus Resources and Terminology Extraction*), devenant le premier corpus de ce type a être rendu disponible à l'ensemble de la communauté. Divers outils ont également été développés pour l'alignement de corpus, la navigation dans ce corpus trilingue, ainsi que l'extraction de terminologie.

Associé au projet EAGLES, le projet Multext (*Multilingual Text Tools and Corpora*, 1996) définit lui aussi des standards et spécifications, entre autres pour l'encodage et le traitement de corpus textuels. Basé sur un plan plus pratique que celui d'EAGLES, Multext a travaillé sur deux corpus (JOC – Journal Officiel de la communauté Européenne – et Dagens Industri 1993) totalisant six langues (anglais, allemand, espagnol, français et italien pour le JOC, suédois pour le Dagens Industri 1993).

Le projet PAROLE[5] (1996-1998) avait pour but de créer un ensemble de corpus pour toutes les langues de l'Union Européenne de l'époque (français belge, catalan, hollandais, anglais, français, finnois, allemand, grec, irlandais, italien, norvégien, portugais et suédois) et d'harmoniser ces corpus d'après certains critères : les textes ne devaient pas être plus vieux que 1970 et contenir dans une certaine proportion les mêmes types de textes (livres, journaux, périodiques, divers). Cet effort d'harmonisation était également appliqué à l'encodage textuel et linguistique des corpus. Chacun des corpus était ainsi codé en respectant à la fois la DTD PAROLE (qui est dérivée de la DTD des guides TEI) et celle de la CES. De plus, environ 250 000 mots étaient étiquetés morphosyntaxiquement, d'après un guide d'annotation défini au cours du projet. Par la suite, en 1999, le projet SIMPLE[6] a permis d'ajouter des informations sémantiques sur les corpus PAROLE.

A partir de 1987 et ce jusqu'à nos jours, le corpus *Le Monde* a été développé progressivement en compilant l'ensemble des articles du journal du même nom. Cela représente environ 50 000 articles par an, constituant la plus grosse base de données de la presse française et très certainement un des corpus les plus utilisés en traitement automatique de la langue française.

Si le développement de corpus monolingues, souvent représentatifs de la grammaire d'une langue donnée, a été longtemps privilégié, les corpus parallèles sont a priori de plus en plus demandés et a fortiori développés en conséquence. Pour aller plus loin, le besoin de plus en plus pressant de grandes quantités de données amène les développeurs à utiliser des corpus comparables (Déjean & Gaussier, 2002), composés de documents écrits dans deux langues différentes et ayant un vocabulaire commun, sans pour autant être un assemblage de documents parallèles.

Ainsi, un large corpus multilingue et parallèle a été développé dans le projet C-STAR[7] (*Consortium for Speech Translation Advanced research*), démarré en 1991, puis poursuivi en de 1993 à 1999 avec le projet C-STAR II.

De même, en 2001 le corpus Hansard[8] a été développé lors du projet *Rewrite*[9] et contenant 1,3 million de paires de phrases alignées provenant d'enregistrement du 36e Parlement canadien. Le format du corpus choisi est très simple, puisqu'il s'agit de fichiers contenant une phrase par ligne.

1 http://www.ilc.cnr.it/EAGLES/home.html

2 http://www.cs.vassar.edu/CES

3 http://www.tei-c.org

4 http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

5 http://www.elda.fr/catalogue/en/text/doc/parole.html

6 http://www.ub.es/gilcub/SIMPLE/simple.html

7 http://www.c-star.org

8 http://www.isi.edu/natural-language/download/hansard

9 http://www.isi.edu/natural-language/projects/rewrite/index.html

Démarré en 1996, et se poursuivant encore actuellement, le corpus Europarl[1] (*European Parliament Proceedings Parallel Corpus*) inclut 11 langues extraites d'enregistrements du Parlement Européen. Le corpus multilingue contient à l'heure actuelle 44 millions de mots par langue. Ce corpus fait partie d'un plus important ensemble de corpus (OPUS[2]) développé sous licence Open Source, regroupant différents corpus multilingues alignés de domaines techniques (Open Office, opensubtitles.org, messages du système KDE, manuel KDE, manuel PHP) ou institutionnels (Parlement Européen, constitution européenne, agence médicale européenne).

Le JRC-Acquis[3], actuellement dans sa version 3.0, est l'un des corpus parallèles les plus importants, créé à partir de plus de 20 langues européennes incluant, de manière quasi unique, des combinaisons rares (telle que la paire Estonien-Grec). Les documents bilingues alignés respectent le format XML des standards de la TEI, et chacun d'entre eux est classé selon des domaines bien spécifiques. Le nombre de mots varie entre 20 et 60 millions par langue présente.

Dorénavant, les très grands corpus sont de plus en plus employés, notamment dans le cadre d'approches statistiques, demandeuses de grosses masses de données. La production de telles ressources se fait notamment en utilisant des données provenant d'Internet, que ce soit pour des données monolingues ou bilingues. On parle désormais de corpus TERABYTE, comme les corpus WT10g, GOV ou TERABYTE utilisés lors des campagnes TREC[4] (tâches *terabyte*).

2.1.1.2 Spécifications

Il est possible de classifier les corpus textuels dans de nombreuses catégories, les critères de définition d'un corpus pouvant se soustraire à :

- ses objectifs (entraînement, développement, évaluation, analyse, etc.),
- son contexte (domaine, langue, tâche, etc.),
- sa structure (monolingue, multilingue, parallèle, aligné, complet, etc.),
- son format (brut, annoté, etc.) ou encore son origine (Web, transcriptions, journalistique, etc.).

Par ailleurs, les corpus généraux sont représentatifs du langage de même que leur grammaire, vocabulaire, construction sémantique, etc. Par contre, les corpus spécialisés (par exemple : médical ou juridique) sont représentatifs d'un domaine spécifique, sans nécessairement l'être pour le langage correspondant ; le vocabulaire ou la grammaire auront du mal à correspondre lors de l'application à un autre domaine.

Les spécifications d'un corpus passent par une étude préalable de l'utilisation qu'il en sera faite et répondant aux critères de définition fournis plus haut. Cela signifie que les données constituant le corpus sont acquises de manière à se rapprocher le plus possible des données réelles, et à répondre au cadre de la tâche d'application du corpus, et ce de la meilleure manière possible. Par exemple, si l'on considère un corpus utilisé en recherche d'information pour des systèmes de question-réponse sur des documents journalistiques, le choix peut se porter sur la correction en amont de l'orthographe des documents, ou au contraire laisser le système automatique s'adapter à des conditions plus réalistes. Quelle que ce soit l'utilisation finale du corpus constitué, ce genre de spécification doit être convenu au préalable, avant toute production de corpus.

En 1996, un des groupes de travail du projet EAGLES (TCWG – Text Corpora Working Group) a pour but d'établir des recommandations sur la typologie des corpus textuels. Certaines caractéristiques attribuables à un corpus ont été établies, avec des valeurs par défaut (le corpus devenant *spécialisé* lorsque ces valeurs par défaut sont modifiées), concernant :

- Sa taille : par défaut, elle est *large*, en terme de mots. Elle doit par ailleurs refléter la difficulté à établir ce corpus. D'une manière plus pragmatique, la taille d'un corpus textuel utilisé pour un traitement automatique va dépendre du type d'utilisation du corpus qu'il en est fait, mais aussi du domaine et de la tâche applicatifs, et des objectifs souhaitant être atteints. Par exemple, un corpus spécialisé de petite taille sera généralement plus pertinent qu'un corpus général volumineux.
- Sa qualité : par défaut, elle est *authentique*. Le langage employé est celui d'une communication ordinaire, sans aucune circonstance artificielle. La frontière est complexe à élaborer, toujours est-il qu'une quelconque intervention linguistique doit être notifiée. De plus, un corpus imparfait ne doit pas être négligé, dès lors que ses imperfections sont connues et prises en compte lors de l'interprétation des résultats.

[1] http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl

[2] http://urd.let.rug.nl/tiedeman/OPUS

[3] http://langtech.jrc.it/JRC-Acquis.html

[4] http://trec.nist.gov

- Sa simplicité : par défaut, il s'agit d'un *texte brut*. Concrètement, l'emploi de chaînes de caractères ASCII est considéré comme valeur par défaut. Il est certain que le critère de *simplicité* a évolué depuis EAGLES, et s'est « complexifiée » par la même occasion. Il est dorénavant plus coutume d'employer des format balisés (SGML, XML, etc.) et l'encodage de caractères UTF-8, permettant l'emploi de caractères non latins (voir partie 2.1.1.3).
- Sa documentation : par défaut, le corpus est *documenté*. Cela concerne les détails sur les composants d'un corpus. Cette documentation est, de manière représentative, fournie par la DTD des formats SGML ou XML, procurant de manière séparée du corpus les informations permettant de comprendre sa structure. Entre autres, cela permet dans certains cas de séparer les données brutes des annotations de ces données.

Dans le cadre de travaux d'évaluation de technologies, il est possible de distinguer de façon simplifiée deux grandes catégories de corpus. Certains corpus sont utilisés pour l'entraînement de systèmes et d'autres pour tester leurs fonctionnalités post-développement dans un cadre d'évaluation clairement défini. La fonction principale d'un corpus d'entraînement est de présenter le maximum de possibilités linguistiques (liées au domaine applicatif) afin que le système puisse s'adapter à un maximum de cas de figure présentés lors de son utilisation. Il est généralement nécessaire d'avoir un corpus d'entraînement d'une taille très importante, afin d'obtenir un maximum de cas pratiques. Par ailleurs, la plupart des corpus sont utilisés dans un but d'apprentissage à partir de modèles probabilistes ou statistiques, afin de développer des systèmes. A partir de calculs plus ou moins complexes, des probabilités sont déduites et permettent de reproduire des modèles de langues. Ainsi, en théorie, plus la taille des corpus utilisés sera importante, plus les résultats obtenus seront de meilleure qualité, dès lors que le degré de précision des probabilités s'affine. *A contrario*, la taille minimale des corpus est difficile à déterminer, puisqu'elle dépend le plus souvent du domaine d'application et de sa complexité.

Les spécifications tiennent également compte du processus de création d'un corpus :
- Collecte des données, incluant l'acquisition des droits (*copyright*) sur ces données et le choix de leur origine (Internet, transcription de parole, etc.),
- Sélection des données pertinentes, d'après les objectifs de la tâche,
- Possible nettoyage et formatage (simplicité, encodage, structure) du corpus,
- Validation du corpus formaté,
- Annotation du corpus, le cas échéant,
- Validation des annotations, le cas échéant,
- Rédaction d'une documentation complète sur le contenu et la structure du corpus.

Ce n'est qu'après avoir appliqué toutes ces étapes successives que le corpus peut-être rendu utilisable.

2.1.1.3 Formats de codage

Longtemps considéré comme le standard en la matière, l'encodage de caractères ASCII n'est plus vraiment utilisé à l'heure actuelle. Il a tout d'abord été abandonné par l'ISO (International Organization for Standardization) qui a défini les standards ISO-8859-1 à ISO-8859-10, afin d'inclure des caractères indispensables comme les caractères accentués des principales langues européennes. Seulement, le développement de l'informatique et des communications a créé un nouveau besoin de représentation des caractères d'autres langues telles que les langues asiatiques, ce que ne permettaient pas les standards ISO. Pour palier à ce problème, le format Unicode a été spécifié, rendant possible l'emploi de nombreux caractères, mais pêchant encore par la taille accrue des documents. Ainsi, l'encodage UTF-8[1] a été spécifié, qui permet de réduire les défauts du format Unicode. De part ses nombreux avantages (encodage d'immenses quantités de caractères, taille, correspondance exacte avec l'encodage ASCII pour les Etats-Unis, etc.), l'encodage UTF-8 est finalement passé dans les mœurs, étant devenu le standard de l'Internet (l'UTF-8 est utilisé par défaut pour les documents XML), et donc de l'informatique.
En conséquence, mais aussi par facilité (l'emploi de corpus multilingues ou la production de ressources pour des langues non européennes), l'UTF-8 est l'encodage le plus utilisé lors de la construction de corpus textuels.

Le formatage des documents textuels constituants les corpus peut se faire de trois manières différentes :
- Format *brut* du texte, sans aucun élément autre que le contenu informatif,
- Format *balisé* du texte, permettant d'enrichir le corpus d'information complémentaires à l'aide d'un système de balisage (SGML, XML, XHTML, etc.),
- Format *compilé* du texte, lisible la plupart du temps par un type de logiciel précis (formats Word, RTF, PDF, ODT, etc.).

[1] Créé par Ken Thompson en 1992

Le second format est le plus couramment utilisé, car il permet l'annotation de corpus, permettant le rajout d'annotations afin d'enrichir le contenu d'informations complémentaires (ou meta informations). Il est alors possible de distinguer à tout moment le texte brut de ses annotations. Différents types d'annotations peuvent ainsi être rajoutés au corpus brut initial, tels que :

- Des informations documentaires : langue employée, identification de segments, méthode, taille, etc.
- Des informations structurales : titre, paragraphe, phrase, mot, etc.
- Des informations linguistiques : syntaxe, sémantique, etc.

Par exemple en annotation syntaxique, la construction d'un corpus annoté de référence permet d'illustrer l'intérêt d'une telle démarche : ce corpus est constitué d'énoncés clairement délimités par des balises, permettant lors d'une évaluation de le comparer aux résultats des systèmes. En effet chaque corpus contiendra une annotation manuelle des énoncés quant au traits syntaxiques des mots les constituants..

Afin d'incorporer de manière simple ces annotations, plusieurs systèmes ont été employés, essentiellement le SGML (Standardized Generalized Markup Language) développé en 1960, et le XML (eXtended Markup Language) développé en 1998. D'une structure plus simple, c'est ce dernier qui s'est imposé pour la création de corpus. Le XML est un langage de balisage générique pour lequel un document balisé est attaché à un schéma (la DTD – *Document Type Definition*). Chaque document attaché à un même schéma doit respecter le jeu de balises que ce dernier offre.

Dans un premier temps, la norme TEI (Text Encoding Initiative)[1], application du format SGML, définit une DTD à même de traiter les documents textuels dans leur ensemble tout en prenant exemple sur de nombreux types de textes (prose, dictionnaire, théâtre, poésie, etc.) et plusieurs domaines (publication électronique, analyse littéraire, recherche documentaire, etc.). La DTD ainsi conçue pour la TEI, modulaire, permet notamment de définir de nouvelles balises ou de redéfinir celles existantes. Cette possible personnalisation de la DTD est alors considérée comme une application de la TEI, c'est-à-dire qu'elle est appliquée dans un cadre bien spécifique. C'est le cas pour le C*orpus Encoding Standard* (CES) qui définit un standard d'encodage de corpus dans le cadre du projet MULTEXT[2], associé à EAGLES et la collaboration de VASSAR/CNRS[3]. A partir de la DTD de la TEI (Ide & Véronis, 1996), le CES fournit ainsi un ensemble de DTD spécifiques au codage des corpus textuels, accompagnées de recommandations quant à leur usage. Le standard a par la suite été redéfini en XML : XCES[4].

Le CES distingue les données brutes des données annotées (et contenant des informations linguistiques uniquement), qui doivent être séparées : les deux types de données sont placées dans des documents bien distincts, le lien entre données et données annotées se faisant *via* des liens hypertextes. Par ailleurs, trois DTD sont définies, issues des DTD de la TEI : CesDoc (encodage des données brutes), CesAna (encodage des données annotées) et CesAlign (alignement de textes parallèles multilingues). Finalement, le CES définit trois niveaux d'encodage des documents, d'une structure grossière (jusqu'au niveau du paragraphe) à une structure plus fine (au-delà des éléments internes au paragraphe) ; ces trois niveaux d'encodage ont un coût croissant et une automatisation décroissante.

En dehors de MULTEXT, le CES a été employé dans le projet PAROLE[5] qui a approfondi la définition des modèles sur deux types de données spécifiques : les unités morphologiques et les unités syntaxiques. Une DTD est proposée pour permettre l'encodage des données brutes analysées et annotées.

Si les principes (XML, DTD, niveau d'encodage, etc.) du CES sont communément admis, il faut reconnaître que la séparation des données (brutes et annotées) ne se fait que très rarement : la plupart du temps les données sont regroupées dans un même document, tandis qu'elles sont attachées à une même DTD commune.

2.1.1.4 Formats de stockage

Les formats de stockage sont de deux types : le stockage du contenu et le stockage physique.

Le format de stockage du contenu dépend des caractéristiques mêmes du corpus :

[1] http://www.tei-c.org

[2] http://aune.lpl.univ-aix.fr/projects/multext

[3] http://www.cs.vassar.edu/~ide/research/

[4] http://www.xces.org

[5] http://www.ub.es/gilcub/SIMPLE/simple.html

- Taille (un corpus très volumineux va être divisé en plusieurs fichiers pour plus de commodité lors de son utilisation). Les corpus en recherche d'information sont souvent divisés de manière à n'avoir qu'une centaine de documents par fichier, afin de ne pas avoir de fichiers trop volumineux, mais aussi de ne pas avoir un trop grand nombre de fichiers (l'un comme l'autre des cas pouvant poser des problèmes pour les outils).
- Homogénéité : on pourra placer les documents dans des répertoires ou des fichiers différents selon qu'il contiennent des parties bien distinctes ou non. Par exemple, la structure d'un corpus bilingue anglais-français peut être telle que les documents anglais sont placés dans un répertoire « EN » et les documents français dans un répertoire « FR ».
- Structure : le corpus peut être constitué d'un document par fichier ou de plusieurs documents par fichiers, le plus souvent regroupés selon le contexte ou la thématique. Il est très facile d'imaginer un corpus annoté, composé de données provenant de domaines variés (courriers électroniques, journaux, discours, etc.) : il devient alors judicieux de décomposer le corpus en plusieurs sous corpus, dans plusieurs fichiers distincts, afin d'observer ou de paramétrer le système en fonction des besoins. Au contraire, dans d'autres cas d'utilisation (une évaluation en contexte par exemple) il pourra être plus judicieux de ne pas distinguer ces différents corpus.
- Etc.

La documentation du corpus aura tendance à être placée à part de manière bien distincte. Elle doit être facilement accessible.

D'une manière générale, les corpus textuels « bruts » sont placés dans des fichiers « textes » (avec ou sans extension *txt*), les corpus avec encodage balisé dans des fichiers avec l'extension XML, SGML, etc. Un corpus textuel peut également être composé de fichiers compilés non reformaté, de type Word ou PDF, dans des cas d'utilisation en contexte réel.

La plupart du temps, le format de stockage physique dépend d'une part du volume des documents composant le corpus, et d'autre part de ces caractéristiques, selon que le corpus est homogène ou contient des parties bien distinctes. Pour de petits corpus, les données peuvent être déposées sur un serveur FTP pour être directement téléchargées *via* le réseau. Ce stockage peut également se faire sur CD-ROM, ou DVD-ROM si le corpus est plus volumineux. L'emploi de disque dur est l'ultime recourt lors de corpus extrêmement volumineux.

2.1.1.5 Standards de validation

La qualité d'un corpus est un élément essentiel lors de sa production, puisqu'il sera utilisé par la suite et que les résultats obtenus par des outils ou logiciels vont dépendre de son état. La question de la qualité se pose aussi lorsque des informations sont perdues ou altérées, ou encore rajoutées. Elle passe aussi par la qualité de l'encodage des données.

La validation des corpus textuels s'établit autour de trois axes majeurs : la validation de la documentation, puis une validation formelle, et enfin la validation du contenu linguistique. De tels formalismes de validation ont été rassemblés, développés ou étendus par le comité de validation d'ELRA. La sous-section « Standards » des pages du site ELRA[1] dédiées à la validation présente notamment les standards à suivre pour les ressources écrites (McEnery et al. 1998). Mis à part quelques exceptions au niveau du format, les validations du contenu se font manuellement, afin d'éviter tout risque d'erreur.

La *validation de la documentation* vérifie que les documents fournis parallèlement à la ressource sont appropriés, de même que le contenu correspondant aux spécifications de sa production. Les critères principaux sont que :
- La documentation doit être claire et suffisante pour pouvoir utiliser la ressource,
- La documentation doit être (au minimum) en anglais,
- La documentation minimale doit contenir la description de sujets de types : copyright, personne à contacter, formats des fichiers et encodage des caractères (comprenant les conventions pour nommer ces fichiers), taille de la ressource, langues utilisées, domaine, applications visées, ainsi que les informations caractéristiques d'un domaine (la documentation d'un corpus pour la traduction automatique ne sera pas nécessairement la même que pour de la recherche d'information).

La *validation formelle* vérifie la facilité d'utilisation de la ressource, notamment par le format et la structure de celle-ci. Elle comprend :
- L'accès à toutes les données (les fichiers peuvent-ils s'ouvrir ?),
- L'absence de virus informatiques,

[1] http://www.elra.info

- Le format des fichiers et l'encodage des caractères doivent correspondre à la documentation (c'est-à-dire aux spécifications),
- La taille du corpus correspond à la documentation,
- Les structures sont correctes,
- L'ensemble ne doit contenir aucun manquement aux données,
- Les données sont valides d'un point de vue légal.

La *validation du contenu* mesure la fiabilité des données brutes linguistiques. Cette validation se fait la plupart du temps sur un échantillon le plus représentatif possible de la ressource. Elle dépend souvent du type de ressource, mais peut être sommairement représentée par :
- Le balisage et la structure des données,
- L'information morphologique,
- L'information syntaxique,
- L'information sémantique.

Lorsque toutes les validations sont terminées, un rapport de validation doit être rédigé, et inclus à la ressource produite.

2.1.2 Ressources orales

2.1.2.1 Historique

- ***Europe***

Une de premières initiatives dans la production de ressource orales en Europe fut le projet Esprit SAM entre 1987 et 1993. Le consortium comportait trente partenaires de huit pays différents. Le travail effectué dans le projet SAM était divisé en trois domaines :
- Evaluation de la reconnaissance vocale
- Evaluation de la synthèse vocale
- Technologies et recherche

Dans ce but, le projet SAM a développé une plate-forme d'enregistrement appelée SESAM et contenant un PC avec une carte d'acquisition OROS et une suite de logiciels SAM, aussi développés dans le projet.
Cette plate-forme permettait aux laboratoires de recherche en Europe de produire des ressources orales dans les mêmes conditions et avec les mêmes standards. Le premier corpus produit dans le cadre du projet SAM, EUROM-0, contient plus de cinq heures de parole.

Le laboratoire LIMSI-CNRS a enregistré en 1991 une base de données BREF de 120 locuteurs répartie en deux groupes : 80 locuteurs prononçant 10 000 mots en parole continue et 40 locuteurs prononçant 5 000 mots. Afin de garantir une distribution phonémique optimale, les textes sont extraits du journal Le Monde à partir d'un corpus de 4 millions de mots. Le contenu de *BREF* est de même nature que celui du *Wall Street Journal* (*WSJ*) enregistré postérieurement aux Etats-Unis. La base de données a été enregistrée avec deux microphones : un premier à proximité (closetalk) et un deuxième à mi-distance posé sur une table (table top). Elle servira à entraîner des systèmes pour des tâches de dictée automatique de textes ainsi que l'évaluation de ces derniers.

En 1993, le projet Polyphone a enregistré via le téléphone plus de 10 000 locuteurs dans plus des dix langues. A chaque locuteur a été transmise une feuille contenant 38 « items » à prononcer (séquences de chiffres, nombres, noms épelés, mots applicatifs, phrases au contenu phonétique contrôlé, questions demandant une réponse spontanée). Polyphone visait à mettre au point des vocabulaires fixes pour les serveurs vocaux ainsi que pour couvrir la variabilité inter-locuteur. Une base de données complémentaire, PolyVar, a été enregistrée pour couvrir la variabilité intra-locuteur. PolyVar consiste en 100 appels par locuteur effectués par 50 locuteurs. Elle permet de tester les systèmes adaptables aux locuteurs et les systèmes de vérification du locuteur.

Les normes pour les ressources orales utilisées aujourd'hui en Europe sont basées sur le travail effectué dans le projet EAGLES (1993) – the Expert Advisory Group on Language Engineering Standards[1]. Un groupe de travail, Spoken Language Working Group (WG5), a établi un jeu de standards appelés « Working Standards » pour faciliter la production des ressources orales à grande échelle.

[1] http://www.ilc.cnr.it/EAGLES/home.html

Le premier grand projet à utiliser ces standards fut SpeechDat(M)[1] en 1995 avec 8 bases de données enregistrées par le réseau téléphonique fixe (chacune contenant 1000 locuteurs) et une base de données enregistrée par le réseau téléphonique mobile (300 locuteurs).

SpeechDat(M) est suivi en 1997 par Speechdat(II) avec 25 bases de données enregistrées en Europe, chacune comprenant entre 50 et 5000 locuteurs. Speechdat(II) a permis de mettre au point la norme pour les ressources téléphoniques d'aujourd'hui en Europe, souvent appelée le standard SpeechDat. Ce standard comprend l'utilisation du format SAM pour les fichiers de transcriptions, un lexique avec la phonétique en SAMPA, un jeu d'étiquettes pour les transcriptions ainsi que la structure utilisée pour stocker la ressource.

Le succès de SpeechDat continue en 1998 avec SpeechDat(E) qui comporte 5 bases de données enregistrées dans les langues d'Europe de l'Est suivantes : russe, tchèque, slovaque, polonais et hongrois.

En 1999, le projet SpeechDat-Car a enregistré 9 bases de données en voiture avec 600 locuteurs chacune. Les enregistrements ont été effectués par le réseau téléphonique mobile.

En 2000, le projet Speecon adapte la norme Speechdat pour des ressources enregistrées par microphone, aussi appelés « desktop ». Ce grand projet comprend des enregistrements effectués par microphone dans des environnements différents tels que bureau, parc, rue et voiture. Le projet Speecon comportait initialement 18 langues avec 600 locuteurs par langue et enregistrées depuis 4 microphones. Depuis, six autres bases de données Speecon, enregistrées en Amérique du Nord et en Asie ont vu le jour.

Les normes utilisées dans Speecon ont été adaptées pour des conditions d'enregistrements avec microphone. De nouvelles étiquettes pour marquer les environnements, les types de microphone, l'emplacement de microphone, etc. ont été introduites. De plus, une plate-forme d'enregistrement a été développée pour faciliter les enregistrements et pour favoriser des conditions similaires (même microphone, même outil d'enregistrements, etc.).

Après les projets Speechdat en Europe, c'est au tour de l'Amérique du Sud de faire l'objet de ce type d'enregistrements, dans le projet SALA (Speechdat Across Latin America). Le projet a produit 6 bases de données contenant chacune 600 locuteurs. Un nouveau projet, SALA II continue les enregistrements en 2003 avec 13 bases de données réalisées en Amérique du Sud et en Amérique du Nord. Les deux projets sont effectués via le réseau téléphonique mobile.

En 2002 le projet OrienTel a enregistré des bases de données dans 6 pays autour de la Méditerranée principalement pour la langue arabe. Dans chaque pays, 3 bases de données sont réalisées : en arabe standard, arabe dialectal et dans la langue d'affaires du pays (anglais ou français). En outre, des enregistrements dans deux pays supplémentaires ont permettis d'ajouter 4 bases de données au projet d'origine : arabe levantin, hébreu, grec et turc. Avec le projet Orientel une nouvelle difficulté se pose : le traitement des langues non latines. Pendant le projet, un jeu de caractères phonétiques pour l'arabe a été développé en SAMPA.

En 2005, le projet LILA a permis d'étendre les ressources normalisées d'après SpeechDat à cinq bases de données enregistrées en Asie comprenant entre 1 000 et 2 000 locuteurs chacune. Au cours du projet, l'encodage UTF-8 et des étiquettes pour la romanisation des langues non latines ont été introduits.

Au-delà des enregistrements par téléphone ou microphone décrits ci-dessus, il existe un troisième type de ressource orale : la télé-radio–diffusion, plus communément reconnue sous sa terminologie anglaise « broadcast news ». Une des premières ressources de ce type, l'IBNC – Italian Broadcast News Corpus, a pu établir les normes pour la production des ressources du même type. Des sources provenant de la télé ou de la radio sont enregistrées et transcrites avec des règles de transcriptions très strictes. Ces ressources sont souvent de taille importante, soit des dizaines à des milliers d'heures d'enregistrements. D'autres ressources « broadcast news » ont été produites telles que la base NetDC et le Corpus oral d'actualités radiophoniques NEMLAR.

[1] http://www.speechdat.org/speechdt/speechdat_m/home.html

- ***Etats-Unis***

Outre-Atlantique, DARPA (Defense Advanced Research Projects Agency) lançait en 1984 un programme sur le traitement de la parole et du langage naturel pour mettre en place des campagnes d'évaluations périodiques. Les campagnes, ouvertes aux laboratoires publics et industriels, comprenaient des séminaires pour permettre aux participants de partager leurs connaissances et compétences. DARPA fut un des premiers a s'investir dans le domaine de campagnes d'évaluation. Des méthodes, des protocoles, des ressources et des outils linguistiques ont été mis en place pour mesurer les progrès accomplis dans le domaine. Les résultats démontrent une amélioration continue au fil des années.

Avec la croissance du domaine, il s'est avéré bientôt nécessaire de gérer la collecte et la diffusion des ressources linguistiques. Dans ce but, LDC (Linguistic Data Consortium) fut créé en 1992 pour mettre ces ressources à disposition pour la recherche et développement. Parmi les ressources orales disponibles à LDC on trouve : TIMIT Acoustic-Phonetic Continuous Speech Corpora, ARPA Continuous Speech Recognition Corpora (WSJ etc), Air Travel Information System (ATIS) Corpora, Texas Instruments Speaker-Independent Connected-Digit Corpus (TIDIGITS), Air Traffic Control Corpus (ATC0), SPIDRE Speaker Identification Corpus, YOHO Speaker Verification Corpus, OGI Multi-Language Corpus, OGI Spelled and Spoken Telephone Corpus, WSJCAM0: Cambridge Read News Corpus.

- ***Inde***

Avec le programme TDIL (Technology Development in Indian Languages) le ministère de la communication et de la technologie d'information du gouvernement de l'Inde commençait à produire des ressources linguistiques pour les langues indiennes dans les années 90. Le résultat du premier projet fur la création d'un corpus de 45 millions de mots dans 15 langues différentes. Ces corpus ont servi de base pour les recherches et les développements de la technologie dans des différents domaines tels que OCR, reconnaissance de la parole, serveurs vocaux, etc.
L'objectif de TDIL c'est le développement des outils et des technologies pour faciliter la communication entre homme-machine, la création des ressources multilingues ainsi que l'intégration de ces ressources dans des produits et services. Au fil des années TDIL a développé des ressources et des outils tels que des polices, des éditeurs de textes, des correcteurs d'orthographe, de l'OCR (reconnaissance optique de caractères), de la synthèse de la parole et des standards.

Début 2008, le centre LDC-IL (LDC for Indian Languages) a été établi pour la collecte et la diffusion des ressources linguistiques dans les langues indiennes. Le centre situé à Mysore est géré par le CIIL (Central Institute for Indian Languages). Leur objectif est d'apporter une aide aux chercheurs et producteurs dans le monde entier pour les corpus et technologies de la langue relatifs aux langues indiennes.

Dans le cadre du projet LILA, trois ressources téléphoniques ont été réalisees en Inde : 2 000 locuteurs en hindi en tant que langue maternelle, 1 800 locuteurs en hindi (deuxième langue) et 1 800 locuteurs en anglais. Le contenu est basé sur les projets précédents tel que OrienTel et SALA mais adapté aux langues asiatiques. Deux autres ressources ont été enregistrées dans le même projet : 1 000 locuteurs en Corée du Sud ainsi que 1 800 locuteurs en Chine.

Le gouvernement en Inde a mis en place un programme à long terme pour le développement des technologies de la langue :

- 1976 – 1990 : Phase A. Adaptation des technologies et la création des compétences dans les laboratoires de recherche.
- 1991 – 2000 : Phase B. Développer des outils de base pour le traitement de l'information, les interfaces, et des outils de conversion. Le programme de TDIL a été établi pendant cette phase.
- 2001 – 2010 : Phase C. L'objectif de cette phase est de développer des technologies collaboratives, des corpus et leur analyse ainsi que des outils de gestion de contenu.

Lors de la phase C, TDIL a introduit la Vision 2010 qui comporte des objectifs à court terme, moyen terme et long terme. Les axes majeurs sont les outils pour le traitement des langues, les corpus multilingues et parallèles, la traduction automatique, la reconnaissance et la synthèse de la parole, l'OCR, la localisation et la standardisation. L'objectif à long terme est de développer des systèmes de traduction automatique « Speech to Speech » (parole-parole).

- ***Japon***

Le partage de ressources linguistiques pour le bénéfice de tout le monde a été reconnu depuis longtemps au Japon. Malgré cela le développement de telles ressources a progressé lentement. En 1994, le LRSI (Linguistic Resources Sharing Initiative) a été lancé pour la production, la collecte et la distribution des ressources linguistiques. Ensuite GSK (Gengo Shigen Kyookai, Language Resources Association) a été établi en 1999 mais après un démarrage difficile, un plan de 3 ans pour un soutien financier a été dressé en 2005. Le NII (National Institute of Informatics), à la fois en tant que centre national de l'informatique et en tant que l'un des instituts inter-universitaire de recherche, vise à approfondir le domaine de l'informatique et de construire une infrastructure d'information scientifique. Dans le cadre de la promotion de ces missions, NII a décidé de lancer le SRC (Speech Resources Consortium) qui a pour objectif la création de références en matière de ressources d'information et en particulier les ressources orales. NII oeuvre pour la promotion de ce consortium avec GSK.

- ***Chine***

La recherche sur la synthèse de la parole et la reconnaissance du chinois a commencé au milieu des années 1980. Par la suite, beaucoup de recherche dans le LN a été réalisée dans les années 1990. Dans la même période, des projets pour la création de ressources linguistiques de grande taille ont été lancés en Chine, Taïwan et Hong Kong. En Chine, le programme 863 a créé plusieurs corpus dans divers domaines, tels que des corpus pour la reconnaissance vocale, la synthèse de la parole, les systèmes de traitement de textes parallèles (pour le chinois, l'anglais et le japonais), les systèmes d'indexation et les systèmes de dialogue.

En Chine, il existe plusieurs organismes voués au développement de corpus pour le traitement de la langue en chinois. Parmi eux, ChineseLDC (Chinese Linguistic Data Consortium) est une agence nationale des chercheurs dédiés à la création et le développement des ressources linguistiques chinoises et pour promouvoir les technologies de la parole et de la langue. ChineseLDC a commencé avec un projet soutenu par le Programme 973. L'objectif de ChineseLDC est de mettre en place un catalogue général qui est composé des meilleures ressources linguistiques chinoises et qui sont comparables aux ressources actuellement disponibles sur le marché international. Afin d'atteindre cet objectif, ChineseLDC crée et collecte des données linguistiques telles que des lexiques, des corpus, ainsi que des données pour créer des standards et des critères de production.

Le CCC (Chinese Corpus Consortium) a été fondée en 2004 pour la distribution du corpus en langue chinoise. Le CCC a été parrainé par un groupe d'universités, d'instituts et d'entreprises privées. L'établissement a pour objectif de fournir des corpus pour la reconnaissance de la parole automatique (ASR), la synthèse (TTS) de la parole, du traitement de langage naturel, analyse de la perception, l'analyse phonétique, l'analyse linguistique, et d'autres tâches connexes. Le siège social est à Beijing, en Chine et CCC est soutenu par le département de ressources linguistiques chinoises de HTEA (High-tech Enterprises Association). Sous la direction de la HTEA, CCC travaille pour promouvoir la standardisation et l'industrialisation des ressources linguistiques chinoises.

2.1.2.2 Spécifications

- ***Types de ressources***

Les ressources orales peuvent être divisées en trois catégories :
- Ressources téléphoniques
- Ressources enregistrées par microphone
- Ressources de télé-radio-diffusion

Les ressources téléphoniques contiennent des données de parole réalisées par un locuteur appelant d'un téléphone fixe ou mobile. Ces données sont enregistrées sur ordinateur à l'aide d'une carte d'acquisition et d'un logiciel spécifique. Le réseau utilisé par le locuteur dépend du réseau utilisé dans le pays (PSTN, GSM, CDMA, WCDMA, etc.) et la communication est transmise par le RNIS dans la plupart des cas.

Les ressources enregistrées par microphone font usage d'un à plusieurs microphones et l'enregistrement s'effectue dans un endroit précis tel que bureau, train, voiture, ou encore en extérieur pour capter également l'environnement avec des interférences : bruits, paroles, vents, etc.

La configuration pour les enregistrements par microphone est déterminée par son application. Par exemple les ressources du projet Speecon utilisent 4 microphones, 2 portés par le locuteur, 1 à mi-distance (table top) et 1 à distance longue (far-field), pour enregistrer les locuteurs dans différents environnements (voiture, parc, rue, bureau, etc). Les ressources Speecon sont faites pour le développement des produits grand public commandés par la voix (réfrigérateurs, jouets, systèmes de navigation, etc). Dans d'autres applications, comme par exemple la voiture, l'utilisation de plusieurs microphones (*microphone array*) est souvent nécessaire. Pour les applications de dictaphone, , ou encore pour les systèmes de dialogue, les enregistrements s'effectuent le plus souvent dans un environnement calme comme un bureau avec l'usage d'un seul microphone.

D'autres ressources enregistrées par microphone sont les ressources pour la synthèse de la parole (TTS) ou des ressources pour les applications interactives. Pour la synthèse de la parole, un ou deux locuteurs avec une très bonne élocution sont en général choisis pour lire des phrases afin de couvrir une très grande partie des diphones et triphones (unités acoustiques) ainsi que des noms communs et des noms propres. La durée d'enregistrement est plus longue que les autres bases de données, de l'ordre de 5 à 10 heures. Les transcriptions sont souvent phonétiques et contiennent un jeu d'étiquettes étendues.

Dans les applications interactives les corpus lus ne sont pas suffisants pour créer de tels systèmes. Pour mieux capter la parole avec des locuteurs confrontés à des conditions réalistes et comportant des hésitations et reprises, l'enregistrement se déroule avec une machine simulée par un « compère » (intervenant humain), selon la technique du Magicien d'Oz (MOz). Le locuteur suit un scénario préétabli dans lequel le compère répond soit en prononçant lui-même la réponse soit en la déclenchant manuellement. La technique du MOz n'est pas sans limite : le comportement de la machine est difficilement simulé par l'être humain. Cette technique reste cependant indispensable au début des recherches des applications interactives.

La troisième catégorie, les ressources de télé-radio-diffusion, sont enregistrées par télévision ou par radio, souvent avec une programmation automatique sur de longues périodes de temps. Différentes émissions (débats, journaux, interviews, etc.) peuvent faire partie de l'enregistrement, ainsi que des enregistrements de chaînes diverses. En général, les émissions sont enregistrées avec une périodicité régulière, c'est-à-dire une seule fois par jour ou par semaine.

Il existe un autre catégorie de ressources orales qui peut être considérée à part de par sa nature, les ressources pour le VOIP (*Voice Over IP* – voix sur réseau IP). Actuellement il n'existe pas de standard pour ce type de ressource, car les études réalisées dans ce domaine utilisent souvent les ressources orales enregistrées par téléphone pour simuler l'environnement. Il reste à définir les paramètres comme le canal (réseau, internet), les interfaces (Skype, MSN messenger, etc.) et les outils d'enregistrements.

- ***Contenu***

Une ressource complète contient quatre parties principales :
- L'enregistrement audio de la parole
- Les transcriptions
- Le lexique
- La documentation

Les transcriptions sont effectuées manuellement ou semi-automatiquement une fois que les enregistrements ont été réalisés. Elles reflètent les énoncés du locuteurs, même si ceux-ci sont erronés, et utilisent un jeu d'étiquettes pour marquer certains événements dans le flux de parole :
- Mots mal prononcés
- Mots tronqués par l'enregistrement
- Distorsions par le canal en cas d'enregistrements téléphoniques mobiles
- Hésitations
- Bruits de bouche (souffle, toux, etc.)
- Bruits stationnaires (musique, bruit de fond, parole en arrière plan, etc.)
- Bruits ponctuels (claquement de porte, sonneries, etc.)

Le lexique contient tous les mots de la ressource triés par ordre alphabétique. Chaque entrée dans le lexique correspond à une ligne avec le mot, le nombre d'occurrences (répétitions dans le corpus) et la phonétique de ce mot.

La documentation comporte un certain nombre de sections pour décrire la ressource et préciser comment la ressource a été produite :
- Format et structure
- La collecte et le design du corpus (non utilisé dans les ressources de télé-radio-diffusion)

- Définition des items dans le corpus : chiffres, phrases, mots, etc. (non utilisé dans les ressources de télé-radio-diffusion)
- Transcriptions
- Romanisation (pour les langues non latines)
- Lexique
- Informations sur les locuteurs (âge, accent, etc.)
- Environnement de l'enregistrement

En général, pour les grands projets avec plusieurs langues, avant d'entamer la production, un document contenant les spécifications pour la langue cible est produit. Ce document appelé LSP (Language Specific Pecularities) contient le contenu du corpus lié à la langue (dates, nombres, monnaies, noms propres, etc) ainsi que la phonétique pour la langue en SAMPA et la romanisation pour les langues non latines. Des déviations éventuelles de spécifications à cause de la langue sont également décrites dans ce document.

2.1.2.3 Formats de codage

Dans une ressource orale il y a deux type des données : les signaux de parole et les transcriptions textuelles. Pour les signaux, le format dépend en grande partie de la ressource en question.

Le format utilisé pour les ressources téléphoniques est A-law en Europe et Mu-law aux Etats-Unis. Pour les ressources enregistrées par microphone, le format peut varier mais en général le signal n'est pas codé mais stocké en format brut (« raw ») et au moins en 16 bits et 16 kHz. Les ressources de télé-radio-diffusion sont dans la plupart des cas codés en 16 bits et 16 kHz et en format brut mais parfois en Wav ou en format Sphere.

Le texte des transcriptions est encodé en ISO pour les langues latines et UTF-8 pour les langues non latines, comme l'arabe ou les langues asiatiques.

Des compressions peuvent être effectuées sur les signaux à condition que cela n'engendre aucune perte. Des outils tels que Flac et Shorten sont parfois utilisés pour obtenir une compression sans perte. Ceci est appliqué plus souvent sur les ressources de télé-radio-diffusion.

En ce qui concerne le lexique, la phonétique utilisée est en SAMPA[1] (Speech Assessment Methods Phonetic Alpahbet), telle que développée dans le projet SAM pour contourner le problème d'encoder les caractères API (Alphabet phonétique international) sur un ordinateur.

2.1.2.4 Formats de stockage

En ce qui concerne les ressources téléphoniques, le format utilisé pour stocker les transcriptions est défini par le format SAM du projet Esprit. Ce format décrit les ressources avec des étiquettes pour les informations nécessaires :

- La ressource (nom, langue, identifiant, encodage de caractères)
- Le signal (nom du fichier, répertoire, lieu et date d'enregistrement, longueur)
- L'encodage du signal (fréquence d'échantillonnage, ordre de bits, nombre de bits, nombre de canaux)
- Le locuteur (homme/femme, âge, accent, numéro d'identifiant du locuteur)
- L'enregistrement (type de réseau, microphone, lieu et conditions d'enregistrement)
- Les transcriptions (texte affiché au locuteur, la parole transcrite avec marqueurs)

Les ressources enregistrées par microphone ainsi que par téléphone utilisent souvent le format SAM.

En revanche, les ressources de télé-radio-diffusion utilisent souvent le format XML ou .trs de l'outil de transcription Transcriber pour stocker les transcriptions ainsi que les paramètres comme :

- Segmentation
- Tours de parole
- Identification de phrases
- Noms des journalistes
- Marqueurs

[1] http://www.phon.ucl.ac.uk/home/sampa/

Avant leur distribution, les ressources orales sont stockées sur un support qui dépend de la taille informatique de la ressource ; on parle ici d'une ressource « hors ligne ». Une ressource hors ligne est une copie de l'original et accessible par un support informatique, tel qu'un CD-ROM, un DVD-ROM ou un disque dur. Tel est le cas pour la plupart des ressources, car un accès en ligne n'est pas aussi pratique à cause de la taille importante de ressources orales. Une ressource peut aussi être accessible entièrement (accès total), ou partiellement pour permettre d'accéder uniquement à une partie de la ressource. Grâce au libre accès, tout le monde peut accéder à la ressource et à tout moment. Mais la majorité des ressources réservent un accès restreint : une ressource accessible uniquement à un groupe préalablement établi ou réservée à une distribution commerciale, en général à des coûts permettant de couvrir les coûts de la production.

En termes de stabilité et d'évolution, les ressources peuvent être divisées en deux catégories : les ressources statiques ou les ressources dynamiques. Une ressource statique ne change pas une fois finalisée. Au contraire, les ressources dynamiques peuvent évoluer et des modifications ou mises a jour sont effectuées afin de les améliorer.

2.1.2.5 Standards de validation

La validation d'une ressource est un procédé d'assurance qualité permettant de vérifier et certifier que la ressource respecte des critères de validation. Ces critères peuvent être définis *a priori*, avant que la ressource ne soit produite ou *a posteriori* après la collecte de celle-ci, ou encore en cours de production.

La validation peut être interne, c'est-à-dire réalisée par l'entité produisant la ressource ou externe, menée par une tierce partie n'ayant pas contribué à la production de celle-ci. Ainsi la validation externe possède l'avantage d'être réalisée de façon indépendante sans conflit d'intérêt.

Dans les grandes initiatives de collecte avec plusieurs partenaires (telles que SpeechDat, Speecon, SALA, OrienTel, LILA), la validation est souvent réalisée par un organisme externe et indépendant. Cette validation suit un protocole préétabli pour s'assurer de la qualité finale de chaque ressource dans le projet. Ce protocole peut être divisé en quatre étapes :

- Pré-validation A
- Pré-validation B
- Validation
- « Pre-release » validation

Avant que la production commence, un corpus avec les items à enregistrer et un jeu de « prompts » (énoncés) à lire par le locuteur (soit sur papier soit sur écran) est produit. Ce corpus, le jeu de prompts et le lexique du corpus sont validés au cours de la première étape : la pré-validation A. Outre les vérifications du corpus et des prompts, le lexique est contrôlé par un expert phonéticien natif de la langue. L'expert vérifie que la transcription phonétique est correcte sur un échantillon de 1 000 entrées lexicales. un maximum de 5% d'erreurs est accepté. A l'issue de cette validation, lorsque nécessaire, le producteur reçoit un rapport pour corriger les défauts identifiés dans le corpus (répétitions d'items, couverture phonétique, design, etc).

Quand le pré-validation A est approuvée, le producteur peut passer à la phase d'enregistrement d'une mini-base de données de 10 locuteurs avec les mêmes conditions que prévues pour la ressource finale. Ces 10 sessions d'enregistrements, ainsi que la transcription, un lexique, la documentation et les tableaux de statistiques sont envoyés pour être validés lors de la phase de pré-validation B.
Le producteur reçoit un rapport avec les détails de cette validation et pour éventuellement faire des corrections avant la production de la base de données complète commence.

Une fois les enregistrements et les transcriptions sont terminés ainsi que la documentation et le lexique, le producteur enverra toute la base de données ou une partie de celle-ci si sa taille est importante, pour la validation finale. Lors de cette validation les parties suivantes sont vérifiées :

- La documentation
- La structure
- La complétude de la ressource
- La qualité du signal
- Les fichiers d'annotation
- Le lexique
- La répartition des locuteurs et de l'enregistrement
- Les transcriptions

La documentation doit contenir les informations administratives et techniques liées à la base de données : le contenu (les items et leur répartition), la répartition des locuteurs (âge, homme/femme, accent), la répartition sur les environnements concernés, la plateforme d'enregistrement, les outils et les conditions d'enregistrement. Il y a aussi une section sur les annotations : type, procédure, jeu d'étiquettes ; ainsi que le lexique avec des tableaux de fréquence de phonèmes pour les parties phonétiquement riches (phrases et mots).

La validation comporte aussi une vérification de la structure de la ressource : le respect de l'arborescence et des normes ISO9660 pour le stockage sur CD ou DVD, le format SAM, la nomenclature des fichiers et répertoires ainsi que les fichiers qui doivent obligatoirement être présents.

Ensuite le contenu de la ressource et sa complétude est validé. Ceci comporte une vérification des sessions, des locuteurs et des items. Tous les items sont comparés en fonction des spécifications pour vérifier leur exactitude et complétude. Néanmoins, une certaine marge est tolérée. Ces valeurs sont en général calculées sur un échantillon de la ressource et ensuite extrapolées sur un intervalle de confiance. De même, la répartition des locuteurs est comparée aux spécifications : nombre, accent, homme/femme. Enfin, la richesse phonétique est aussi vérifiée pour les phrases et mots phonétiquement riches.

Une vérification des fichiers est également effectuée. Cette partie consiste à vérifier trois parties : le nombre des fichiers manquants, les fichiers contenant une transcription vide (max. 5%) et les fichiers ne contenant que des mots tronqués ou mal prononcés.

Lors de la validation des signaux de la parole, des vérifications sont faites pour l'encodage (nombre de bits, fréquence d'échantillonnage) et la distribution des fichiers par rapport à leur taille, taux de clipping, moyenne, et SNR (rapport signal-bruit). De même, la période de silence en début et fin du signal est vérifiée : un maximum de 10% des signaux peut avoir un silence plus court que la durée indiquée dans les spécifications.

Ensuite, les fichiers d'annotation sont vérifiés pour s'assurer que le format SAM est respecté et que les champs ne contiennent pas des valeurs illégales.

Le lexique est aussi validé pour contrôler la couverture des phonèmes, le format du lexique ainsi que la complétude et le compte des mots.

La validation de la distribution comporte les locuteurs (les tranches d'âges, nombre d'hommes/femmes, les accents) ainsi que les enregistrements (par environnements et accents).

La validation des transcriptions est composée de deux parties : la première est effectuée automatiquement avec des logiciels pour vérifier uniquement la présence des caractères orthographiques (ni les chiffres ni les symboles ne sont autorisés), l'absence de caractères illégaux et que seuls les marqueurs valables sont utilisés. La deuxième partie est effectuée par un expert natif de la langue et comporte la vérification de 2 000 entrées de transcriptions. Les erreurs sont notées et communiquées dans le rapport de validation. Pour les erreurs sur les marqueurs de bruit, 20% d'erreurs est accepté alors que pour la parole seul 5% est accepté.

Si le taux d'erreur dépasse le seuil défini dans les spécifications, il peut s'avérer nécessaire de refaire une partie de la base de données pour ensuite repasser une ou plusieurs étapes dans la validation.

Avant la distribution et pour certifier que toutes les modifications après la validation ont été effectuées, une dernière validation est réalisée: la validation de la version avant publication, nommée « pre-release ». Le seul document qui s'ajoute à cette version est la version finale du rapport de validation.

2.1.3 Ressources multimodales

2.1.3.1 Historique

Les ressources multimodales sont des ressources intégrant plusieurs modalités de communication différentes telles que : la voix, la gestuelle, la posture, le regard, l'écriture, l'image, etc.

Un corpus multimodal combine ainsi des données de différentes natures : ressources visuelles (images, vidéos), ressources sonores (enregistrements audio, divers types de sons), ressources textuelles, etc. Ce type de corpus contient nécessite la production de différents types d'annotations : marqueurs visuels, transcriptions audio, segmentation audiovisuelle, etc.

Le périmètre des domaines d'étude qui nécessitent la production de ce type de ressources est très large :

identification et suivi de personnes ou d'objet dans une vidéo, reconnaissance de gestes, analyse des interactions humaines, reconnaissance des émotions, reconnaissance automatique de caractères, etc.

Depuis plus de dix ans, nombre de grands projets internationaux ont nécessité la collecte de bases de données multimodales. C'est le cas notamment des projets d'analyse et d'indexation automatique de documents audiovisuels, tels que :

- le *broadcast news* et autres programmes télévisés : campagnes d'évaluation internationales TRECVID[1] organisées par le NIST (Smeaton, 2006), projet Informedia[2] de l'université Carnegie Mellon aux Etats-Unis (Hauptman, 2005), etc.
- les films de cinéma : projet européen MUSCLE[3] (Spachos, 2008),
- les enregistrements de sessions du parlement européen : projet européen Reveal-This[4] (Pastra, 2006).

Toutefois, il s'agit là de traiter des données déjà produites par un tiers (chaînes de télévision, maisons de production, archives parlementaires, etc.). La constitution des corpus consiste alors à rassembler les données déjà disponibles (après négociation des droits avec leur propriétaire), et éventuellement à les sélectionner et à les reformater pour les besoins spécifiques des traitements à mettre en oeuvre. Il ne s'agit pas là d'un travail de production de données à proprement parler. En revanche, ces projets nécessitent généralement un important travail d'annotation des données en vue d'une campagne d'évaluation par exemple.

Plus récemment, des moyens croissants sont investis dans des projets d'un autre type. Il s'agit de projets portant sur l'analyse automatique des interactions personne-personne et personne-machine, mais aussi des interactions de la personne et de son environnement quotidien. Les technologies mises en jeu sont très variées : détection d'activité, reconnaissance des gestes, reconnaissance des expressions faciales, suivi 3D d'une ou plusieurs personnes dans un espace donné, etc.

Il n'existe pas en général de ressources sur étagère satisfaisant aux besoins de ces nouvelles technologies, et chaque projet consacre la majeure partie de son budget à collecter ses propres données *ad hoc*. Ces recherches ont ainsi déjà contribué et continuent à contribuer fortement au développement d'infrastructures et la définition de protocoles pour la collecte de larges bases de données multimodales.

Un certain nombre de grands projets se sont en particulier intéressés à l'analyse des interactions humaines, dans le cadre de réunions de travail notamment. Cela nécessite la collecte de large corpus d'enregistrement de réunions en salle close (séminaires, cours, réunions interactives). Si la collecte de ce type de corpus s'est au début davantage focalisé sur la composante audio, comme ce fut le cas pour les corpus ICSI (Janin, 2003) et ISL (Burger, 2002), de nombreux corpus de données véritablement multimodaux, faisant intervenir un grand nombre de canaux audio et vidéo, ont depuis été créés, notamment dans le cadre de grands projets internationaux tels que :

- les projets européens M4[5], AMI[6] et CHIL[7],
- les projets Smart Space[8] et VACE[9] du NIST.

Les bases de données audio-visuelles produites dans le cadre de ces projets étaient souvent destinées à la mise en œuvre de campagnes d'évaluation. Citons notamment les campagnes d'évaluation CLEAR[10] (Mostefa, 2006 ; Stiefelhagen, 2007) organisées conjointement par les projets CHIL (Moreau, 2008) et VACE (Chen, 2005).

Une grande partie des bases de données multimodales existantes résultent de l'enregistrement de réunions interactives (plusieurs personnes prenant chacune une part active à une réunion, dans des salles équipées de multiples capteurs audio et vidéo). Dans chacune des salles utilisées, les réunions sont filmées sous différents angles et enregistrées par différents types de microphones (distants, personnels, etc.) De nombreux problèmes ont dû être résolus tant au niveau de la capture, du stockage, que de l'annotation des données recueillies, permettant ainsi d'affiner les procédures et les spécifications de la collecte de données, au fur et à mesure de l'avancée de ces projets (Casas, 2004).

[1] TRECVID: http://www-nlpir.nist.gov/projects/trecvid/

[2] Informedia Project: http://www.informedia.cs.cmu.edu/

[3] Projet MUSCLE: http://www.muscle-noe.org/

[4] Projet Reveal-This: http://www.reveal-this.org/

[5] M4 project (Multimodal Meeting Manager): http://www.dcs.shef.ac.uk/spandh/projects/m4/

[6] AMI (Augmented Multiparty Interaction): http://www.amiproject.org

[7] CHIL (Computers in the Human Interaction Loop) : http://chil.server.de

[8] NIST Smart Space: http://nist.gov/smartspace

[9] VACE (Video Analysis Content Extraction) : http://www.informedia.cs.cmu.edu/arda/index.html

[10] CLEAR (Classification of Events, Activities, and Relationships Evaluation) : http://www.clear-evaluation.org

Un autre type de données multimodales concerne l'enregistrement d'activités humaines à domicile (réalisés dans des *smart homes*, c'est à dire des appartements ou des maisons équipées d'une batterie de capteurs audio et vidéo). Des bases de données de ce type ont été et sont collectées dans le cadre de plusieurs projets européens et américains récents ou en cours, tels que, entre autres :

- le projet AwareHome[1] de l'université GeorgiaTech (Kidd, 1999),
- le HomeLab[2] de Phillips Research à Eindhoven, utilisé pour étudier les interactions humaines avec des interfaces et des appareils domestiques intelligents (Aarts, 2002),
- le projet PlaceLab[3] du MIT (Intille, 2006).

D'autres projets de ce type sont à l'heure actuelle en phase de lancement. C'est le cas du projet européen NETCARITY[4] qui met actuellement en place une structure de collecte pour constituer de larges bases de données audiovisuelles portant sur l'enregistrement des activités quotidiennes d'individus âgés ou handicapés dans un environnement domestique (Cappelletti, 2008). Le but est de développer des systèmes automatisés d'aide à domicile.

Enfin, notons qu'il existe un grand nombre de champs d'étude plus circonscrits qui nécessitent également la constitution de ressources multimodales bien spécifiques, tels que, entre autres :

- la reconnaissance de parole audiovisuelle (Chitu, 2008),
- la reconnaissance audiovisuelle des émotions (Colletta, 2008),
- l'authentification biométrique multimodale (projet M2VTS[5], utilisant les caractéristiques du visage et de la voix de la personne à authentifier), etc.

La suite de ce chapitre a pour but de donner, d'une part, un aperçu des pratiques mises en oeuvre par ces différents projets en matière de collecte de corpus de données mutlimodales et de proposer, d'autre part, une méthodologie générale permettant de guider la mise en place d'un projet de production pour ce type particulier de données.

2.1.3.2 Spécifications

Un travail de spécification est nécessaire en préambule de tout projet de collecte de données. Dans le cas de la collecte de données multimodales, ces spécifications doivent définir :

- Le(s) scénario(s) retenu(s) pour collecter les données: environnement (extérieur, intérieur, degrés de luminosité, etc.), cadrage vidéo, interaction entre les sujets, etc.
- La nature et la masse des flux de données requis : nombre d'images, de flux audio, de flux vidéo, etc.
- L'infrastructure de collecte à mettre en œuvre : les différents outils de capture requis et leur disposition spatiale.
- Le format des données collectées.
- La nature et le format des annotations à réaliser sur les données brutes : transcriptions audio, segmentation vidéo, apposition de marqueurs sur des images, etc.

Un aspect fondamental pour la collecte de ce type de données, est de faire en sorte que les corpus recueillis soient cohérents et réalistes en regard des applications envisagées. Par exemple : souhaite-t-on enregistrer des séminaires interactifs ou non, s'il s'agit de collecter des données de type « réunion » ? Et s'il s'agit de séminaires interactifs, quel degré d'interactivité est-il souhaité ? Combien de personnes différentes souhaite-t-on enregistrer ? etc. Il s'agit donc, avant de procéder à la mise en place des instruments de collecte, de définir avec soin un ou plusieurs *scénarios* types qui dépendent étroitement de la nature des tâches traitées par le projet (reconnaissance d'objet, identification de personnes, détection d'événement, etc.).

En particulier, il convient de définir si les scènes enregistrées doivent être entièrement scénarisées à l'avance (exemple : réunions pour le projet AMI ou activités domestiques pour le projet NETCARITY) ou spontanées (exemple : réunions CHIL) en fonction des besoins propres au projet. La difficulté est de mettre en place des scénarios de collecte reflétant la réalité des comportements et des situations dans l'environnement concerné, tout en assurant une diversité suffisante en termes d'environnements (il est souhaitable de disposer de plusieurs installations de collecte dans des lieux différents) et d'événement (exemple : variété des événements acoustiques) afin de réaliser des évaluations qui soient pertinentes (Zancanaro, 2004).

[1] Aware Home Research Initiative: http://awarehome.imtc.gatech.edu/

[2] Phillips HomeLab: http://www.research.philips.com/technologies/misc/homelab/

[3] PlaceLab Initiative: http://architecture.mit.edu/house_n/placelab.html

[4] Projet NETCARITY: http://www.netcarity.org/

[5] M2VTS Project (Multi-modal Biometric Person Authentification): http://www.tele.ucl.ac.be/PROJECTS/M2VTS/

Enfin les spécifications doivent définir les différentes étapes chronologiques du processus de collecte des données :

- La collecte d'une base de données de *pré-validation* permettant, à partir d'un petit échantillon de données, de vérifier la pertinence de l'infrastructure de collecte mise en place et d'identifier d'éventuels problèmes.
- La collecte des données à proprement parler.
- La validation des données.
- L'annotation des données brutes.
- La validation des annotations.
- La rédaction d'une documentation complète décrivant les données collectées.

2.1.3.3 Infrastructure de collecte de données audiovisuelles

La collecte d'un corpus de données audiovisuelles suppose la capture simultanée d'un grand nombre de canaux audio et vidéo de différentes natures. A titre d'exemple, les salles d'enregistrement multi-capteurs (*smart rooms*) mises en place dans le cadre du projet CHIL, étaient équipées du jeu de capteurs suivant (Casas, 2004) :

Capteurs audio :

- un « microphone array » 64 canaux linéaire,
- 3 microphones 4 canaux « en T »,
- 3 microphones de tables,
- une série de microphones personnels portés par un ou plusieurs locuteurs participants.

Capteurs vidéo:

- 4 caméras fixes aux 4 coins de la pièce,
- une caméra panoramique grand angle fixée au plafond,
- une caméra active "pan-tilt-zoom".

Il n'existe cependant pas de norme bien définie pour mettre en place une telle infrastructure de collecte multimodale. Chaque projet a ses spécificités, et d'autre *smart rooms*, comme celle mise en place dans le cadre du projet M4 (McCowan, 2003), repose sur une autre architecture multi-capteur.

Flux audio
Il n'y a pas de recommandations particulières quant au choix du matériel utilisé pour les prises de sons à distance (*far-field microphones*) ou rapprochées (*near-field microphones*). Chaque site de collectes utilise différentes marques de microphones personnels, de microphones de table ou de microphones 4-canaux, dits « en T ».

Cependant, il est communément fait usage de *microphone arrays* (séries de microphones alignés) permettant une prise de son multi-canal pour faire du *beamforming* (formation de faisceaux) et de la localisation spatiale de source sonore. Pour ce type d'équipement, la norme *Mark III array*[1] s'est largement imposée. Il s'agit d'un *microphone array* linéaire de 64 canaux développé par le NIST.

Les infrastructures d'enregistrement d'activités humaines domestiques (dans des *smart homes*) excluent l'usage de microphones personnels (*close talking microphones*), uniquement pertinents dans le cadre de réunions de travail ou de colloques. Dans ce cas, la prise de son est réalisée au moyen de microphones distants (*far field microphones*). Le projet NETCARITY a ainsi fait le choix d'utiliser des microphones multidirectionnels 4 canaux, dits *en T*, fixé à une hauteur de 2 mètres sur les murs de son *smart home* (Cappelletti, 2008).

Flux vidéo
Là encore, il n'y a pas de standard particulier quant au choix du matériel utilisé. Les différents projets de collecte de données multimodales cités précédemment ont aussi bien utilisé des caméras numériques que digitales, avec des niveaux de résolution allant de 288×360 à 1024×768 pixels par image, et des taux d'échantillonnage allant de 15, 25 et jusqu'à 30 images par seconde.

Dans les spécifications du projet CHIL (Casas, 2004), il est toutefois recommandé d'utiliser une résolution vidéo minimum de 640×480 pixels nécessaire à la mise en œuvre de certaines technologies telles que le reconnaissance d'objet, de personne, etc.

[1] The NIST MarkIII Microphone Array: http://www.nist.gov/smartspace/cmaiii.html.

Pour les prises de vue dans un environnement domestique, le projet NETCARITY a mis en place des web-cams PTZ (pan-tilt-zoom) de résolution 640×480, offrant la possibilité d'être commandées à distance et de permettre des prises de vue grand-angle dans des pièces relativement exiguës (Cappelletti, 2008).

Lorsqu'un jeu de plusieurs caméras fixes est utilisé, il est nécessaire de calibrer toutes les caméras à l'aide d'un système de coordonnées de référence afin de pouvoir ultérieurement déduire les coordonnées 3-D d'un point de l'espace, à partir de plusieurs prises de vue. La calibration se fait le plus souvent manuellement à l'aide d'une mire visible par les différentes caméras filmant la scène. Il existe toutefois des procédures itératives et automatiques de calibration, notamment via l'outil *Multi-Camera Self-Calibration*[1] (Lathoud, 2004).

Enfin, il est communément recommandé d'accompagner les enregistrements vidéo de brèves séquences « à vide » où la scène est filmée sans qu'aucune personne ni aucun objet mouvant n'apparaisse (ex : salle de réunion vide, parking désert, etc.). Il est indispensable d'inclure ces données dans les corpus multimodaux pour permettre aux algorithmes de traitement de l'image de modéliser l'arrière plan des différents lieux de collecte.

Synchronisation
Au-delà du choix du jeu de capteurs, le choix de l'architecture globale du système de collecte a un impact primordial sur la qualité des corpus produits lorsqu'il s'agit de données multimodales.

Les capteurs utilisés pour la collecte de données audiovisuelles doivent s'intégrer dans une plateforme capable de coordonner l'ensemble du processus d'enregistrement.

Pour reprendre l'exemple du projet CHIL, les spécifications pour la collecte de données étaient les suivantes :

- Pour l'audio : les différents canaux (hors MarkIII) étaient enregistrés sur un ordinateur via une carte RME Hammerfall HDSP9652 I/O. Les 64 canaux des « microphones array » NIST Mark III étaient acquis au moyen d'un autre ordinateur, via une connexion Ethernet sous la forme de paquets IP multiplexés.
- Pour la vidéo: un ensemble de machines dédiées assure la sauvegarde simultanée de tous les flux vidéo sous forme de séquences d'images compressées au format JPEG.

Cet exemple illustre bien le fait que la capture simultanée d'une telle quantité et d'une telle variété de flux de données nécessite le support d'une infrastructure informatique complexe.

L'un des principaux problèmes à résoudre dans le cas des enregistrements de données multimodales est de garantir l'alignement temporel permanent entre un grand nombre de flux collectés simultanément. Cela nécessite la mise en œuvre de procédures de synchronisation des différents canaux audio et vidéo.

Pour répondre à ce problème, en particulier dans le cadre des projets Smart Space et VACE, le NIST a développé un système (NIST Smart Data Flow architecture) permettant la capture simultanée d'un grand nombre de flux audio et vidéo tout en assurant la synchronisation de tous les flux (via le *NIST atomic clock signal*) à quelques millisecondes près (Garofolo, 2004).

De la même façon, AMI utilise différents appareils de synchronisation émettant des signaux de référence sur lesquels se calent les différents flux de données audio-visuelles collectées (Moore, 2002).

Cependant de telles infrastructures restent lourdes à gérer et il est fréquent qu'un problème technique oblige à re-synchroniser tout ou partie des données à la main, ce qui représente un travail long et fastidieux. Par précaution, il est donc recommandé dans tous les cas d'avoir recours en début et fin d'enregistrement à l'usage d'un « clap » cinématographique émettant un signal lumineux et sonore facilitant une (re-)synchronisation manuelle des données si nécessaire (Martial, 2001 ; Moreau, 2007).

2.1.3.4 Formats de données collectées

- ***Données audio***

Le format de codage audio le plus communément employé est le format SPHERE développé par le NIST.

C'est en particulier le format utilisé dans le projet CHIL pour coder les 64 canaux enregistrés par les "microphone arrays" NIST Mark III. Ce format permet aisément en particulier d'encoder les différents canaux

[1] Svoboda, T.: Multi-Camera Self-Calibration : http://cmp.felk.cvut.cz/~svoboda/SelfCal/

dans un seul fichier.

Les données audio collectées par le NIST (VACE, NIST Smart Space), initialement encodées au format NIST-SMD (48-KHz/24-bit) sont ensuite également converties au format SPHERE (en les sous-échantillonnant à 16-KHz/16-bit) en vue de leur stockage et de leur distribution.

Dans les projets CHIL et AMI, les données audio des microphones isolés (microphones mono-canaux ou microphone 4 canaux dits « en T ») sont toutefois codés au format WAV (44.1 kHz pour CHIL, 16kHz pour AMI, et 24 bits).

- ***Données vidéo***

Dans les projets CHIL et NETCARITY, les flux vidéo sont codés sous forme de séquences d'images compressées au format JPEG avec des taux d'échantillonnage allant de 10 à 30 images par seconde. Chaque séquence d'images JPEG collectée dans le cadre de CHIL est accompagnée de 2 fichiers de spécification : l'un pour fournir un certain nombre de caractéristiques de l'enregistrement (taux d'échantillonnage etc.), l'autre pour donner la liste des *time stamps* associés à chaque image de la séquence.

Les formats vidéo MPEG sont largement utilisés. Les corpus de données vidéo des campagnes d'évaluation TRECVID[1] sont disponibles au format MPEG-1. La norme MPEG-2 est utilisée pour encoder les flux vidéo dans nombre de projets : Smart Space, VACE, AMASS++[2], etc.

Le NIST (projets Smart Space et VACE) encode d'abord ses données vidéo au format propriétaire NIST-SMD, au moment de leur enregistrement. Les vidéos collectées sont ensuite converties du format SMD au format MPEG-2 en vue de leur stockage et de leur distribution.

Dans le projet AMI, on trouve également des données vidéo au format DIVX AVI.

Le stockage des données vidéos requiert en général beaucoup plus d'espace que les autres types de médias. On peut estimer que 30 minutes de données vidéo de qualité (par exemple 30 minutes d'un bulletin d'information télévisé) requièrent autour de 10 Go d'espace mémoire. A titre de comparaison, un signal audio stéréo de qualité CD et de durée comparable nécessite autour de 300 Mo de stockage. Les données télétexte du bulletin d'information n'occupe quant à elle qu'un espace de l'ordre de 10 Ko. Les données vidéo peuvent être compressées, mais cela ne peut se faire sans une réduction plus ou moins sensible de la qualité des images. Avant de mettre en oeuvre un projet de collecte de données vidéo, il faut donc réfléchir au compromis souhaité entre qualité et taille des données, et dimensionner en conséquence l'infrastructure de stockage à mettre en oeuvre.

2.1.3.5 Formats des annotations

Le travail d'annotation représente une part très importante de la création des ressources multimodales puisqu'il s'agit d'annoter un très grand nombre de signaux différents. Il s'agit en général de produire les méta-données suivantes :

- Segmentation : séquence vidéo (scènes), type de son (musique, parole, bruit, silence), tours de parole, etc.
- Transcriptions manuelles d'un ou plusieurs canaux audio : transcriptions orthographiques dans le cas de la parole, entités nommées, tours de paroles des différents locuteurs, annotations d'événements sonores, etc.
- Annotations manuelles des images d'un flux vidéo à l'aide de marqueurs spécifiques aux technologies à évaluer : marquage des points clés (yeux d'un visage) ou des zones d'intérêt (rectangle encadrant un véhicule) de l'image.
- Enfin, certains projets ont généré des annotations vidéo 3D, dérivées des annotations 2D réalisées pour différentes caméras filmant la même scène sous différents angles (Lathoud, 2004).

Là encore, chaque projet a proposé ses propres outils et procédures d'annotation.

- ***Transcriptions audiovisuelles***

[1] TREC Video Retrieval Evaluation http://www.itl.nist.gov/iaui/894.02/projects/trecvid/

[2] AMASS++ : http://www.cs.kuleuven.be/~liir/projects/amass/

Il existe un grand nombre d'outil de production et de formats associés pour les transcriptions de données audio.

Pour le projet CHIL, la transcription audio s'est faite en plusieurs passes, en commençant par utiliser le logiciel Transcriber[1] : transcription orthographique en condition « *near-field* », affinée par une deuxième écoute de la même séquence en condition « *far-field* » (Mostefa, 2007). Enfin une troisième passe permettait l'annotation des événements acoustiques et classes de sons à l'aide de l'outil Annotation Graph Tool Kit (AGTK)[2], qui permet, contrairement à Transcriber, l'annotation de plusieurs événements sonores simultanés.

Ces outils produisent des transcriptions au format TRS (format XML spécifique à l'outil Transcriber) ou AG (format XML spécifique à l'outil AGTK). D'autres formats de transcription audio sont communément utilisés, notamment les formats STM et CTM (formats de transcription définis par le NIST).

Un outil également très populaire pour la labellisation de segments audio est le logiciel Praat[3]. Les transcriptions sont produites au format Textgrid spécifique à l'outil Praat.

Toutefois ces outils ne concernent que l'annotation de la modalité audio. L'annotation de données multimodales nécessite souvent de recourir à des outils permettant de visualiser simultanément le son et l'image. C'est le cas par exemple de l'annotation de l'état émotif d'une personne (Colletta, 2008), qui peut être caractérisé par des indices aussi bien verbaux (prosodie) que non verbaux (posture, expression faciale).

Il existe un certain nombre d'outils freeware permettant une annotation véritablement multimodale d'un document audiovisuel, l'un des plus connu étant le logiciel ANVIL[4], très largement utilisé par la communauté multimodale. Les transcriptions audiovisuelles produites sont stockées au format ANVIL (format XML spécifique à l'outil).

Dans le projet AMI, les transcriptions ont été réalisées à l'aide de l'outil NXT (NITE XML Toolkit)[5] et sont stockées au format XML spécifique à cet outil. L'outil NITE (Natural Interactivity Tools Engineering) propose une plateforme pour l'annotation et l'analyse de conversation, flexible et configurable pour l'encodage de conversations et de corpus multimodaux.

Enfin, le standard MPEG7[6] (Multimedia Content Description Interface) peut être utilisé pour décrire l'ensemble des annotations attachées à un flux de données audiovisuelles. La norme MPEG7 est un standard ISO/IEC qui définit un ensemble de descripteurs, écrits en langage XML, destinés à la description des méta-données associées à tous types de données multimédia (en particulier : descripteurs destinés à la caractérisation des images, à la description d'une segmentation vidéo, etc.). C'est le choix qui a été retenu pour l'annotation des données vidéo dans le cadre du projet MUSCLE (Spachos, 2008).

- ***Annotations de marqueurs visuels***

L'annotation manuelle de marqueurs visuels sur une image ou une séquence de trames vidéo (marquage de la position d'un objet, orientation de la tête, etc.) requiert des outils plus spécifiques, dont il est difficile de trouver des versions génériques, librement accessibles. De même, il n'y a guère de standards bien définis et largement utilisés pour le format des annotations produites. La spécificité d'un projet de collecte oblige en général à développer une interface et un format d'annotation *ad hoc*.

Ce fut le cas, dans le cadre du projet CHIL par exemple, pour l'annotation des visages sur les séquences vidéo. Une interface de marquage *ad-hoc* a été créée dans le cadre du projet. Un format de fichier a été défini pour stocker les annotations visuelles produites par cet outil (simple fichier texte avec formatage sur chaque ligne des coordonnées spatiales des labels visuels).

2.1.3.6 Procédures de validation

- ***Validation des données brutes***

[1] Transcriber Tool: http://trans.sourceforge.net

[2] The AGTK Annotation Tool: http://agtk.sourceforge.net.

[3] PRAAT: http://www.fon.hum.uva.nl/praat/

[4] ANVIL: http://www.anvil-software.de/

[5] NITE XML Toolkit (NXT): http://weblex.ens-lsh.fr/projects/xitools/logiciels/NITE/nite.htm

[6] MPEG7: http://www.cselt.it/mpeg/standards.html

La mise en place d'une procédure de validation des données recueillies implique la définition d'un ensemble de critères de qualité minimaux auxquels les données collectées doivent satisfaire.

Une première phase, dite *pré-validation*, consiste à produite au préalable un petit nombre de données, et à vérifier la conformité de ce corpus (dit *corpus de pré-validation*) aux critères de qualité retenus.

Le projet VACE a utilisé une telle procédure de pré-validation :
- Collecte d'un *micro-corpus* (il s'agissait de données vidéo).
- Envoi du *micro-corpus* aux participants.
- Discussion entre participants pour partager leur retour d'expérience
- Révision du protocole de production.

La collecte des données proprement dite (corpus de développement, d'entraînement et d'évaluation) peut ensuite commencer, tout enregistrement ne satisfaisant pas aux exigences spécifiées dans le protocole devant être refait.

Dans le cadre du projet CHIL, une procédure de validation interne a été mise en place, autour des critères de qualité suivants (Moreau, 2007) :
- *Collecte des données vidéo* : la désynchronisation maximum autorisée entre les différents flux vidéo était de 200ms.
- *Collecte des données audio des « microphone arrays »* : la désynchronisation maximum autorisée entre les différents canaux (suite à une perte de paquet) était de 200ms.
- *Collecte des données audio pour les autres micros* : la désynchronisation maximum autorisée entre les différents canaux (suite à une perte de paquet) était de 50ms.

Chaque « CHIL-room » devait d'abord produire un enregistrement de pré-validation afin, le cas échéant, de modifier son infrastructure de collecte pour qu'elle soit conforme à ces critères. Par la suite, tout enregistrement ne satisfaisant pas aux exigences ci-dessus devait être refait.

- ***Validation des annotations***

En ce qui concerne la validation des transcriptions audiovisuelles, la procédure généralement utilisée est de faire revérifier une transcription de façon croisée par 1 ou 2 opérateurs différents du transcripteur initial (*cross-validation*). Si la masse de donnée est très importante, ce travail peut être réalisé sur la base d'échantillons de transcription pris au hasard.

Dans le cas de tâches d'annotation très spécifiques faisant intervenir une forte part de subjectivité (comme c'est le cas par exemple pour l'annotation de l'état émotif d'une personne à partir d'indices sonores ou visuels) une autre stratégie consiste à faire annoter une même séquence deux fois, par deux annotateurs différents (Colletta, 2008). Au cours d'une deuxième passe, un troisième opérateur (différent des 2 précédents) est alors chargé de comparer manuellement les 2 séquences d'annotations produites et de prendre une décision finale en cas de désaccord entre les 2 annotateurs initiaux.

Il peut être toutefois trop fastidieux et trop coûteux de valider manuellement certaines annotations produites pour un corpus multimodales. C'est notamment le cas de l'annotation de séquences vidéo à l'aide de marqueurs visuels (par exemple : marquage de la position des yeux ou des mains de toutes les personnes visibles sur les trames d'une séquence vidéo). Si la vérification trame par trame des labels produits n'est pas envisageable, un certain nombre de mesures peuvent être prises pour s'assurer de la qualité des annotations produites.

Tout d'abord, il est conseillé de demander au producteur des données brutes (séquences vidéo) qu'il accompagne chaque fichier de données d'un certain nombre d'informations descriptives complémentaires destinées à faciliter le travail des annotateurs, et à le rendre plus fiable :
- Description du contexte : lieu, date, sujet de la réunion s'il s'agit d'un enregistrement de réunion, etc.
- Description précise des objets à annoter dans la séquence. Par exemple, s'il s'agit d'annoter les visages des personnes présentes, il est nécessaire de produire une liste descriptive de ces personnes : identité, sexe, caractéristiques physiques, habillement ou toute autre information permettant de les repérer aisément à l'image.

La qualité de ces informations et de leur présentation (dans un manuel de production à l'usage des annotateurs) permet de limiter les risques d'erreurs et de confusion (par exemple : le risque qu'un annotateur intervertisse par inadvertance les identifiants de 2 personnes à annoter).

Ensuite, une fois produites, les annotations visuelles bas niveau peuvent se prêter à une passe de vérification automatique. Une telle procédure de validation des marqueurs faciaux (cadre du visage, yeux, etc.) a été mise en place au cours du projet CHIL (Moreau, 2007). Les fichiers d'annotation étaient scannés à l'aide d'un script spécialement développé pour le projet, permettant de détecter les erreurs grossières (ex : inversion des yeux, œil placé en dehors du cadre du visage, etc.). Finalement, au cours d'une 2ème passe, un opérateur humain vérifiait manuellement les trames identifiées comme suspectes par l'outil automatique (variation brutale de la position d'un marqueur d'une trame à l'autre, forme peu habituelle d'un marqueur de visage, etc.). La personne chargée de la validation était différente de l'annotateur initial.

2.2 Standards et bonnes pratiques pour la diffusion des ressources linguistiques

Pour un meilleur partage d'information et d'archivage des ressources linguistiques (on parle souvent d'interopérabilité), il s'est vite avéré indispensable de définir en ensemble de critères de description standardisés. Le terme communément usité pour définir ces formats de description est « Méta-données ». Les méta-données sont en fait des données sur des données. Celles-ci servent par exemple au catalogage et à la recherche des informations disponibles sur différents sites.

Différents schémas de méta-données ont été devéloppés pour décrire des ressources linguistiques. Parmi les schémas existants, nous développons ici :

- OLAC (*Open Language Archives Community*) : communauté pour la création d'une bibliothèque virtuelle de ressources linguistiques internationale,
- IMDI (*International Standards for Language Engineering Metadata Initiative*): initiative pour la standardisation des métadonnées utilisées pour décrire les ressources linguistiques,
- Catalogue de ressources linguistiques ELRA: catalogue recensant plus de 900 ressources linguistiques comportant des descriptions formalisées, mis en place par l'Association européenne pour les ressources linguistiques (ELRA),
- Catalogue LDC : catalogue de ressources linguistiques en particulier produites par le LDC (Linguistic Data Consortium) ou résultant de projets financés par le gouvernement américain.

2.2.1 OLAC (Open Language Archives Community)

OLAC[1] est un groupement collaboratif d'institutions et individus, fondé en décembre 2000, ayant pour but de créer une bibliothèque virtuelle mondiale de ressources linguistiques. Les axes principaux de travail sont:

- le développement d'un consensus sur les bonnes pratiques en termes d'archivage numérique des ressources linguistiques,
- le développement d'un réseau de centrales de données et services pouvant interagir ensemble pour un meilleur archivage et une meilleure mise à disposition des ressources.

En effet, force est de constater que les informations disponibles sur les ressources linguistiques restent très disparates. On peut les trouver à de multiples endroits : listes de diffusions, index web, catalogues d'archives et éditeurs… OLAC vise ainsi à joindre toutes ces informations en développant une infrastructure unique de recherche de ressources linguistiques.

L'outil principal d'OLAC est basé sur la création d'un ensemble de méta-données standardisé et agréé par un grand nombre d'acteurs du domaine permettant de décrire de façon uniforme des ressources linguistiques. Au-delà de ces méta-données et de leur usage idéalement déployé, OLAC a produit un outil de recherche permettant de récupérer toutes les informations disponibles sur le web archivées sous la forme de ces méta-données.

Le jeu de méta-données OLAC est disséminé via un protocole de récuperation de méta-données (*metadata harvesting protocol*) de l'Open Archives Initiative (OAI)[2]. Les utilisateurs peuvent ainsi accéder aux méta-données via le système d'indexation d'archives basé sur les méta-données du Dublin Core (compatible OLAC) ou via le le système OLAC basé sur le jeu de méta-données OLAC.

Il est important de noter que même si un utilisateur exploite le jeu de méta-données Dublin Core et OLAC, cela n'empêche pas de conserver ses archives dans son format d'origine. En effet, l'outil de récupération (harvester) n'utilise que le schéma exporté dans le format OLAC.

Actuellement, l'outil de recherche d'OLAC se présente sous la forme d'une mode de recherche textuel simple accessible sur internet permettant une recherche sur l'ensemble des archives utilisant le jeu de méta-données OLAC (voir impression écran ci-après).

[1] OLAC : http://www.language-archives.org

[2] Open Archives Initiative: http://www.openarchives.org

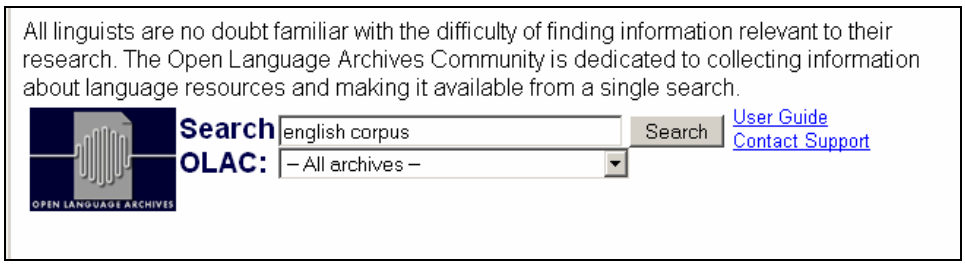

Figure 1 : Impression écran de l'outil de recherche OLAC

2.2.2 IMDI (International Standards for Language Engineering Metadata Initiative)

L'Europe n'est pas non plus en reste en termes de descriptions de ressources. En effet, une initiative du même acabit qu'OLAC a été créée. Ici encore, un jeu de méta-données a été produit, ainsi qu'un outil permettant de naviguer entre les différentes ressources décrites dans ce format.

Le projet IMDI[1] est en partie basé sur des conventions et standards existants de la communauté des ressources linguistiques. Notamment, il se base sur l'usage régulier de « headers » au sein de corpus tels que Childes[2] ou la banque de données ESF (Second Language Databank)[3]. Ces « headers » fournissent en effet en tête de fichier un ensemble de description du contenu du fichier. IMDI s'inspire également en grande partie sur les travaux de la TEI (Text Encoding Initiative) et de sa version adaptée aux corpus CES et xCES.

Ainsi, le schéma IMDI proposé est fourni sous un format XML, format largement utilisé dans la communauté des données numériques.

L'outil de recherche développé dans IMDI se distingue de l'outil OLAC par le fait que l'on accède à un outil de navigation accessible via une application Java (Java WebStart) et donnant accès à l'ensemble des archives en liste. Depuis cette liste, l'outil permet de naviguer entre les fichiers de chaque ressource archivée. Pour chaque fichier, sont affichées les informations renseignées dans les descripteurs IMDI à la discrétion du fournisseur.

En-dehors de l'outil de recherche, IMDI a également développé deux autres outils :
- IMDI Editor : outil permettant de créer des descriptions de ressources selon le schéma IMDI,
- IMDI CVEditor : éditeur de vocabulaire contrôlé.

2.2.3 Catalogue de ressources linguistiques ELRA

Dès la création d'ELRA, la question de présentation des ressources linguistiques mises à disposition par ELRA s'est posée. En effet, la mission d'ELRA étant basée sur la distribution de ressources linguistiques, il fallait trouver un moyen de décrire de façon claire les ressources disponibles et publier ces informations sur divers supports, mais principalement sur internet. Ainsi, ELRA a tout d'abord publié un premier catalogue de ressources linguistiques en ligne sous un format HTML non dynamique.

Très rapidement, ELRA a été amenée à utiliser avec ses fournisseurs des formulaires de description très détaillés afin d'obtenir des descriptions à même de répondre aux questions des utilisateurs potentiels (standards utilisés, types d'annotation, nombre d'entrées, etc.). Ainsi naissait le premier jeu de méta-données ELRA, quoi que disponible uniquement dans un format Word peu exportable.

Afin d'alimenter le catalogue en ligne de façon plus systématique, et grâce à sa participation au projet européen INTERA (Integrated European language data Repository Area)[4], un jeu de méta-données plus approfondi a pu être élaboré. La phase suivante pour ELRA était de pouvoir implémenter ces méta-données dans un catalogue dynamique.

[1] IMDI: http://www.mpi.nl/IMDI

[2] CHILDES: http://childes.psy.cmu.edu/data/

[3] ESF : http://www.mpi.nl/world/tg/lapp/esf/esf.html

[4] INTERA : http://www.mpi.nl/INTERA/

Finalement, le catalogue ELRA tel que disponible actuellement[1] au format PHP/MySQL a intégré l'ensemble du jeu de méta-données existant qui permet d'alimenter les informations sur les ressources de façon plus systématique et orientée-utilisateur. Les informations ayant été considérées comme importantes sont notamment :

- Identification des données,
- Description des données (par catégorie de données),
- Auteurs et éditeurs des données,
- Objectifs à l'origine de la creation des données et applications visées,
- Sources des données et méthodologie de production,
- Précision et niveau de confiance des données,
- Information sur la distribution.

Le besoin ressenti par les communautés OLAC et IMDI de rassembler des informations non seulement sur les ressources disponibles mais plus largement sur les ressources existantes a également été mis en évidence par ELRA. ELRA étant spécialisée dans l'identification de ressources linguistiques, sa tâche n'a pas été d'utiliser un outil pour rechercher les ressources existantes, mais plutôt de rechercher directement ces ressources et d'archiver les informations disponibles dans un supra-catalogue : le catalogue universel.

La figure ci-dessous montre de quelle manière le catalogue universel rassemble l'information sur les ressources existantes et leurs distributeurs et fournisseurs potentiels. Lesdits distributeurs et fournisseurs potentiels sont de différentes natures : ce sont des centres de données (BAS, LDC...), des projets et partenaires de consortiums (LC-STAR, OrienTel,...) ou encore des groupes de recherche produisant des ressources en-dehors du cadre figé de projets, etc.

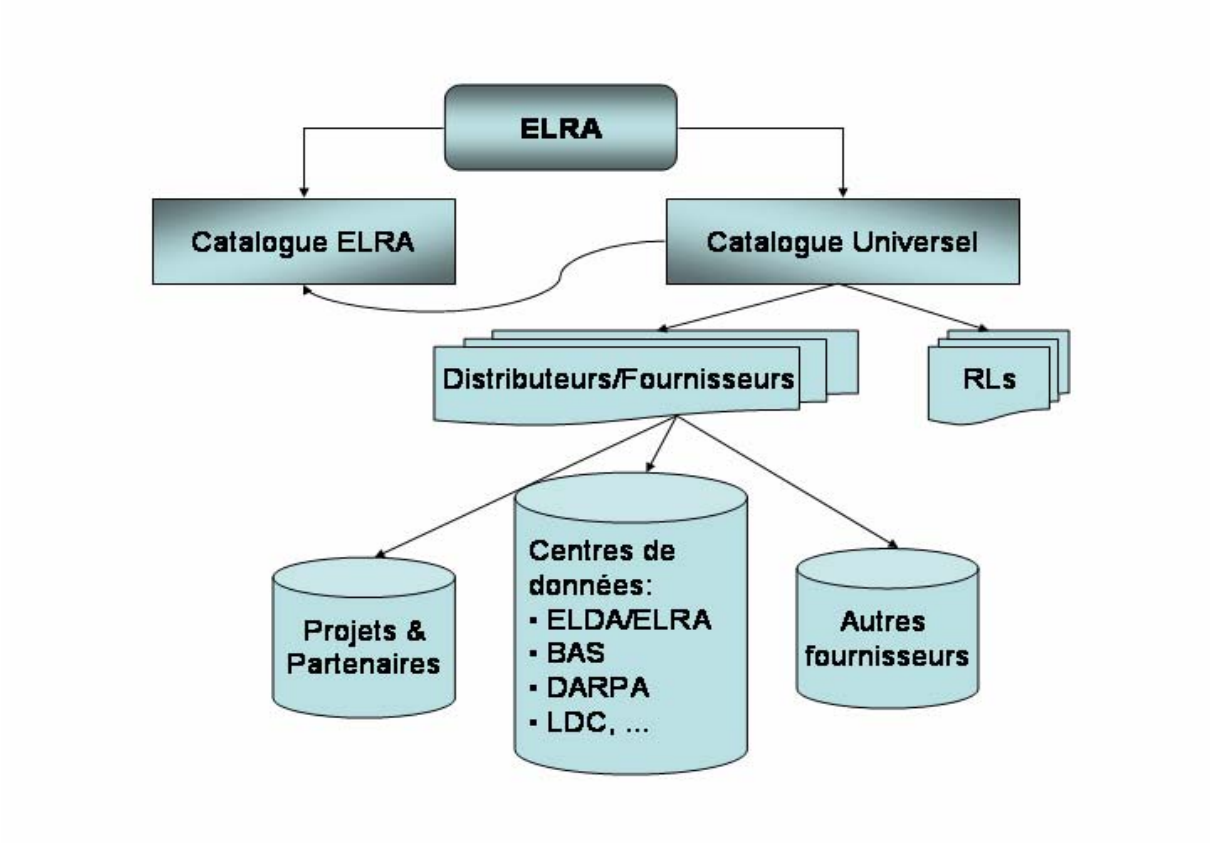

Figure 2 : Relation entre le catalogue ELRA et le catalogue universel

Le design du catalogue universel est basé sur celui du catalogue ELRA et utilise donc le même jeu de descripteurs que le catalogue des ressources disponibles.

[1] Catalogue de ressources linguistiques ELRA : http://catalogue.elra.info

2.2.4 Catalogue LDC

Comme ELRA, le LDC (Linguistic Data Consortium) s'est employé très vite depuis sa création en 1992 à distribuer ses ressources via un catalogue en ligne. Ce catalogue[1] utilise un grand nombre de méta-données standardisées qui sont également utilisées dans l'initiative OLAC (voir section 2.2.1).

L'usage de ce méta-données a permis notamment la mise en place d'un outil de recherche en ligne dans le catalogue LDC (voir figure ci-dessous).

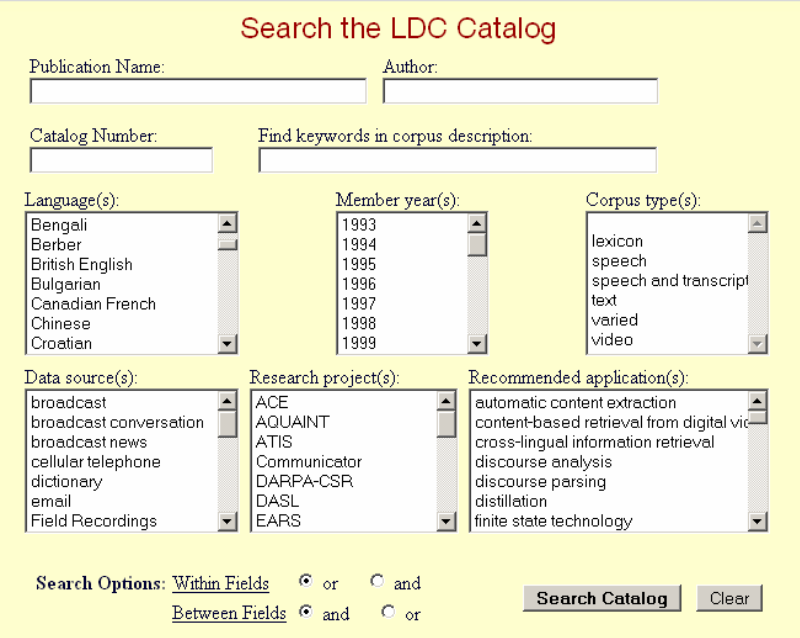

Figure 3 : Moteur de recherche en ligne du catalogue LDC

Le catalogue LDC est structuré par type de ressource. A chaque type de ressource est attribué un certain nombre de descripteurs (ou méta-données). La structure du catalogue peut être schématisé comme suit :

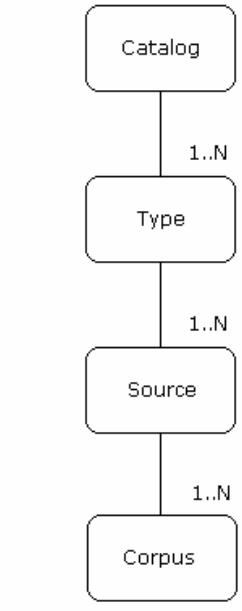

Figure 4 : Structure du catalogue LDC

Le jeu de méta-données LDC contient notamment la référence catalogue, le nom de la ressource, la source des données, le projet dans lequel a pu être produite la ressource, la ou les langues traitées, etc.

[1] Catalogue de ressources linguistiques LDC : http://www.ldc.upenn.edu/Catalog

3. Qualification juridique des ressources linguistiques et droits afférents

3.1 Introduction

L'objectif du guide est, sur la base d'un « état de l'art », de proposer aux producteurs de ressources linguistiques des recommandations techniques, juridiques et stratégiques permettant d'assurer la réutilisation des ressources.

L'idée de la nécessité d'une application aux ressources linguistiques d'un objectif cadre de « réutilisabilité » ('Reusability') s'est imposée récemment en tant que telle, à l'issue de l'Atelier de Grosseto de 1995. La mise en place du Linguistic Data Consortium (LDC, 1992) et de la European Language Resources Association (ELRA, 1995) répondait ainsi à une volonté de fournir à l'Industrie comme aux Académies des ressources partageables.

Le principe de réutilisabilité tend à ce que le développement et l'exploitation de ressources linguistiques ne soit plus seulement confinés dans le cadre privatif de multiples projets parallèles mais que soient au contraire assurées les conditions d'une mise à disposition de ces ressources à la Communauté. S'il a vocation à s'appliquer très généralement aux ressources exploitées et produites par la Recherche[1] et par le développement en Informatique[2], le principe de réutilisabilité présente un intérêt particulièrement manifeste dans un domaine où est nécessaire – notamment à raison de l'exploitation sur la base de traitements statistiques – la constitution de larges corpus dans un maximum de langues.

La poursuite de cet objectif comporte des implications techniques manifestes, notamment la définition de standards de compatibilité encadrant la définition et le développement de ressources linguistiques. La mise en oeuvre de ces standards n'assure néanmoins que des possibilités : l'effectivité de la mise en commun implique que soient écartés des schémas restrictifs ou exclusifs de distribution de nature à anéantir, de fait, les possibilités assurées au plan technique ; abordée au plan juridique, cette question de définition de cadres pour une large diffusion semble, en dépit de sa relative nouveauté, pouvoir s'appuyer sur les solutions expérimentées dans des domaines voisins. C'est dans une telle logique qu'il nous a été demandé d'orienter nos recherches notamment vers l'analyse comparée des Licences Libres et l'évaluation de la possibilité et de l'opportunité de leur application à la distribution de ressources linguistiques.

La présente étude comporte ainsi deux aspects : premièrement, une analyse du régime juridique, des contraintes et risques auxquels la création, l'exploitation et la distribution de ressources linguistiques sont soumis en application du droit commun ; deuxièmement, dans une optique prospective et finalisée, l'appréciation de la pertinence d'une application de licences libres, au regard des contraintes et risques juridiques issus du droit commun, d'une part, et de la poursuite d'un objectif de réutilisabilité des ressources linguistiques, d'autre part.

3.1.1 Notion(s) de « ressource linguistique »

Le droit commun ne consacre pas la notion de ressource linguistique par la détermination de règles spécifiques. Se pose ainsi une problématique essentielle de qualification juridique des ressources linguistiques. Une première approche d'une telle question est la recherche d'une définition unitaire et précise de la notion.

Les bases d'une définition doivent être recherchées dans le domaine de la pratique contractuelle des professions concernées ; les démarches communautaires pour la promotion de la diversité linguistique ont cependant été l'occasion d'un emploi de la notion, permettant de dessiner la perspective des autorités communautaires quant à la question.

[1] Lancé par Creative Commons, le projet Science Commons vise spécifiquement à lever les barrières entravant la mise en commun des ressources scientifiques et des fruits de la recherche. La démarche repose sur la combinaison de trois initiatives dont une vise expressément à assurer la « réutilisabilité » des ressources issues de la recherche scientifique : en premier lieu au plan technique, par un repérage et un marquage standardisé des données en vue de développer un « protocole de données en accès libre » (open-access data protocol) ; le dit protocole devrait constituer, dans un second temps, un fondement pour le développement d'instruments juridiques.

[2] En Science Informatique et en Ingénierie Logicielle, la réutilisabilité comporte notamment l'idée d'une possibilité de réutiliser un segment de code source en lui adjoignant de nouvelles fonctionnalités, au prix de modifications minimes ou inexistantes.

3.1.1.1 Pratique professionnelle et définitions dans le domaine contractuel

Le contrat utilisé dans les rapports entre le distributeur ELDA et ses fournisseurs définit la ressource linguistique comme « un ensemble de bases de données acoustiques orales et/ou textuelles et/ou graphiques, d'outils, d'algorithmes ou d'autres informations décrites en annexe A du présent contrat. » On remarque que cette définition retient une approche purement matérielle et énumérative : l'éclectisme des formes listées et surtout le caractère non exhaustif de la liste (la définition se terminant par la mention d'une notion, particulièrement générique et indéterminée d' "autres informations" – renvoyant à des précisions fournies, au cas par cas, en annexe du contrat) reflètent la difficulté d'une définition unitaire, au plan matériel, de la notion de ressource linguistique.

La licence libre LGPLLR, conçue spécifiquement pour régir la distribution de ressources linguistiques, définit ces dernières comme « une collection de données relatives à la langue, préparées de façon à être utilisées par des programmes d'application » : au contraire de la précédente, cette définition est très générique au plan matériel mais ajoute une considération fonctionnelle et finaliste (« préparées de façon à être utilisées par des programmes d'application »). Inclure un aspect fonctionnel dans la définition semble justifié par le terme de "ressource", impliquant un rapport passif à l'intervention de moyens extérieurs d'exploitation. La précision apportée par la définition LGPLLR est cependant probablement trop restrictive dans l'optique d'une définition générale des ressources linguistiques, ces dernières étant fréquemment, mais pas systématiquement développées en considération d'une exploitation logicielle : les ressources peuvent ainsi être utilisées, dans les milieux académiques, à des fins pédagogiques ou scientifiques n'impliquant pas nécessairement l'exploitation par des logiciels.

3.1.1.2 Perspective Communautaire sur la notion de « ressource linguistique »

Dans les textes de droit commun, le terme de ressource linguistique ne semble apparaître en tant que tel que dans une décision du Conseil de l'Union Européenne en date du 21 novembre 1996, **«** Concernant l'adoption d'un programme pluriannuel pour promouvoir la diversité linguistique de la Communauté dans la société de l'information » [1].

A titre essentiel, la décision manifeste, dans ses considérants, les objectifs de contribuer à lever les barrières linguistiques entravant – notamment pour les PME – l'accès au Marché Intérieur, de fournir aux citoyens un accès équitable à l'information, et de soutenir les langues en risque de marginalisation. Le Conseil estime, dans son considérant (7), « que la politique linguistique relève de la compétence des Etats membres, dans le respect du droit communautaire; que, cependant, la promotion du développement des outils modernes de traitement de la langue et de leur utilisation est un domaine d'activité où une action communautaire est nécessaire pour permettre la réalisation d'économies d'échelles substantielles et la cohésion entre les différentes zones linguistiques ».

L'objet de la décision est ainsi la mise en place d'un programme de soutien financier, d'une part, « à la création d'un cadre de services pour les ressources linguistiques » (ligne d'action 1), d'autre part à « l'utilisation de technologies, de ressources et de normes linguistiques et leur intégration dans des applications informatiques ». Dans une optique stratégique, les considérants de la décision visent un soutien tant aux développeurs de ressources, technologies et normes linguistiques qu'aux industries exploitant ces ressources pour l'offre de services multilingues et l'intégration dans des applications informatiques. Dans le cadre de la définition de la ligne d'action 1, semblant correspondre aux acteurs du domaine des ressources linguistiques stricto sensu, par opposition aux exploitants de ces ressources objets de la ligne d'action 2, il est énoncé que « les ressources linguistiques comme les dictionnaires, les banques de données terminologiques, les grammaires, les recueils de textes et d'enregistrements vocaux sont une matière première essentielle pour la recherche en linguistique, le développement d'outils de traitement de la langue intégrés dans des systèmes informatiques, l'apprentissage des langues et l'amélioration des services de traduction. »

Cette affirmation propose une association intéressante de listes typologiques des formes matérielles et finalités des ressources linguistiques. La liste des types matériels, moins extensive que celle proposée par ELDA, paraît cependant plutôt conservatrice et restrictive en considération de cette dernière.

[1] Décision du Conseil n°96/664/CE, 21 novembre 1996, JOCE 28 novembre 1996, vol. 306, p. 40)

Au-delà de ces éléments de définition, la décision manifeste une certaine prise en considération de questions ayant une portée relativement à la problématique de réutilisabilité des ressources linguistiques ; le considérant 1 énonce ainsi : « l'exploitation intégrale des ressources disponibles est actuellement entravée par le fait qu'elles sont principalement monolingues, souvent difficiles à localiser et que leurs spécifications de base sont parfois divergentes, ce qui limite leur utilisation plus large. » Selon l'article 2 de la décision les actions entreprises incluent « la stimulation de l'utilisation […] des normes linguistiques et leur intégration dans les applications informatiques », sans précision cependant sur la notion de norme linguistique. Les considérations plus précises sur l'établissement de normes se focalisent en réalité sur l'objectif de normes multilinguistiques des produits et prestations[1], notamment pour les industries des technologies de l'information et de la communication[2].

3.1.2 Problématiques et annonce de plan

Comme l'illustre la recherche d'une définition unitaire, la notion de ressource linguistique recouvre des réalités variées et complexes : la qualification juridique des ressources linguistiques et la détermination des éventuelles conditions de leur protection juridique impliquera une approche analytique (Section 2).

Les droits portant sur les ressources linguistiques ou leurs composants étant identifiés, la spécificité des modes d'exploitation des ressources linguistiques soulève en second lieu l'appréciation de l'exercice – et de la possible mise en cause – par les exploitants des droits portant sur les ressources linguistiques ou leurs composants (Section 3).

Indépendamment des questions liées à l'existence de droits sur les oeuvres exploitées, le contenu même des ressources est, en raison notamment de la diversité de ces dernières, de nature à présenter un caractère illicite et/ou attentatoire aux droits des personnes (Section 4).

Sur la base des contraintes et risques juridiques identifiés, d'une part, de l'objectif général de réutilisabilité des ressources linguistiques d'autre part, sera proposée une analyse des licences libres, de la possibilité, des apports et des limites de leur application à la distribution de ressources linguistiques (Section 5).

3.2 Qualification et protection juridiques des ressources linguistiques

La protection juridique des ressources linguistiques distribuées est dans les faits déterminée par les rapports contractuels – contrôle servi par la concentration du marché de la distribution des ressources en particulier. L'identification des risques juridiques dans l'exploitation, au sens large, de ressources linguistiques, ainsi que les perspectives d'une dissémination des rapports de distribution conduisent néanmoins à rechercher, au travers d'une qualification juridique des ressources linguistiques, les fondements de leur protection en droit commun.

Considérées en elles-mêmes et indépendamment de leurs supports matériels, les ressources linguistiques ont en commun de présenter la nature immatérielle de données. Ainsi, au-delà des droits portant potentiellement sur les supports matériels des ressources linguistiques, la possibilité d'une protection juridique de ces dernières dans le cadre de leur exploitation doit, compte tenu de leur nature immatérielle, être recherchée sur le terrain du Droit de la Propriété Intellectuelle.

Les ressources linguistiques se caractérisent, dans cette optique, par une double diversité : diversité de leurs natures et diversité de leurs rapports d'intégration.

3.2.1 Ressources linguistiques primaires

3.2.1.1 Le Corpus et les documents qu'il comporte

En amont des éventuels processus successifs de modification et/ou intégration conduisant à la production de ressources linguistiques complexes, les corpus sont constitués par des documents ou ensembles de documents caractérisés par le seul fait qu'ils comportent l'usage d'une langue déterminée.

[1] Cf ligne d'action 1 : « L'objectif de cette ligne d'action est de soutenir, pour toutes les langues européennes, la mise en place d'une infrastructure européenne de ressources multilingues et d'encourager la création de ressources linguistiques électroniques. »

[2] Considérant (16) : considérant qu'il convient d'encourager les industries des technologies de l'information et des communications à établir des normes qui prennent en compte la diversité linguistique et à les intégrer dans leurs produits et applications;

La seule condition de rattachement à cette catégorie étant l'emploi d'une langue déterminée, sont concernés des documents de toute nature (documents écrits, audio ou audiovisuels), manifestant tout mode d'expression dans la langue définie (documents d'expression verbale ou multimodale, à l'écrit ou à l'oral) et empruntant tous supports de distribution ou diffusion (presse, émissions radiophoniques, télévisuelles, cinématographiques …).

Le document peut être créé directement dans une optique linguistique, par exemple d'illustration d'un fait de langue ou d'une phonétisation ; dans de nombreux cas, cependant, des documents créés en considération d'un objet propre et extralinguistique (par exemple informatif ou artistique) connaîtront une seconde vie par une réutilisation en tant que ressources linguistiques ; les documents concernés sont ainsi de nature et origine diverses : journalistiques ou littéraires, radiophoniques ou télévisuels …

Selon l'article L.111-1 du Code de la Propriété Intellectuelle,

« L'auteur d'une oeuvre de l'esprit jouit sur cette oeuvre, du seul fait de sa création, d'un droit de propriété incorporelle exclusif et opposable à tous. »
« Ce droit comporte des attributs d'ordre intellectuel et moral ainsi que des attributs d'ordre patrimonial, qui sont déterminés par les livres Ier et III du présent code. »
Les conditions de protection par le droit d'auteur sont approchantes dans tous les pays. En droit français, la personne physique qui a créé une œuvre à caractère original bénéficie sur cette dernière, du seul de sa création, de droits patrimoniaux et moraux[1].

Selon une formule constante de la Cour de Cassation, évaluer l'originalité reviendra à rechercher dans l'oeuvre "l'empreinte de la personnalité" de son auteur. Dans les faits, cette appréciation se fondera largement sur la nouveauté de l'œuvre et sur « l'effort de création » ou « effort personnalisé ». La jurisprudence se contente en pratique d'une faible personnalisation ; d'après l'article L.111-1 CPI, la destination et le caractère méritoire de la création sont, par ailleurs, en principe sans portée.

L'article L112-1 du Code de la Propriété Intellectuelle dresse une liste (non limitative) des « œuvres de l'esprit » protégées. Cette liste confirme notamment que la protection s'étend à des œuvres à finalité artistique aussi bien qu'informative, sur tous médias et de longueurs diverses (l'article L112-1 1° CPI vise ainsi « les livres, les brochures et autres écrits »).

Les œuvres à objet artistique telles que les fictions littéraires, télévisuelles ou cinématographiques seront considérées, dans l'immense majorité des cas, comme portant l'empreinte de la personnalité de l'auteur. L'originalité d'un écrit scientifique – au sens d'empreinte de la personnalité de l'auteur – est possible mais plus difficile à déceler que celle d'une œuvre de fiction par exemple.

La qualification en document public écarte la protection : textes publics ou décisions de justice ne seront pas protégés par un monopole de l'auteur (L112-5-3° c)), au contraire des cours, discours et plaidoiries. Au regard d'oeuvres à but informatif tel que les oeuvres journalistiques, un document pourra être protégé à condition qu'il ne se résume pas à une information brute (ex : une dépêche « brute » de l'Agence France Presse) ou à une simple compilation d'informations. Prises en elles mêmes, les informations sont en principe librement exploitables et ne peuvent être protégées ; une protection peut en revanche être reconnue sur une matérialisation, une mise en forme donnée de ces informations – à condition que cette mise en forme présente le caractère d'originalité caractéristique des œuvres d'auteur. La compilation d'informations pourra ainsi être reconnue comme originale et protégeable dès lors que, par sa présentation, l'auteur lui aura donné une physionomie propre portant l'empreinte de sa personnalité.

Les corpus, qui ont fréquemment pour source des éditeurs, peuvent ainsi faire ou non l'objet de droits d'auteur, selon qu'ils présentent l'originalité requise ou qu'ils apparaissent au contraire comme purement informatifs ; les oeuvres d'auteur peuvent par ailleurs être tombées dans le domaine public (depuis 1997, l'article L. 123-1 CPI prévoit une échéance de la protection des droits patrimoniaux de l'auteur fixée à 70 ans après le décès de ce dernier). Au regard de l'existence potentielle de droits d'auteur de nature patrimoniale portant sur tout ou partie du corpus fourni, le fournisseur – habituellement éditeur – s'affirme normalement, en application des contrats types, titulaire ab initio ou cessionnaire des droits patrimoniaux considérés.

[1] Le système anglo-saxon du copyright, qui tend au contraire à protéger le producteur ayant financé l'œuvre, pose à la protection une condition additionnelle de dépôt de l'œuvre auprès d'un organisme agréé.

Les droits moraux que le droit français reconnaît à l'auteur sont perpétuels et incessibles (art. L.121-1 al.3 CPI) et présentent un caractère d'ordre public. Le droit de divulgation paraît sans portée réelle quant à l'utilisation d'une œuvre dans un corpus, l'œuvre faisant habituellement l'objet à ce titre d'une réutilisation, ou provenant d'une reproduction à l'occasion d'une diffusion publique (enregistrement de diffusions radiophoniques et télévisuelles notamment). Le droit de repentir et de retrait est d'utilisation très rare ; la nécessité d'une indemnisation d'un cessionnaire visant une exploitation de l'œuvre à titre de corpus – et non en tant que telle – paraît devoir être, au mieux, symbolique. Le droit à la paternité impliquera la mention systématique de l'auteur en association avec les reprises de son œuvre.

Le droit au respect de l'œuvre est il de nature à affecter l'utilisation d'une œuvre à titre de ressource linguistique ? Des modifications de l'œuvre, une reproduction ou représentation dans des conditions techniques altérant l'œuvre, une présentation tronquée et/ou hors contexte sont en principe de nature à porter atteinte au droit de l'auteur à un respect de son œuvre ; néanmoins, le droit moral de l'auteur à un respect de son œuvre est reconnu sous réserve d'abus, et on peut penser que l'exploitation de l'œuvre à titre de ressource linguistique comporte nécessairement l'affirmation d'une perspective et d'un contexte spécifiques ne portant pas les exploitants à percevoir l'œuvre en tant que telle. Les droits moraux de l'auteur ne semblent donc globalement pas de nature à affecter l'exploitation de l'œuvre dans une stricte fonction de ressource linguistique.

3.2.1.2 Bases de données

L'article 1.2° de la directive 96/9 du 11 mars 1996 donne de la base de données une définition reprise par la loi du 1er juillet 1998 et insérée à l'article L. 112-3 alinéa 2 du Code de la propriété intellectuelle : « on entend par base de données un recueil d'oeuvres, 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[1] ». Il est donc nécessaire que les données aient fait l'objet d'un classement et ne soient pas uniquement compilées. Des ressources linguistiques telles qu'un lexique terminologique ou phonologique devraient ainsi relever du droit applicable aux bases de données.

L'architecture de la base (principe d'organisation systématique des données), ses outils d'interrogation, l'éventuel logiciel d'exploitation et son interface graphique constituent les éléments structurels de la base ; ils sont protégés par droit d'auteur s'ils présentent l'originalité requise[2].

Les efforts et investissements nécessaires à la collecte, à la classification et à l'organisation des données de la base sont protégés par un droit spécifique, dit droit sui generis des producteurs de bases de données. Le domaine d'applicabilité de cette protection est déterminé par une notion fondamentale, représentée par l'adjectif « substantiel » : les données d'une base pourront être protégées par ce droit à la condition que leur collecte et organisation[3] aient nécessité un investissement « substantiel » en termes de ressources financières, matérielles ou humaines (art. L. 341-1 al. 1er du CPI) ; cette possibilité de protection étant retenue, pourra être sanctionnée l'extraction (reproduction) ou la réutilisation (diffusion) de la totalité ou d'une partie « substantielle » du contenu de la base (art. L.342-1) ; une extraction ou une réutilisation non substantielle est également interdite lorsqu'elle est répétée et systématique et vise à reconstruire la totalité ou une partie substantielle de la base (art. L.342-2). Le caractère substantiel est apprécié aux plans quantitatif et qualitatif, par les juges du fond.

A noter que dans l'optique de la sanction d'une reprise illicite des données, la nature de ces dernières est indifférente ; le fait qu'il s'agisse de données du domaine public et/ou de données rendues accessibles au public par le producteur de la base n'exclut pas la qualification en extraction et/ou utilisation illicite.

Présentant fréquemment un pur caractère informatif, les informations constituant le contenu d'une base de données relèveront classiquement du domaine public, de données exploitables par tous – sans préjudice des règles encadrant, le cas échéant, le traitement de données à caractère personnel ; rien ne s'oppose néanmoins à ce que ces données soient protégées par le droit d'auteur si, prises isolément et en elles mêmes, elles présentent l'originalité requise.

---

1 Concrètement, est requis un outil d'exploitation de la base tel qu'une grille de recherche, un lexique, un thésaurus

2 La possibilité d'une protection de la structure de la base de données sur le fondement du droit d'auteur est visée au considérant n°15 de la directive 96/9 ; elle n'apparaît pas explicitement dans la Loi du

3 Au contraire, les efforts relatifs à la production même des données de la base ne peuvent être pris en compte dans l'appréciation de l'investissement.

3.2.2 Ressources linguistiques dérivées

3.2.2.1 Les modifications apportées à la ressource primaire par le développeur de ressource linguistique

3.2.2.1.1 La création et la modification de bases de données

Comme vu supra pour les ressources linguistiques premières constituées par des bases de données, deux dimensions de ces dernières sont susceptibles d'être protégées sur deux fondements juridiques distincts : indépendamment du caractère plus ou moins automatique de la collecte, de l'identification et de l'annotation de données linguistiques en vue de leur inclusion au contenu d'une base de données, la définition en amont de l'architecture de la base, de la logique d'organisation des données, peut être protégée par le droit d'auteur si elle présente un caractère original et personnalisé ; d'autre part, le caractère substantiel de l'investissement humain, financier et matériel consenti pour la constitution du contenu de la base de données fonde, en application du droit sui generis des producteurs de bases de données, une action à l'encontre de quiconque extrait ou utilise indûment une partie substantielle des données de cette base.

Compte tenu de sa portée sur une exploitation industrielle des ressources linguistiques d'une part, des liens étroits entre spécifications des bases de données et des logiciels d'autre part, la protection du travail conceptuel et créatif sur l'architecture d'une base de données, si elle est actuellement assurée sur le même fondement du droit d'auteur, nous semble par ailleurs pouvoir relever de la problématique juridico-politique très controversée, au niveau européen, du possible octroi de brevets sur les logiciels.

3.2.2.1.2 Le commentaire linguistique de corpus

Même si une telle qualification ne sera sans doute pas automatique, des œuvres telles qu'un cours, une dissertation, un commentaire linguistiques semblent devoir être généralement reconnues comme des oeuvres de l'esprit protégées par un droit d'auteur. Dans une ressource linguistique dérivée composée d'une ressource première et du commentaire linguistique de celle-ci, les éventuels droits d'auteur sur la ressource première se combineront avec les éventuels droits d'auteur sur l'apport constitué par le commentaire ; dans un tel cas d'œuvre composite et de pluralité d'auteurs, les droits sur l'œuvre dérivée seront reconnus selon une logique distributive, le premier comme le second auteur restant titulaire des droits afférents à la partie dont il est l'auteur.

3.2.2.1.3 L'annotation linguistique de corpus

L'analyse linguistique d'un document peut cependant, d'autre part, se matérialiser par des annotations à caractère succinct et systématique, tendant au marquage, après identification, de l'appartenance à certaines classes de nature ou fonction linguistique : de telles annotations visent notamment la création de ressources pouvant être exploitées automatiquement par les moyens informatiques des Technologies de la Langue et constituent, à ce titre, une part essentielle des développements de ressources linguistiques modernes.

Il peut paraître inopportun, au regard de ce type d'apports systématiques et non complétés de commentaires personnalisés, de parler d'oeuvre de l'esprit portant « l'empreinte de la personnalité de l'auteur » et protégeable à ce titre par droit d'auteur.

3.2.2.2 Rapports d'intégration et titulaires des droits

Les développements précédents ont présenté les droits portant potentiellement sur les différentes ressources linguistiques et qui étaient susceptibles d'être mis en cause par leur utilisation, modification ou distribution. Au plan pratique, cette identification vise à permettre à la personne qui, non titulaire de ces droits, veut exploiter une ressource selon des modalités impliquant leur exercice, à solliciter du titulaire la cession ou la licence des droits requis.

Compte tenu de la pluralité de personnes intervenant ordinairement dans le développement d'une ressource linguistique, l'identification nécessaire du ou des titulaires de ces droits soulève deux problématiques connexes : l'identification du ou des titulaires des droits selon que la ressource considérée a été développée dans le cadre de rapports entre plusieurs intervenants ; la persistance de droits sur les ressources premières modifiées ou intégrées et leur mise en cause potentielle dans l'exploitation de la ressource dérivée.

Deux axes de reconnaissance de droits peuvent ainsi conjuguer leur complexité dans l'effort de détermination des titulaires et du régime d'exercice de leurs droits : la question de la reconnaissance de droits dans l'axe vertical de modifications successives d'une oeuvre se conjugue avec celle d'une attribution des droits, selon un axe horizontal, entre les différentes personnes ayant participé à la création d'apports et susceptibles à ce titre d'être reconnus titulaires de droits d'auteur.

A notre sens, la protection des efforts créatifs et investissements à l'œuvre dans la constitution de telles ressources doit être recherchée sur le fondement du droit applicable aux bases de données. La notion de « recueil d'oeuvres, de données ou d'autres éléments indépendants » dont dépend, selon l'article L.112-3 al. 2 CPI, la qualification en base de données, ne nous parait pas empêcher le rattachement de certaines annotations linguistiques linéaires à une protection en qualité de base de données – dès lors que divers documents sont assemblés et annotés selon une même méthodologie fondant des catégorisations. En particulier, la linéarité sous laquelle se présentent, formellement, les documents annotés, ne nous semble pas devoir écarter la qualification en base de données : si, en raison de leur linéarité, ces documents n'apparaissent pas de prime abord comme « disposés de manière systématique ou méthodique » au sens de l'article L.112-3 al.2 CPI, une telle organisation systématique et non linéaire est destinée à être obtenue automatiquement par l'application de programmes informatiques, sur la seule base des informations apportées par les annotations (metadata, descripteurs hypertexte) : ainsi, la présentation des annotations au sein du texte et sous une forme linéaire ne semble être qu'une forme possible de présentation du contenu d'une base de données.

A tout le moins, l'effort d'annotation semble nettement relever de l'investissement substantiel, en termes de ressources humaines, matérielles et financières, requis pour la protection du contenu d'une base de données sur le fondement du droit des producteurs de bases de données.

3.2.2.2.1 Qualification juridique de l'œuvre et existence de rapports entre les différents acteurs de la production

En droit français, le titulaire initial des droits d'auteur sur une œuvre individuelle est l'auteur lui-même, au sens de la personne physique qui a créé l'œuvre par ses efforts personnels[1]. Une clause explicite issue d'un contrat de travail ou d'un contrat de commande peut prévoir la cession instantanée et automatique des droits patrimoniaux de l'auteur. Expresse, la clause doit en outre porter sur une ou plusieurs oeuvres déterminées : la cession globale d'oeuvres futures est nulle en droit français.

L'exigence d'une clause explicite reflète le principe selon lequel l'existence d'un contrat de commande ou de travail n'emporte pas cession automatique des droits au commanditaire ou employeur – sous réserve de l'inscription de ces rapports contractuels dans le cadre de la création d'une « œuvre collective » (infra) : l'auteur reste titulaire des droits et le commanditaire ne peut exploiter l'oeuvre que conformément à la licence accordée par le contrat – l'autorisation de l'auteur devra être obtenue pour des exploitations ultérieures telles que des modifications et intégrations successives.

L'article L113-2 CPI vise trois cas d'œuvres manifestant une pluralité d'auteurs : l'œuvre de collaboration, l'œuvre composite et l'œuvre collective.

Les oeuvres des divers auteurs peuvent d'abord s'inscrire comme contributions personnelles destinées à se fondre à une oeuvre qu'une personne physique ou morale a initié et qu'elle édite, publie et divulgue sous sa direction et sous son nom. L'œuvre sera alors qualifiée de collective.

Le critère de fusion des contributions personnelles en un ensemble, d'application fortement subjective, n'apporte guère de prévisibilité et sécurité juridique[2] ; le critère déterminant, dans les faits, est l'application au processus de création d'une logique hiérarchique avec des instructions et orientations dictées du haut vers le bas[3]. Les créations du salarié peuvent relever ou non, selon les cas, de la qualification en "oeuvre collective" ; dans le cas négatif, l'existence de rapports salariaux est sans portée sur la titularité des droits.

[1] le système anglo-saxon du copyright tend au contraire à la protection de l'investisseur

[2] le caractère fusionnel est ainsi rejeté pour un film de cinéma, qualifiée d'œuvre de collaboration ; il est au contraire retenu pour un journal, même composé d'articles séparés et signés de leurs auteurs respectifs

[3] Le critère n'est cependant pas non plus dénué d'approximations et incertitudes : ainsi, dans les faits, un éditeur de livres impose couramment ses vues à des coauteurs, sans que l'œuvre résultante ne soit pourtant qualifiée d'œuvre collective - une raison en est probablement l'image traditionnelle de la création littéraire et artistique comme fondée sur la personnalité propre et l'indépendance d'esprit.

En dehors de tels rapports de subordination commune, l'œuvre seconde sera dite soit « de collaboration » soit « composite » selon que l'auteur de l'oeuvre première aura, ou non, participé aux modifications et à la création de la seconde. Pour une qualification en œuvre de collaboration, une participation active et créative semble requise, le simple exercice par l'auteur de l'œuvre première de son droit d'autoriser ou non les modifications de cette œuvre conduira ainsi normalement à la qualification en « œuvre composite ».

Ordinairement, le cas dans lequel une œuvre créée à des fins propres et non linguistiques est « recyclée » en corpus et intégrée en tout ou partie à une œuvre seconde à finalité linguistique semble devoir relever de la définition de l'oeuvre composite, en l'absence de collaboration de l'auteur de la première oeuvre à la création de la seconde. Si l'œuvre première a été créée à des fins linguistiques – dans le cas notamment d'une œuvre créée ab initio pour servir de corpus et dans celui d'une ressource linguistique dérivant déjà de ressources incorporées – semblent plausibles des qualifications en œuvre de collaboration ou en œuvre collective.

Outre l'identification des titulaires des droits, cette qualification de l'œuvre créée par plusieurs auteurs est d'importance au regard de la détermination du délai à l'échéance duquel l'œuvre considérée relèvera du domaine public et sera librement exploitable (dans le respect des droits moraux des auteurs). Le point de départ de ce délai dépend en effet de la qualification de l'œuvre : sous réserve d'une cession de ceux-ci, l'auteur est titulaire, sa vie durant, des droits protégeant l'oeuvre ; après son décès, les droits patrimoniaux et moraux sont transmis à ses héritiers, les premiers pour une durée de soixante-dix ans (délai à l'échéance duquel l'oeuvre tombe dans le domaine public), les seconds à titre perpétuel (à l'exception du droit de repentir) ; pour l'œuvre de collaboration, le point de départ du délai est le décès du dernier des collaborateurs ; une œuvre collective tombera en revanche bien plus tôt dans le domaine public car le délai court à compter du jour de divulgation de l'œuvre.

3.2.2.2.2 Intégration de ressource et persistance des droits

Une œuvre d'auteur comportant la modification ou intégration d'une oeuvre première sans la participation de l'auteur de cette dernière à la création de l'œuvre seconde est qualifiée d' « œuvre composite » par l'article L.113-2 du Code de la Propriété Intellectuelle ; l'article L.113-4 indique clairement, au regard des œuvres ainsi qualifiées, que les droits éventuels sur l'œuvre première persistent et doivent être respectés au travers des procès de modifications ou intégrations successives correspondant à la création et l'exploitation des ressources dérivant - de façon parfois très indirecte - de cette œuvre première.

Il sera donc nécessaire d'obtenir cession ou licence des droits relatifs à cette œuvre première pour toute utilisation, modification ou distribution de l'œuvre dérivée comportant la représentation ou reproduction d'éléments significatifs[1] de l'œuvre première, ainsi que pour la création de ressources ultérieurement dérivées - dans la mesure où subsiste dans l'œuvre obtenue des éléments significatifs de l'oeuvre première.

Au regard de l'intégration de tout ou partie de bases de données, ces règles issues du Droit d'Auteur sont applicables à la réutilisation d'éléments structurels originaux de la base, tels que son architecture, ses outils d'exploitation, son logiciel d'exploitation, l'interface graphique de ce dernier ; les documents constituant le contenu de la base sont par ailleurs susceptibles d'être protégés par des droits d'auteur à raison de leur originalité et de relever ainsi des règles vues aux paragraphe précédents.

En revanche, la reprise de contenu de la base de données relève du droit sui generis des producteurs de bases de données : la personne protégée sera non l'auteur mais le producteur de la base de données, au sens de la personne à qui peut être attribuée l'initiative et la supervision de la collecte et organisation des données ; le producteur ne bénéficiera pas de droits réels portant sur les données de la base et de nature à être opposés, comme des droits d'auteur, à toute représentation ou reproduction de ces données ; son action présente un caractère personnel et son bien-fondé est subordonné à la preuve du caractère substantiel tant de l'effort de développement de la base que de celui du volume des données extraites et utilisées.

Le fait que les parties de la ressource première se réduise éventuellement à une portion congrue à l'issue de modifications successives peut finir par écarter la mise en cause des droits sur cette ressource première : dans le cas d'une œuvre protégée par droit d'auteur, la brièveté des reprises peut relever du droit de citation ou de commentaire de l'œuvre ; dans le cas d'une base de données, la réduction du contenu réutilisé pourra rapidement aboutir à ce que l'extraction et réutilisation soit considérée comme ne portant pas sur une partie substantielle du contenu de la base de données initiale.

3.2.2.2.3 Régime de l'exercice des droits en cas de pluralité d'auteurs

[1] Sur la licéité de reprises sans autorisation de l'auteur, sur le fondement notamment de leur volume limité, voir les sections 3.3 et 3.4

A la qualification juridique de l'œuvre s'attache la détermination du régime d'éventuel exercice en commun de ces droits ; ce régime conditionne les modalités de prise de décision quant aux cessions et licences de droits nécessaires à l'utilisation, la modification et/ou la distribution de l'œuvre. En pratique, se pose essentiellement la question de la possibilité d'un interlocuteur unique.

Au regard du régime de l'oeuvre composite, l'auteur de l'œuvre dérivée est titulaire des droits sur cette œuvre dérivée. L'auteur de l'œuvre première reste cependant titulaire des droits sur cette dernière et son autorisation est nécessaire pour la représentation ou la reproduction de l'œuvre ou partie d'œuvre incorporée ; il conserve ainsi la possibilité d'exiger qu'une rémunération lui soit versée chaque fois que ces droits sont exercés.

Dans le cas d'une œuvre de collaboration, l'exercice des droits d'auteur est fondé sur l'application de principe de la règle contraignante des décisions prises à l'unanimité (article L113-3 CPI) ; l'article L113-3 al. 4 prévoit cependant que « Lorsque la participation de chacun des coauteurs relève de genres différents, chacun peut, sauf convention contraire, exploiter séparément sa contribution personnelle, sans toutefois porter préjudice à l'exploitation de l'oeuvre commune. » De plus, un recours au juge est possible sur le fondement de l'abus de droit quand l'un des coauteurs, en raison de son opposition manifeste, paralyse l'exploitation de l'œuvre.

L'article L.113-5 du CPI pose une présomption (simple) selon laquelle est propriétaire de l'œuvre collective et est investie des droits de l'auteur sur cette dernière la personne physique ou morale sous le nom de laquelle elle est divulguée. Prise isolément, la qualification d'une ressource en « oeuvre collective » assurera donc normalement à une personne unique la titularité des droits. Néanmoins, les incertitudes de la qualification en œuvre collective par les juges font qu'il sera prudent de prévoir, à titre supplétif, une clause de cession des droits qui seraient ainsi reconnus malgré la probabilité apparente de qualification en œuvre collective.

Il est important de comprendre que les qualifications d'œuvre et les reconnaissances de droits d'auteur ne sont pas exclusives les unes des autres et peuvent se cumuler.

Une œuvre complexe, résultant d'une succession d'intégrations et/ou de travaux collaboratifs ou collectifs, pourra combiner plusieurs qualifications. Par exemple, l'exploitation, à titre de corpus non annoté, d'une œuvre d'auteur pour la création d'une œuvre d'auteur dérivée à objet linguistique tendra à une première qualification de la seconde en « œuvre composite » - l'auteur de l'œuvre utilisée comme ressource première participant rarement à la création de l'œuvre linguistique dérivée ; prises isolément, œuvre première comme apports de l'œuvre dérivée peuvent être par ailleurs le fruit du travail de plusieurs auteurs et relever de qualifications cumulatives en œuvre de collaboration ou œuvre collective. L'intégration ultérieure de l'œuvre seconde à une œuvre troisième – par exemple un logiciel exploitant la ressource – pourra apporter un niveau additionnel de droits d'auteurs ...

Face à la multiplicité d'auteurs potentiellement reconnue, une logique distributive s'applique à la détermination des droits de chacun et des modalités d'exercice de ces droits. Au regard des droits patrimoniaux de l'auteur, des cessions de droits seront fortement indiquées, en pratique, pour éviter un enchevêtrement inextricable de droits ; en revanche, ces possibilités de cessions ne peuvent écarter la persistance des droits moraux que le droit français, particulièrement, reconnaît aux auteurs sur leurs œuvres respectives (ces droits moraux étant définis comme inaliénables et imprescriptibles) : la question des titulaires multiples de droits sur une œuvre complexe et du régime d'exercice en commun de ces droits est donc d'importance même dans le cas de cession systématique des droits patrimoniaux des auteurs.

3.3 Droits exercés dans l'exploitation d'une œuvre à titre de ressource linguistique

Dans la mesure où l'utilisation d'œuvres à titre de ressources linguistiques relève du régime juridique général applicable à l'exploitation d'œuvres d'auteur, elle est de nature à impliquer l'exercice de différents droits patrimoniaux de l'auteur – droits qui devront être acquis par cession ou licence ; dans cette perspective, les droits ainsi potentiellement requis dépendront des modalités d'exploitation de l'oeuvre (3.3.1). Cependant, la spécificité de l'exploitation d'œuvres à des fins linguistiques appelle à envisager la possibilité que cette dernière relève des limites reconnues, dans l'intérêt du public, au monopole de l'auteur (3.3.2).

3.3.1 Monopole d'exploitation de l'auteur et mise en cause de droits patrimoniaux

3.3.1.1 Définition des droits patrimoniaux reconnus à l'auteur

Les droits patrimoniaux reconnus à l'auteur ou aux auteurs de l'œuvre protégée sont le fondement de son exploitation ; l'exploitation commerciale de l'oeuvre implique normalement l'exercice de ces droits et est en principe réservée au titulaire des droits d'auteur, c'est à dire l'auteur lui même ou le cessionnaire éventuel de ces droits (les droits d'auteur d'ordre patrimonial peuvent faire l'objet d'une cession) ; l'exploitation de l'œuvre par une personne non titulaire des droits nécessite une licence délivrée par l'auteur, c'est-à-dire une autorisation expresse et préalable de mettre en œuvre les droits visés par la licence. Tout droit non expressément accordé par la licence ne peut être exercé.

En droit français, ces droits patrimoniaux sont traditionnellement définis comme constitués d'un droit de représentation et d'un droit de reproduction de l'oeuvre.

La représentation de l'œuvre est définie comme « la communication de l'œuvre au public par un procédé quelconque » (art L122-2 CPI). La reproduction est définie comme « la fixation matérielle de l'œuvre par tous procédés qui permettent de la communiquer au public d'une manière indirecte» (L122-3). Cette distinction classique, peu parlante – notamment au regard de la diffusion par Internet – tend à être remise en cause au plan international ; ainsi la directive communautaire 2001 sur le droit d'auteur, intégrant les deux traités OMPI de 1996, distingue : la reproduction même de l'œuvre ; la distribution de l'œuvre ainsi fixée matériellement ; la communication de l'œuvre au public (équivalant à la représentation en droit français).

Un droit de suite (art L 122-8 CPI) permet en outre à l'auteur de percevoir une rémunération, au-delà de la vente initiale du corpus, sur les représentations et distributions ultérieures de l'œuvre.

L'exercice et l'éventuelle mise en cause de ces droits, dans le cadre de l'exploitation de l'œuvre à titre de ressource linguistique, dépendent des modalités d'exploitation de l'œuvre et notamment du rapport entre cette exploitation et le public.

3.3.1.2 Notion de public(s)

Au regard de l'exploitation de ressources linguistiques, la notion de public soulève trois questions essentielles.

La confrontation des notions de ressource linguistique et de public soulève en premier lieu la question préalable et générale de l'existence d'un "public", au sens juridique du terme, pour des oeuvres exploitées à titre de ressource linguistique et non en tant que telles. L'existence d'un "public" ne semble pas pouvoir être globalement écartée en considération des finalités linguistiques de l'exploitation d'œuvres : pour l'identification de la communication d'une œuvre à un public, le critère retenu est celui, d'ordre matériel, de la possibilité offerte à un public de prendre connaissance de l'oeuvre ; le critère conduira à apprécier si une part substantielle de l'œuvre a été communiquée et ne s'attachera pas aux finalités de cette communication. En ce sens, la qualification de l'utilisateur final, notamment, en "public" ne semble pouvoir être globalement écartée que sur la base du caractère fragmentaire de la reprise de l'œuvre ou de son inaccessibilité en raison de son intégration au logiciel (infra, 3.3.3).

En second lieu, la dichotomie public / privé et la notion de sphère privée en particulier sont déterminantes pour l'identification de modes d'exploitation de l'œuvre ne comportant pas l'exercice des droits de reproduction et/ou de représentation : en limitation du monopole reconnu au titulaire des droits d'auteur, l'usage privé et la copie privée de l'œuvre ne nécessitent pas d'accord (art L122-5 1° et 2° CPI). La qualification de sphère privée paraît devoir être envisagée pour deux champs : premièrement, la chaîne des rapports tendant à l'intégration de l'oeuvre dans une ressource linguistique, avant diffusion ou distribution à l'utilisateur final ; ensuite, dans le cas où la première hypothèse serait écartée, la sphère constituée par les rapports "In House" au sein d'une entité unique.

Les schémas logiques fondant, dans le marché des ressources linguistiques, les contrats définissant les droits d'utilisation, de modification et/ou de distribution, comportent notamment une distinction entre fournisseurs, distributeurs et intégrateurs d'une part, "utilisateurs finaux" d'autre part. Y a-t-il lieu de penser que seule la communication à l'utilisateur final doit s'analyser comme portant l'oeuvre éventuellement protégée à la connaissance d'un public et comme exerçant à ce titre le droit de représentation et/ou de reproduction ? La chaîne des fournisseurs et développeurs de ressources linguistiques ne doit elle pas être considérée à cet égard comme présentant une forme d'intégration au sein de laquelle les rapports devraient être considérés comme privatifs – a fortiori dans le cadre d'une entité unique ?

Au plan fonctionnel, comme vu supra, seul semble importer le fait que des personnes aient pris connaissance de l'œuvre, indépendamment de finalités purement linguistiques et professionnelles ayant conduit cette découverte. La notion de sphère privée est quant à elle restrictive et ne couvre qu'une utilisation dans un « cercle de famille », à l'exclusion d'un cercle associatif ou des associés d'une société par exemple ; la copie privée se limite à un but non professionnel et non collectif : la copie à destination des associés d'une entreprise, comme dans le cas d'une réunion d'entreprise, ne sera pas couverte. La communication de l'œuvre, tant à un tiers, qu'à un collaborateur ou qu'au sein d'une société ou organisme, implique donc en principe l'autorisation préalable de l'auteur.

Au-delà de la question des possibilités éventuelles d'une exploitation sans autorisation préalable de l'auteur, la notion de public doit être précisée en ce qu'elle conditionne l'étendue des droits de communication exercés et, en conséquence, l'étendue de la licence nécessaire.

Il est important, à cet égard, de noter que la qualification de diffusion publique n'implique pas nécessairement une diffusion massive. Des groupes limités sont de nature à constituer un public au sens des articles L122-2 et L122-3 CPI ; la diffusion à tout groupe constitutif d'un nouveau public s'analyse ainsi comme un nouvel exercice du droit de représentation et nécessite à ce titre une autorisation distincte du titulaire des droits[1]. Ainsi, dans le cas d'une diffusion d'œuvres par un poste radiophonique ou télévisuel, le droit de représentation sera réputé exercé tant par l'émetteur initial (« chaînes de radio et télévision ») que par l'éventuel utilisateur d'un poste de réception audible et/ou visible d'un public local. De ce fait, une vigilance particulière devra être exercée au regard de la circonscription, dans la rédaction de la licence, du public visé.

3.3.1.3 Modalités de diffusion de la ressource et exercice des droits de reproduction et de représentation

Précisons en premier lieu que la notion de communication au public de la ressource implique la possibilité offerte au premier de prendre concrètement connaissance de la seconde, en tant que telle. En ce sens, l'intégration de l'œuvre à un logiciel exploitant celle-ci sans permettre à l'utilisateur d'en prendre directement et globalement connaissance ne conduit pas à une communication au public.

D'un point de vue fonctionnel, la communication de l'œuvre peut se concrétiser par : le fait de montrer l'œuvre à un tiers à partir de son support matériel original (représentation) ; le fait de fixer l'œuvre sur un support matériel : premier enregistrement ou duplication d'une oeuvre orale, musicale, visuelle, audiovisuelle ; copie sur support matériel de données digitales (gravure sur disque optique, enregistrement sur support de stockage, sur disque optique, sur mémoire de masse …) ; le fait de permettre un accès à l'œuvre sous forme dématérialisée, par Internet ou par un réseau local

La communication de l'œuvre sous forme dématérialisée, par réseau informatique tel qu'Internet, relève de l'exercice du droit de représentation par celui qui rend cette œuvre accessible par ce moyen ; il est à noter cependant que l'enregistrement de données communiquées par réseau relève de l'exercice du droit de copie de l'œuvre, même dans le cas d'un enregistrement automatique et temporaire destiné à la seule visualisation de contenu Internet.

Au regard de la reproduction et distribution de l'oeuvre, la théorie dite de l'épuisement des droits (directive du 11 mars 1996, article 5), tendant à garantir la libre circulation des marchandises dans le Marché Unique Européen, impose une distinction importante quant au support de diffusion de la base de données : dans le cas d'une diffusion de la base de données sur un support matériel tel qu'un CD-ROM, le titulaire des droits ne peut, une fois le bien mis sur le marché, les utiliser pour faire obstacle à la libre circulation de la base de données sur le support matériel en question. Au contraire, l'envoi de données par Internet étant considéré comme une prestation de service et non comme une commercialisation de bien, la règle de l'épuisement des droits ne s'applique pas en la matière : ainsi, « *chaque prestation en ligne est un acte qui devra être soumis à une autorisation pour autant que le droit d'auteur le prévoit* » (directive du 11 mars 1996, 33è considérant)

3.3.2 Limites du monopole de l'auteur et perspectives d'exploitation licite sans autorisation du titulaire des droits.

La possibilité la plus évidente d'exploitation de ressources sans mise en cause de droits patrimoniaux

[1] Arêt CNN du 6 avril 1994 (Arrêt CNN, Civ.1ère, 6 avril 1994, JCP 1994.II.22273, note Galloux), confirmé au niveau européen par l'arrêt CJCE 7 déc 2006, D2007, 1236, obs Edelman)

est naturellement leur appartenance au domaine public. L'exploitation d'une oeuvre protégée ne nécessitera cependant pas toujours l'autorisation préalable de l'auteur. En premier lieu, comme vu précédemment, les modalités de l'exploitation de l'œuvre peuvent ne pas porter atteinte aux droits patrimoniaux de l'auteur en ce qu'elles ne conduisent pas nécessairement à porter l'œuvre à la connaissance d'un public au sens juridique du terme.
Le caractère fragmentaire de la représentation ou reproduction de la ressource première est de nature à écarter la nécessité d'une autorisation. Ainsi, au regard de la reprise de contenu d'une base de données, le droit protégeant le producteur de la base de données n'est opposable que dans le cas de l'extraction ou de l'utilisation d'une partie « substantielle » des données contenues ; l'évaluation de ce caractère substantielle étant menée au plan quantitatif mais également qualitatif, le volume limité de l'extraction n'est pas de nature à assurer systématiquement l'immunité du repreneur. De plus, la protection sur le fondement du droit sui generis des producteurs de bases de données n'exclut pas une protection additionnelle d'éléments de la base pris isolément, sur le fondement du droit d'auteur, à raison de leur originalité.
S'agissant d'œuvres protégées par le droit d'auteur, signalons en premier lieu qu'une reprise limitée à des éléments linguistiques épars ne conduira pas habituellement à porter une partie significative de l'oeuvre à la connaissance d'un public : les mots ou groupes de mots pris isolément n'appartiennent pas à l'oeuvre mais au domaine public. Au-delà de cette hypothèse extrême, le droit d'auteur apporte des limites d'importance au monopole conféré à l'auteur, limites tendant à la possibilité de représenter ou reproduire l'œuvre en question – toujours partiellement, mais de façon significative voire substantielle. Cette possibilité, fondée sur des finalités notamment éducatives ou scientifiques, présente un intérêt majeur au regard des ressources d'une matière linguistique globalement ou fréquemment – selon les perspectives – caractérisée par des finalités scientifiques, éducatives ou informatives.

3.3.2.1 Question de la portée de l'article L.122-5-3° points a) et e) au regard de l'exploitation d'oeuvres à des fins linguistiques.

Selon l'article L122-5-3° points a) et e) du Code de la Propriété Intellectuelle, sont autorisées, sans accord de l'auteur mais sous réserve que soient indiqués clairement le nom de ce dernier et la source :

« a) Les analyses et courtes citations justifiées par le caractère critique, polémique, pédagogique, scientifique ou d'information de l'oeuvre à laquelle elles sont incorporées ; »
« e) La représentation ou la reproduction d'extraits d'oeuvres, sous réserve des oeuvres conçues à des fins pédagogiques […] et des oeuvres réalisées pour une édition numérique de l'écrit, à des fins exclusives d'illustration dans le cadre de l'enseignement et de la recherche, […] dès lors que le public auquel cette représentation ou cette reproduction est destinée est composé majoritairement d'élèves, d'étudiants, d'enseignants ou de chercheurs directement concernés, que l'utilisation de cette représentation ou cette reproduction ne donne lieu à aucune exploitation commerciale (…) »[1]

La question centrale soulevée par ces points a) et e) est d'évaluer dans quelle mesure l'exploitation d'une oeuvre à titre de ressource linguistique répond à des finalités correspondant à celles visées à ces alinéas.

En première approche, on est porté à répondre que les ressources linguistiques ont notamment vocation à être exploitées à des fins scientifiques (recherche fondamentale ou recherche & développement) fondant notamment le développement de ressources linguistiques exploitables commercialement et/ou pédagogiquement. Reconnaître ainsi, globalement et par principe, l'adéquation de l'exploitation linguistique aux finalités scientifiques ou pédagogiques visées soulève néanmoins la question du caractère omnivalent de la justification et de la portée dès lors illimitée, au regard des finalités, des dérogations prévue par l'article L.122-5-3° points a) et e). Une telle applicabilité illimitée serait liée au fait que les finalités linguistiques n'impliquent qu'une approche formelle et extrinsèque de l'oeuvre et ne sont ainsi pas conditionnées par son contenu sémantique et son intérêt propre, intrinsèque.[2]

[1] La limite visée au point e) est d'origine récente : elle a été introduite par la Loi du 1er août 2006 relative aux Droits d'Auteur et Droits Voisins dans la Société de l'Information (Loi DADVSI).

[2] L'analyse linguistique peut impliquer, dans une mesure limitée, la prise en compte du contenu sémantique du corpus et sa mise en relation avec sa dimension formelle ; les finalités artistiques de l'œuvre sont par ailleurs de nature à appeler une approche de l'œuvre à un plan formel, mais distinct dans l'ensemble de celui intéressant l'analyse linguistique ; une convergence paraît possible dans différents cas où l'analyse linguistique pourra appuyer l'analyse artistique de l'originalité des modes d'expression.

En ce sens, une lecture possible quoique restrictive des points a) et e) est qu'ils ne posent comme légitime la reprise d'une partie de l'œuvre qu'à raison de son contenu sémantique[1] ; étendre au contraire leur application à l'exploitation formelle, à titre linguistique, de l'oeuvre est possible mais signifie que, dès lors que les conditions des points a) et e) seraient par ailleurs remplies, toute œuvre pourrait être exploitée à titre linguistique sans licence du titulaire des droits : en pratique, cette règle tendrait à ce que l'exploitation à titre linguistique d'une œuvre, dans les conditions visées, ne donne pas en principe lieu à rémunération du titulaire des droits – sauf à ce que les excès manifestes de la pratique soient considérés comme portant atteinte l'exploitation normale de l'oeuvre ou aux intérêts légitimes de l'auteur[2] ; une telle exploitation à titre gratuit est exclue sur la base de l'article L.122-5 point e), la logique d'application de cette limite étant la conclusion de contrats entre universités et éditeurs de livres pour l'établissement d'une rémunération forfaitaire en compensation de l'exploitation visée.

Au-delà de cette problématique des finalités, les conditions fixées par les points a) et e) doivent elles être considérées comme restrictives au regard d'une exploitation de l'oeuvre à titre de ressource linguistique ?

Les points a) et e) se distinguent fondamentalement en ce que le premier vise une intégration de l'œuvre première à une œuvre dérivée alors que le second envisage l'exploitation d'un extrait de l'œuvre première en tant que telle. Dues à cette différence, les conditions nettement plus contraignantes du point e) – notamment l'exclusion de toute exploitation commerciale – limitent son intérêt pour les acteurs du marché des ressources linguistiques ; le point e) semble en revanche de nature à fonder la réutilisation de ressources linguistiques, sans nécessité de cession ou licence de droits d'utilisation, par les milieux Académiques.

L'exigence posée au point e) d'un public spécialement concerné soulève à cet égard une première question d'interprétation.

Selon une première lecture, plutôt restrictive, cette exigence se fonderait sur la considération du contenu sémantique de l'œuvre considérée et limiterait l'applicabilité du point e), au bénéfice d'enseignants, étudiants et chercheurs en linguistique, aux oeuvres ou parties d'oeuvres à objet spécifiquement linguistique telles que des rapports de recherche, dissertations ou commentaires séparables – par opposition aux ressources primaires d'objet propre. On peut penser, en première approche, que de telles ressources sont de nature à être fréquemment sinon systématiquement reconnues comme des oeuvres conçues, entre autres, à des fins pédagogiques, au sens de l'exclusion du point e) ; l'assimilation globale paraît cependant exclue, dans la mesure où la réalisation d'une œuvre pédagogique implique à titre essentiel une prise en compte de ses destinataires et l'application de méthodologies didactiques.

Selon une seconde lecture, plus extensive, le public visé pourrait être également considéré comme spécialement concerné à raison de la dimension formelle de l'œuvre considérée : compte tenu du cadre fixé (exploitation non commerciale, rémunération forfaitaire sur une base conventionnelle, finalités précises), une telle extension de l'applicabilité du point e) paraît légitime ; si une telle lecture est retenue, il semble néanmoins nécessaire que l'œuvre présente un intérêt linguistique particulier, l'intérêt linguistique courant étant d'une nature très peu limitative paraissant incompatible avec la volonté de restriction exprimée par l'exigence d'un public spécialement concerné.
Le domaine d'application du point a), à savoir l'intégration de partie de l'oeuvre protégée dans une œuvre seconde semble de nature à couvrir largement des ressources linguistiques présentant, comme on l'a souligné, de fréquents rapports d'intégration, notamment par l'apport de commentaires et annotations linguistiques à des corpus bruts. Si la notion d'analyse linguistique est admise, malgré son approche formelle, au titre des finalités visées par le point a), l'application du point a) posera en second lieu la question d'exigences vis à vis de l'oeuvre réalisant l'intégration. Le terme d' « œuvre » utilisé à cet égard par le point a), conjugué avec les finalités visées, tend à ce que l'œuvre seconde présente les caractères d'une œuvre de l'esprit protégée par le droit d'auteur ; néanmoins, rien ne semble imposer, en droit, que l'œuvre intégrante présente les caractères d'originalité ou de personnalisation nécessaires à la protection par le droit d'auteur : on peut ainsi penser que le point a) sera applicable non seulement aux dissertations et commentaires linguistiques présentant des apports créatifs personnels, mais également à des œuvres résultant de travaux plus systématiques tels que des annotations standardisées.

[1] Dns cet esprit, le « public spécialement concerné », visé au point e) comme le seul auquel peut être licitement destiné la représentation ou reproduction envisagée, serait de même déterminé à raison du contenu sémantique de l'œuvre.

[2] Une rémunération du titulaire des droits resterait bien entendu possible et pourrait être notamment justifiée par la prestation de service constituée par la fourniture de l'œuvre.

Au plan matériel, les points a) et e) exigent une reprise plus ou moins limitée de l'œuvre considérée, exigence qui apparaît cependant peu restrictive, notamment dans la perspective d'une exploitation linguistique : la notion de « citation » utilisée au point a) n'apparaît pas aussi limitative qu'on pourrait le penser, la reprise de 10 à 15 % d'un livre étant par exemple admise ; plus encore, le critère jurisprudentiel quant à la longueur licite de l' « analyse » est interprété par la jurisprudence de façon permissive : la reprise de l'oeuvre dans un tel cadre ne doit simplement pas dispenser le public d'aller à l'oeuvre première ... Probablement moins restrictive encore, la notion d'extrait utilisée par le point e) est fortement indéterminée et ne s'oppose guère, lue objectivement, qu'à celle de reprise intégrale.

La restriction matérielle quant à la part de l'œuvre reprise apparaît surtout largement indéterminée et fortement subjective ou relative : elle pourra probablement être appliquée de façon significativement plus restrictive si des conditions telles que celles de la finalité sont globalement estimées comme ne restreignant pas la libre exploitation d'œuvres à des fins linguistiques.

3.3.2.2 Perspectives de dépassement des exceptions catégorielles visées à l'article L.122-5-3°

Les limites posées au monopole de l'auteur restent présentées, en droit français, comme consistant en des exceptions énumérées et devant être lues de façon restrictive, à la différence du droit anglo-saxon consacrant, avec la notion de « fair use » (US) ou « fair dealing » (UK) une limite de portée générale au copyright. Cette lecture analytique et restrictive pourrait cependant être prochainement remise en cause suite à l'introduction en droit français du principe dit du triple test.

Consacré en droit communautaire[1] puis transposé en droit français, au niveau législatif[2], par la loi n°2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information (Loi DADVSI), le principe dit du « triple test » ou « test en trois étapes » implique que les limites prévues au monopole de l'auteur ne soient appliquées que dans la mesure où elles ne portent pas atteinte, par ailleurs, à l'exploitation normale de l'oeuvre ou aux intérêts légitimes de l'auteur.

Comme le souligne le Professeur Alleaume[3], la « propagation à tous les niveaux du fameux triple test, ou test des trois étapes » peut être considérée comme la véritable nouveauté de ces dernières années en matière de Droit d'Auteur : « apparu dans la Convention de Stockholm en 1967, [le triple test] est repris à l'article 13 des accords ADPIC, à l'article 10 du Traité de l'OMPI sur le droit d'auteur, à l'article 16 du Traité de l'OMPI sur les droits voisins, à l'article 5.5 de la directive du 22 mai 2001 sur la société de l'information et aujourd'hui dans le Code de la propriété intellectuelle français. »
L'appellation « triple test » implique que la considération de l'exploitation normale de l'œuvre et des intérêts légitimes de l'auteur (tests 2 et 3) s'ajoutent à la condition initiale de l'applicabilité d'une des exceptions catégorielles explicitement prévues (test 1). Néanmoins, on peut penser que l'apport est de nature à révolutionner la lecture des exceptions au droit d'auteur[4], en ce que les deux considérations ajoutées seraient en elles mêmes suffisantes et pourraient prétendre à devenir l' « exception universelle » du droit d'auteur[5] : dans cet esprit, le législateur pourrait franchir un pas final et affranchir le juge du pilier des exceptions analytiques et catégorielle prévues, pour tendre à une appréciation plus globale du caractère équitable d'une exploitation sans licence, dans la lignée des solutions appliquées en droit anglo-saxon du Copyright.[6]

Au regard de l'exploitation d'œuvres à des fins linguistiques, l'appréciation globale du caractère équitable de l'exploitation permettrait une libre définition jurisprudentielle des régimes et lèverait la question d'un rattachement plus ou moins artificiel et/ou partiel des finalités linguistiques aux finalités pédagogiques, scientifiques et informatives visées à l'article L.122-5-3° ; l'esprit libéral de l'approche devrait tendre à une couverture large ou complète de l'exploitation d'œuvres à des fins authentiquement linguistiques. Le prix en serait – dans un premier temps au moins – une certaine insécurité juridique.

---

1 Article 5-5 de la Directive 2001/29/CE du 22 mai 2001

2 Le principe est introduit au sein de l'article L.122-5-3° du CPI ; il avait été, peu avant, consacré en jurisprudence par l'arrêt C. Cass. « Mulholland Drive » du 28 février 2006

3 C. ALLEAUME, « les nouvelles exceptions du droit d'auteur », LPA, 06 décembre 2007 n° 244, P. 46

4 C. Geiger, La transposition du test des trois étapes en droit français, D., 2006, dossier, p. 2164. ; Ch. Caron, Droit d'auteur et droits voisins, Litec, 2006, n°351

5 P.-Y. Gautier, Propriété littéraire et artistique, PUF, collection droit fondamental, 5e éd., 2004, no 203.

6 L'exception générale tirée de la notion de « fair use » ou « fair dealing » est évaluée sur la base de quatre critères : 1°) les but et caractère de l'usage (fins commerciales ou non lucratives ? fins éducatives ?) ; 2°) la nature de l'oeuvre ; 3°) le volume et l'importance de la partie utilisée par rapport à l'ensemble de l'oeuvre ; 4°) l'incidence de l'usage sur le marché potentiel de l'oeuvre protégée ou sur sa valeur.

3.3.3 Mise en cause de droits moraux de l'auteur sur l'oeuvre

L'exploitation d'une œuvre protégée par droit d'auteur sera habituellement conditionnée par la cession ou licence des droits patrimoniaux de représentation et de reproduction, comme vu précédemment ; certains droits moraux reconnus à l'auteur sur son œuvre sont cependant de nature à affecter l'exploitation, particulièrement en droit français – ce dernier se distinguant par l'intensité et l'étendue du domaine d'application des droits moraux de l'auteur sur son œuvre[1]. La prise en compte de ces droits moraux peut comporter des aspects contraignants, dans la mesure où, d'une part, les droits et obligations considérés s'appliquent à l'égard de toute œuvre ou partie d'œuvre exploitée, indépendamment du nombre des intégrations, de leur imbrication éventuelle et du nombre global des auteurs. La spécificité des droits moraux, à cet égard, provient de leur attribution en propre aux personnes physiques ayant créé l'oeuvre et de leur nature perpétuelle et inaliénable : les obligations liées aux droits moraux s'appliquent donc indépendamment de la cession éventuelle de ces droits ou de l'éventuelle qualification de l'œuvre en « oeuvre collective » : obligation d'identification qui peut impliquer un suivi méticuleux ou, à défaut, un travail non négligeable de recherche.

Le droit de l'auteur à la paternité de l'œuvre implique la mention systématique de son nom et de sa qualité d'auteur au regard de toute œuvre ou partie d'œuvre protégée qui serait exploitée ou intégrée. Si cette obligation tend à garantir l'identification nécessaire au respect des autres composantes du droit d'auteur, elle peut soulever des contraintes notamment au plan technique, dans le cas où il n'est pas possible ou pratique d'insérer ces informations dans l'œuvre concernée, en raison de la nature de cette dernière (ex : documents audio ou audiovisuels) ; dans ce cas, sera requise la fourniture systématique d'un document accessoire comportant l'identification du ou des auteurs.

Le droit moral de divulgation implique que seul l'auteur peut prendre la décision initiale de porter une œuvre à la connaissance du public ; ce droit s'épuise dans la première divulgation de l'œuvre – sous réserve de l'hypothèse, largement théorique, d'une décision de l'auteur de retirer l'œuvre divulguée du marché. Le droit de repentir et de retrait accordé, en droit, à l'auteur lui permet théoriquement, après avoir divulgué l'œuvre ou même cédé ses droits à l'exploitant, d'exiger des modifications substantielles sur son œuvre en cours d'exploitation ou le retrait de l'œuvre et l'arrêt pur et simple de cette exploitation ; ce droit, dont l'exercice implique une indemnisation du cessionnaire des droits, est dans les faits très peu exercé.

Dans la mesure où cette exploitation comporte la divulgation à un public au sens vu précédemment, l'exploitant d'une ressource linguistique comportant une œuvre originale devrait à ce titre s'assurer systématiquement que cette dernière a déjà fait l'objet d'une publication autorisée par l'auteur, ou, à défaut, que l'auteur donne expressément son autorisation à une publication dans le cadre de l'exploitation à titre de ressource linguistique.

Le droit de l'auteur au respect de son œuvre lui permet de s'opposer à toute modification, adjonction, suppression, utilisation dans un contexte inapproprié. L'utilisation dans un contexte inadapté paraît globalement exclue dans le cadre d'une exploitation à des fins linguistiques ; plus problématique semble être la question de la modification de l'œuvre, la création de ressources linguistiques étant de nature à conduire à des altérations considérables.

En fait, le danger paraît pouvoir être considéré comme globalement écarté dans l'optique d'une exploitation à titre de ressource linguistique : le principe du respect de l'esprit de l'œuvre, permettant à l'auteur de s'opposer à une représentation altérée, ne semble devoir s'appliquer que dans le cadre d'une exploitation de l'œuvre en tant que telle, de sorte qu'une exploitation linguistique semble devoir écarter systématiquement la difficulté.

A supposer que cette justification générale ne soit pas retenue, quelles garanties juridiques pourraient être constituées ?

La licence expressément donnée par l'auteur pour une exploitation de l'œuvre à titre de ressource linguistique semble devoir impliquer son autorisation, au plan moral, des formes de présentation de l'œuvre qui en découleront. Par mesure de sécurité juridique, il serait cependant probablement préférable de garantir, dans le cadre contractuel, la conscience de l'auteur des formes prévisibles de représentation et reproduction de l'œuvre impliquées par l'exploitation à titre linguistique à laquelle il donne autorisation. Quoique le droit au respect soit en principe inaliénable et qu'une telle clause contractuelle ne puisse théoriquement pas écarter de façon certaine son exercice, il est admis en jurisprudence que l'exercice de ce droit est facilement susceptible d'abus – abus qui semble pouvoir être facilement établi dans le cas de la remise en cause d'un consentement préalable, exprès et spécifique.

[1] A l'inverse, le droit moral n'existait pas même en droit américain du copyright, ce dernier tendant à une perspective patrimoniale et à la protection de l'investisseur plutôt que de la personne physique ayant créé l'oeuvre ; suite à la ratification de la Convention de Berne […] , une loi fédérale du 1er décembre 1990 a introduit un système complet de droit moral, qui reste cependant d'intensité moindre et de domaine d'application plus réduit qu'en droit français.

Soulignons par ailleurs un point sur lequel il importera d'être vigilant : le droit au respect de l'œuvre est un droit incessible, appartenant en propre à la personne physique ayant créé l'œuvre par ses efforts personnels ; or, dans nombre de cas, la licence accordée pour le recyclage d'œuvres à titre de ressource linguistique sera accordée de façon globale, sur nombre de documents, par une personne n'étant pas l'auteur lui même mais le cessionnaire des droits patrimoniaux sur ces oeuvres. Dans ces conditions, la licence, même spécifiquement accordée pour une exploitation à titre de ressource linguistique, n'est pas de nature à engager l'auteur ou à présumer de l'exercice de son droit de supervision du respect de l'œuvre. Idéalement, le consentement de l'auteur en personne devrait sans doute être à ce titre systématiquement recherché.

Au regard de ressources linguistiques constituées, par définition, par des documents d'auteurs de toutes origines, il est important de souligner que le droit français est applicable à la protection du droit moral de l'auteur étranger : dans le célèbre arrêt « Huston »[1], la Cour de Cassation a affirmé que les dispositions protégeant le droit de l'auteur à la paternité de l'œuvre et au respect de son intégrité (article L11-4 CPI) étaient « des lois d'application impérative » dans tous les cas : se trouve ainsi garanti un seuil minimal de protection des droits de l'auteur étranger en France.

3.4 Protection des données à caractère personnel

3.4.1 Applicabilité aux ressources linguistiques du régime d'encadrement du traitement de données à caractère personnel

3.4.1.1 Domaine d'application : notions de « donnée personnelle » et de « traitement »

La production et l'exploitation de ressources linguistiques est de nature à occasionner le traitement de données à caractère personnel relatives au(x) locuteur(s) : d'une part, en application d'une volonté de repérage et de rattachement de manifestations linguistiques à des données contextuelles ou personnelles[2] ; d'autre part, en raison du contenu naturel des documents de multiples natures utilisables à titre de corpus.

En application de la loi du 6 janvier 1978 modifiée par la loi du 6 août 2004, certains traitements de données à caractère personnel sont encadrés par un régime juridique strict visant la protection de ces dernières ; sont concernés, aux termes de l'article 2 de la loi du 6 janvier 1978, les traitements automatiques et les traitements non automatiques mais qui donnent lieu à la constitution de fichiers.

Pour la suite de ces développements, la notion de traitement de données à caractère personnel sera entendue comme impliquant, dans le sens d'une applicabilité de la loi, un traitement automatisé ou un traitement manuel entraînant la constitution de fichiers.

La définition légale des données à caractère personnel justifiant l'application du régime est extensive et couvre toutes les données permettant directement ou indirectement d'identifier la personne concernée. Selon l'interprétation donnée par la CNIL, doivent, par extension, être considérées comme données personnelles toutes les données dont le recoupement permet l'identification d'une personne (par ex : une date de naissance associée à une commune de résidence). Sont couverts des documents de nature et supports multiples – tels, par exemple, que des documents visuels permettant de reconnaître la personne.

3.4.1.2 Absence d'opposition de principe à la réutilisation de ressources à des fins linguistiques.

L'article 6 de la loi de 1978 vise fondamentalement à assurer une stricte concordance entre, d'une part, les opérations ou séries d'opérations dont les données à caractère personnel font l'objet et, d'autre part, les finalités légitimes justifiant ces traitements.

Au regard du principe même d'un traitement ultérieur, pour la constitution de ressources linguistiques, de données personnelles initialement collectées et traitées à d'autres fins, deux points sont à souligner : premièrement, en application d'un principe de sectorisation, la diffusion des données personnelles est normalement limitée au secteur d'activité concerné ; deuxièmement, le point 2°) de l'article 6 prévoit que les

[1] Cass. 1re civ., 28 mai 1991, *Turner Entertainment Co (Sté) c/Huston (Cts)* ; *D.* 1993, jurisp., p. 197, note J. Raynard.

[2] Ainsi, par exemple, la description des ressources Globalphone comporte des informations relatives aux locuteurs, répondant parfois à des spécifications avancées : age, sexe, origine, mais aussi profession, niveau d'études …

données ne peuvent en principe faire l'objet d'un traitement ultérieur « incompatible » avec les finalités déterminées, explicites et légitimes ayant fondé la collecte de ces données.

La notion utilisée de « compatibilité » ne semble pas d'un caractère restrictif de nature à exclure de façon générale une réutilisation des données dans une perspective d'étude linguistique. Le point 2°) précise d'ailleurs expressément, en ce sens, qu'est considéré comme compatible avec les finalités initiales le traitement qui : d'une part, poursuit des finalités statistiques ou de recherche scientifique ou historique – finalités auxquelles les études linguistiques semblent pouvoir se rattacher ; d'autre part, respecte le régime juridique applicable et n'est notamment pas utilisé pour prendre des décisions au regard des personnes concernées.

En compensation – et confirmation – d'une telle liberté globale de réutilisation à des fins linguistiques, s'appliquera au traitement en résultant le strict régime prévu pour la protection des données à caractère personnel.

3.4.1.3 Question de la pertinence et du caractère non excessif du traitement de données à caractère personnel dans des buts linguistiques.

La finalité et le mode de traitement des données à caractère personnel doivent respecter l'article 6 de la loi, selon lequel « les données traitées doivent être adéquates, pertinentes et non excessives au regard des finalités pour lesquelles elles sont collectées » ; sont appréciées à ce titre les possibilités alternatives de poursuivre les mêmes fins.

Cette exigence d'un caractère adéquat, pertinent et non excessif des données apparaît plus contraignante que la notion vue précédemment de « compatibilité » du traitement linguistique avec les finalités initiales du traitement.

Quoique la finalité linguistique ne soit pas préjudiciable en soi et que des données à caractère sensible puissent présenter un intérêt linguistique, l'exploitation de telles données à des fins linguistiques semble devoir manifester en principe une disproportion entre le caractère sensible de la donnée et son intérêt linguistique – sauf, potentiellement, dans le cas particulier de ressources linguistiques constituées spécifiquement pour couvrir l'expression des appartenances et tensions associées.

3.4.1.4 Opposabilité des droits de la personnalité à l'exploitation de données à caractère personnel dans le cadre d'une ressource linguistique.

En application de la Loi du 6 janvier 1978 modifiée par la Loi du 6 août 2004, la collecte et le traitement de données à caractère personnel doit respecter quatre droits des personnes concernées : un droit d'information sur les conditions de la collecte et ses finalités (art. 32), un droit d'opposition pour raison légitime (art. 38), un droit d'accès aux données collectées (art. 39) et un droit de rectification (art. 40) de ces dernières.

L'affirmation d'un droit d'opposition « pour raison légitime » soulève une interrogation quant à la consistance et la porté de cette dernière notion. Compte tenu du caractère a priori non préjudiciable de l'exploitation orientée vers la linguistique, la restriction semble écarter globalement une opposition efficace de la personne concernée à cette exploitation ; l'opposition ne semble logiquement devoir apparaître « légitime » que dans le cas d'un dévoiement du traitement à finalité linguistique, ce dernier couvrant alors des finalités non compatibles.

On note cependant, dans les faits, que le droit d'opposition énoncé par l'article 38 est reconnu largement, sans que la « raison légitime » visée n'impose en pratique de restrictions – la démarche même d'une opposition reflétant des motifs personnels déterminants. En pratique, l'opposition ne concernant que les données relatives à la personne considérée, pourra être traitée simplement par la suppression des données litigieuses.

3.4.2 Obligations formelles vis-à-vis de la CNIL

3.4.2.1 Obligation générale de déclaration préalable à la CNIL

La déclaration préalable à la CNIL, avant sa mise en place, d'un traitement automatisé[1] de données à caractère personnel peut être qualifié en quelque sorte de « régime de droit commun » applicable en la matière.

[1] Les traitements non automatisés ou manuels ne sont pas en principe soumis à l'obligation de déclaration préalable ; ceux de ces traitements qui concernent des données personnelles à caractère « sensible » (art. 25, I de la loi de 1978) sont cependant soumis à autorisation préalable.

En effet, tous les traitements, qui ne font pas partie des listes de traitements pour lesquels la loi de 1978 prévoit spécialement l'application du régime d'autorisation préalable (art. 25 à 27), doivent être considérés comme étant soumis au régime de la déclaration préalable.

La déclaration doit être effectuée par le responsable du traitement[1], avant la mise en place de ce dernier[2] ; elle doit notamment indiquer les finalités et modalités du traitement et engager le responsable quant à la conformité du traitement à la loi de 1978. (art. 23 et 30).

Contraignante, l'application effective du régime de déclaration préalable est de fait devenue résiduelle grâce aux possibilités de dérogations offertes pour simplification.

Premièrement, les traitements peuvent être dispensés de déclaration, par le législateur ou par la CNIL. Au niveau légal, sont notamment dispensés de déclaration préalable les traitements non automatisés. Deuxièmement, la CNIL a le pouvoir de décider, sur délibération, de dispenser de ces formalités les « catégories les plus courantes de traitement de données à caractère personnel dont la mise en œuvre n'est pas susceptible de porter atteinte à la vie privée ou aux libertés » (art. 24, II) ; au regard de ces mêmes catégories, la CNIL peut alternativement établir des normes simplifiées de déclaration : si le traitement envisagé relève de l'une des ces normes simplifiées, le responsable du traitement aura seulement l'obligation d'envoyer à la CNIL une déclaration de conformité à la norme simplifiée correspondante.

La CNIL met à la disposition du public la liste des traitements automatisés ayant fait l'objet d'une des formalités prévues par les art. 23 à 27 (déclaration, déclaration de conformité à une norme simplifiée, autorisation, déclaration de conformité à une autorisation ou un acte réglementaire unique). Aucun traitement visé n'est rattachable à la logique de création de données linguistiques : une délibération n°2005-284 du 22 novembre 2005 dispense de déclaration des sites webs diffusant ou collectant des données à caractère personnel, sites mis en oeuvre par des particuliers dans le cadre d'une activité exclusivement personnelle : une telle logique restrictive rend improbable l'octroi d'une dispense de déclaration pour des traitements de ressources linguistiques comportant des données personnelles.

D'autre part, les formalités de déclaration préalables peuvent[3] être remplacées par la désignation d'un « correspondant à la protection des données », chargé de superviser de façon indépendante le traitement des données à caractère personnel ; le correspondant à la protection des données est désigné par le responsable du traitement, en considération notamment de ses garanties d'appréciation neutre et indépendante.

3.4.2.2 Régime d'autorisation préalable

L'article 25, I de la loi impose l'obtention d'une autorisation préalable de la CNIL[4] pour le traitement de certaines données personnelles porteuses de risques de discriminations illégitimes[5].

Aux termes de l'article 8, I de la Loi de 1978, « Il est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci. ». Le point III de l'article 8 prévoit cependant que la CNIL « peut autoriser, compte tenu de leur finalité, certaines catégories de traitements », si les données à caractère personnel sont appelées à faire l'objet, à brève échéance, d'un procédé d'anonymisation reconnu conforme aux exigences de la CNIL. Cette autorisation est accordée selon les modalités prévues à l'article 25.

Le régime d'autorisation préalable s'étend en second lieu à des données personnelles du domaine financier ou pénal, susceptibles de conduire à la constitution de « listes noires » ou listes d'exclusion. Aux informations personnelles à caractère financier s'applique en outre un régime strict[6] de sectorisation dans le traitement des

[1] Le responsable juridique de l'organisme qui décide de créer le fichier ou la personne qui a délégation de signature

[2] Les déclarations et les demandes d'autorisation doivent être adressées à la CNIL : 1° Soit par lettre remise contre signature; 2° Soit par remise au secrétariat de la CNIL contre reçu ; 3° Soit par voie électronique, avec accusé de réception qui peut être adressé par la même voie. (art. 8 du décret n° 2005-1309 du 20 octobre 2005)

[3] sauf dans le cas d'un transfert des données vers un Etat non membre de la Communauté Européenne

[4] La demande d'autorisation est formée par le responsable du traitement et transmises selon les modalités vues supra ; la CNIL dispose d'un délai de 2 mois pour se prononcer.

[5] Nous viserons ici les catégories plus spécialement susceptibles d'apparaître dans des ressources linguistiques.

[6] A la différence du principe général de sectorisation posé par l'article 6, qui exige seulement que les données collectées ne soient pas traitées ultérieurement « de manière incompatible » avec les finalités initiales de la collecte.

données de nature à exclure par principe la réutilisation à des buts linguistiques de ressources faisant apparaître de telles données.

La prise en compte des nouvelles technologies se manifeste par l'inclusion dans le domaine d'application de l'article 25 des données génétiques et biométriques. Au regard des évolutions technologiques, la CNIL[1] manifeste un souci particulier de prise en compte des risques soulevés par la convergence des technologies de la biométrie, de la vidéosurveillance et de la géo-localisation et par le développement accéléré du recours à la biométrie en particulier[2].

Selon le point 5° de l'article 25, une autorisation préalable de la CNIL est encore nécessaire avant la mise en place de tout traitement automatisé ayant pour objet l'interconnexion de fichiers relevant de personnes autres que des personnes morales gérant un service public et dont les finalités sont différentes. Le développement de ressources linguistiques comportant le croisement de bases de données issues de différentes sources impliquera une vigilance particulière à l'égard de ce point.

3.4.2.3 Régime de déclaration préalable

Par dérogation, un régime de déclaration préalable peut s'appliquer dans certains cas relevant en principe de l'application du régime d'autorisation préalable.

Au regard des données personnelles « à caractère sensible » visées à l'article 8, I le point II de l'article prévoit qu'il est fait exception à l'autorisation préalable vis à vis de certains traitements : (1°) les traitement pour lesquels la personne concernée a donné son consentement exprès; (4°) le traitement de données à caractère personnel rendues publiques par la personne concernée.

Les ressources linguistiques premières étant fréquemment constituées de la réutilisation de documents informatifs ou artistiques ayant fait l'objet d'une publication, le point 4°) présente potentiellement un réel intérêt pratique pour le développeur de ressources linguistiques ; il soulève cependant la difficulté considérable, en l'absence de rapport direct avec la personne concernée, de déterminer si cette publication est précisément le fait de cette dernière.

Plus généralement, un régime de déclaration préalable s'appliquera par dérogation dans le cas éventuel où la CNIL aura délivré une "autorisation unique" : une telle autorisation couvre, au-delà d'un cas particulier, les traitements portant sur des catégories de données identiques, présentant les mêmes finalités et visant les mêmes destinataires ou catégories de destinataires. Dans un tel cas, le responsable du traitement aura seulement l'obligation d'adresser à la CNIL une déclaration de conformité à l'autorisation unique.

Ainsi, par exemple, face au développement accéléré de la biométrie, la CNIL affirme une réponse différenciée : pour la vérification de l'appartenance d'une personne à un petit groupe dans le cadre d'activités usuelles (contrôle de l'accès aux locaux sur les lieux de travail, gestion des horaires et de la restauration de salariés ou d'élèves), la CNIL a délivré pour chacune de ces finalités une autorisation unique - de sorte que seule une déclaration de conformité est requise ; en revanche, les systèmes biométriques reposant sur la reconnaissance de l'empreinte digitale dans une base centralisée font l'objet d'une demande d'autorisation auprès de la CNIL, et doivent être justifiés par un « fort impératif de sécurité ».

3.4.2.4 Devoir de « mise à jour »

Aux termes des articles 31 de la loi et 11 du décret n° 2005-1309 du 20 octobre 2005, le responsable d'un traitement déjà déclaré ou autorisé doit informer sans délai et par écrit la CNIL : de tout changement affectant les informations précitées que la déclaration ou l'autorisation doit contenir ; de toute suppression du traitement.

3.4.3 Obligations positives vis à vis de la personne concernée

Selon l'article 7 de la loi de 1978, « Un traitement de données à caractère personnel doit avoir reçu le consentement de la personne concernée OU satisfaire à une des conditions suivantes : […] 5°) 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. »

[1] Conférence de presse 9 juillet 2007, « vos informations personnelles ont de la valeur, ne vous en fichez pas ! », Présentation du 27ème rapport d'activité de la CNIL 2006

[2] entre 2005 et 2006, les demandes d'autorisation de mise en oeuvre de dispositifs biométrique ont été multipliées par 10.

Si la rédaction de l'article semble poser le consentement de la personne concernée en principe, l'étendue du champ d'application du point 5° tend, en réalité, à faire de cette exception la règle.[1] Le consentement préalable de la personne ne semble ainsi pas requis vis-à-vis de l'exploitation, dans le cadre de ressources linguistiques, de données personnelles à caractère non « sensible ».

La question d'un contact entre l'exploitant de la ressource et la personne concernée par les données personnelles reste cependant posée du fait de l'obligation d'information. Si elle est la condition effective de l'exercice notamment du droit d'opposition, l'obligation d'information est systématique et d'une portée générale indépendante de la question de la légitimité du traitement ou d'une opposition éventuelle.

L'obligation d'information ne pose globalement pas de difficulté dans le cas d'un rapport direct entre le responsable du traitement et la personne visée : si les données sont recueillies directement auprès de la personne concernée, l'obligation d'information prévue à l'article 32 doit être accomplie à cette occasion et dans les conditions prévues.

Elle est en revanche susceptible de présenter un caractère significativement contraignant pour le développeur ou exploitant de ressources linguistiques, dans le cas d'une réutilisation de ressources ayant été originellement constituées à des fins propres, extralinguistiques – cas dans lequel le responsable du traitement n'est pas en relation directe avec la personne concernée.

Même en cas de rapport indirect, l'obligation d'information s'impose en principe : si les données n'ont pas été recueillies directement auprès de la personne concernée, l'obligation doit être accomplie par le responsable du traitement dès l'enregistrement des données ou, si les données seront appelées à être transmises à des tiers, au moment de la première communication.

En dérogation, la contrainte est cependant susceptible d'être écartée : dans le cas d'informations recueillies indirectement, il est fait exception à l'obligation d'information dans le cas où l'accomplissement de cette dernière est impossible ou exigerait un effort disproportionné par rapport à l'intérêt de la démarche. On peut penser que le traitement à finalité linguistique de données personnelles ne présentant pas de caractère sensible bénéficiera généralement de la dérogation tirée du caractère disproportionné de l'effort par rapport à l'intérêt de la démarche.

3.4.4 Obligation d'anonymisation des ressources

Comme vu supra, le point III de l'article 8 pose entre autres comme condition au traitement des données personnelles à caractère sensible visées au I que les dites données soient « appelées à faire l'objet à bref délai d'un procédé d'anonymisation préalablement reconnu conforme aux dispositions de la présente loi par la Commission nationale de l'informatique et des libertés »

Plus généralement, l'exigence, posée par le point 5°) de l'article 6, que les données ne soient pas conservées – sous une forme permettant l'identification des personnes – au-delà de la durée nécessaire aux finalités poursuivies, impose en pratique, de façon générale, une procédure d'anonymisation rapide.

Compte tenu de la définition légale extensive des « données à caractère personnel », cette procédure d' « anonymisation » devra conduire à effacer toutes les données permettant d'identifier directement ou indirectement la personne concernée, dans la mesure où elles n'apparaissent pas nécessaires à la constitution de la ressource linguistique. Si, dans la perspective d'un traitement suivi dans le temps, la conservation de références apparaît nécessaire, l'utilisation d'une fonction de hachage[2], transformant les données nominatives en données numériques pourra être utilisée pour obtenir une référence exploitable mais anonyme.

Dans la perspective de la constitution de fiches et bases à entrées de données multiples, on prendra spécialement garde aux informations élémentaires, non concernées si prises isolément, mais dont le recoupement permettrait d'identifier indirectement la personne concernée : cette circonstance, expressément assimilée par la CNIL à la définition des « données personnelles », présente un caractère peu manifeste impliquant une vigilance toute particulière pour des développeurs et exploitants de bases linguistiques comportant des recoupements de données.

[1] En ce sens, Voir notamment les travaux préparatoires à la loi du 06/08/2004 et le Rapport n° 1537 déposé le 13/04/2004 à l'Assemblée nationale

[2] http://www.cnil.fr/index.php?id=1536 CNIL.fr > approfondir > dossiers > technologies > Anonymisation

La CNIL recommande[1] différentes mesures face au risque d'identification par recoupement de fichiers : les données dont le croisement risque de lever l'anonymat devraient être réparties dans des fichiers ou des systèmes informatiques distincts ; le logiciel d'exploitation des bases de données en cause ne devrait permettre un recoupement libre des données qu'à titre exceptionnel, uniquement pour un petit nombre d'experts nommément habilités – les autres utilisateurs ayant accès aux informations pertinentes via des requêtes préprogrammées ; la procédure de collecte et saisie des données devrait, dans la mesure du possible, être répartie auprès de personnels ou organismes différents.

3.4.5 Conditions particulières applicables au transfert de données personnelles vers un Etat tiers

On entend ici par « Etat tiers » un pays non membre de la Communauté Européenne. Du fait de l'harmonisation des législations nationales des Etats membres en matière de protection des données à caractère personnel, opérée par la Directive 95/46/CE, le transfert vers un autre Etat membre n'est pas soumis à des conditions autres que celles du régime général applicable en droit interne.

Le transfert de données à caractère personnel vers un Etat tiers constitue, en lui-même, un « traitement » régi par la loi ; il est en principe interdit si l'Etat tiers destinataire ne garantit pas un niveau suffisant de protection de ces données (art. 68 et 70). Il est cependant fait exception à l'interdiction en cas de consentement exprès de la personne concernée (art. 69).

La protection en cause est celle de la vie privée et libertés et droits fondamentaux des personnes à l'égard du traitement dont ces données font l'objet ou peuvent faire l'objet.

Les Etats garantissant ou non un niveau de protection suffisant font l'objet d'une liste dressée par la Commission Européenne ; les Etats-Unis étant considérés comme ne présentant pas un niveau de protection suffisant, le transfert de données à caractère personnel vers des ressortissants est subordonné à l'application par ces derniers des « principes « de la Sphère de Sécurité » (Safe Harbor principles) ».

3.5 Ressources linguistiques, réutilisabilité et licences libres

3.5.1 « Licences Libres » ('Free Licences')

L'objectif de réutilisabilité des ressources, présenté en introduction, constitue – conjointement avec les contraintes et risques juridiques issus de l'application du droit commun, la considération centrale dans l'étude et la proposition de règles contractuelles pertinentes pour une "distribution nouvelle" des ressources linguistiques. Un esprit de libre distribution de ressources utiles à la Communauté, répondant à priori à la poursuite de l'objectif de réutilisabilité, est au cœur de la définition des célèbres Licences Libres ('Free Licences') telles que GNU GPL ou Creative Commons, développées et appliquées à l'origine dans le domaine du logiciel.

Comme le relève la partie introductive de la Licence GNU GPL – de loin la plus utilisée des licences libres –, son objet est d'écarter le risque que l'application à des logiciels techniquement « ouverts » ('open source') de règles de propriété intellectuelle ne revienne à en faire des logiciels propriétaires : problématique voisine de celle posée, au plan juridique, à l'objectif de réutilisabilité des ressources linguistiques.

Il est à ce titre intéressant d'analyser ces licences libres et d'en rechercher les possibles apports et incompatibilités dans l'optique d'une transposition au domaine des ressources linguistiques ; à ces fins, les principales ressources libres seront confrontées à la fois aux problématiques relevées dans nos développements précédents et à la référence constituée par les licences actuellement appliquées par ELRA.

3.5.1.1 Principes fondateurs des Licences Libres

Les licences libres reposent sur la combinaison de deux principes fondateurs communs correspondant à un esprit de partage et de large diffusion de l'œuvre considérée.

Un premier principe est la liberté d'utiliser, d'analyser, de modifier et de redistribuer de la chose acquise[2] sous une telle licence ; le droit de redistribution de l'œuvre est particulièrement au cœur de l'esprit des licences libres, comme l'illustre nettement le système de licences Creative Commons (CC) : le modèle Creative Commons repose sur la libre combinaison de diverses options pour la création de licences différentes ; la seule mention qui s'impose dans tous les cas est la mention 'sharing', correspondant à la libre redistribution.

[1] id.

[2] Il est important de noter que nous ne visons à ce stade que la redistribution de la chose telle qu'elle a été acquise, sans modification ou intégration ; les principes applicables dans le cas de modifications sont exposés plus loin.

Néanmoins, comme l'illustre la liberté de combinaison des options dans le système de licences Creative Commons, liberté appliquant le principe 'share what you want, keep what you want', les droits d'utilisation, modification et redistribution peuvent n'être que partiellement concédés, ou leur exercice subordonné à des conditions : ainsi, le droit de distribution des modifications apportées et produits dérivés de l'œuvre est fréquemment, mais pas toujours, concédé : l'option CC 'no modification' s'oppose aux modifications apportées dans une perspective de distribution de ces dernières ; l'option CC 'no commercial' s'oppose à une exploitation de l'œuvre à titre commercial. Ces restrictions ne mettent cependant pas en cause le principe général d'une concession des droits d'utilisation, analyse, modification et redistribution, les droits de modification et de redistribution étant en tout état de cause reconnus respectivement dans un cadre privé et à titre non commercial.

Pris isolément, le premier principe ne régit cependant que la concession des droits au licencié de premier rang et ne garantit pas la transmission de ces droits aux probables sous-licenciés ; cette question est réglée par le second principe fondamental inhérent à la notion de copyleft, dit principe de « contamination » : dans la mesure où il exerce le droit de redistribution, le licencié a l'obligation d'appliquer à la redistribution la même licence libre que celle sur la base de laquelle il a acquis la chose considérée ; la licence libre et la liberté de redistribution qu'elle comporte se reportent ainsi indéfiniment, dans la limite du domaine d'applicabilité de la licence.

Les deux principes présentés constituent les deux aspects du principe consacré par la licence GNU GPL[1] sous l'appellation « copyleft[2] » : cette appellation fait, par jeu de mot, référence au principe – central, en droit anglosaxon – de « copyright » sur lequel le principe de copyleft se fonde. S'il est défini en opposition au principe de copyright, le principe de copyleft repose en effet sur l'assertion[3], en premier lieu, de copyright et de droits de propriété intellectuelle sur l'oeuvre considérée. La reconnaissance de tels droits permet à l'auteur ou son ayant-droit de subordonner l'exercice des droits concédés au respect de conditions destinées à garantir le maintien de modalités libérales de partage et mutualisation : un produit qui ne serait pas à la base protégé par copyright serait, à titre de ressource du domaine public, librement distribuable mais également appropriable, pourrait librement être exploité à titre gratuit comme à titre commercial, sans que l'auteur ne puisse intervenir. Certaines licences libres s'approchent cependant, par leur caractère particulièrement peu contraignant, du régime applicable aux oeuvres relevant du domaine public : ainsi, dans le système de licences CeCILL – licences libres spécifiquement adaptées au droit français – le modèle CeCILL-B abandonne l'application du principe de copyleft moyennant, en compensation, une obligation accrue de mention de l'auteur et de la licence ; la licence BSD (développée pour le 'Berkeley Software Distribution, une distribution de système d'exploitation proche d'Unix) et ses dérivées accordent de même, dans le domaine du logiciel, une liberté générale d'utilisation et redistribution du code source et du code binaire, avec ou sans modification, moyennant le seul respect d'obligations de mention systématique des conditions applicables et de l'auteur et moyennant le respect d'une obligation négative dite de 'non endorsement'[4].

A titre complémentaire, les licences libres se caractérisent en outre par un troisième principe, répondant à une volonté de garantir concrètement, dans les faits, les conditions d'effectivité des droits et libertés accordées : une application notoire de cette volonté est l'obligation systématique, dans le domaine du logiciel, de systématiquement fournir ou mettre à disposition le code source du logiciel – le code binaire de la forme compilée et exécutable du logiciel n'étant globalement pas analysable et donc modifiable. L'effectivité de l'application des licences libres est par ailleurs recherchée au plan de l'information des intéressés quant aux conditions applicables, par des obligations de distribution systématique des informations conjointement à la chose distribuée, mais aussi par des efforts de clarté, visibilité et intelligibilité de ces indications : un apport fondamental du célèbre système de licences de Creative Commons est ainsi de symboliser par des signes graphiques simples (icônes) chacun des grands principes optionnels dont la combinaison définit le régime juridique applicable. Creative Commons propose par ailleurs un principe de « marquage » ('marking') en application duquel sont détaillées, en fonction de la nature du document considéré, les meilleures solutions techniques pour faire visiblement apparaître la mention de la licence applicable[5]. Ces solutions présentent un

[1] Est également répandue l'appellation 'Share-alike' (Partagez à l'identique) employée par les licences Creative Commons ; principes de copyleft et de share-alike sont globalement assimilables, quoiqu'ils ne se recouvrent pas exactement.

[2] Généralement traduit en français par « gauche d'auteur ».

[3] Rappelons qu'en droit du copyright, contrairement au droit d'auteur français, la protection de l'œuvre n'est pas acquise du seul fait de sa création mais nécessite un dépôt.

[4] Obligation négative s'opposant à ce que le nom des auteurs de l'oeuvre originale et des éventuelles modifications soient utilisés de façon à faire ou laisser penser que ces auteurs sont à l'origine ou cautionnent des modifications ultérieurement apportées par des tiers.

[5] Creative Commons propose ainsi différents marqueurs officiels « licence Creative Commons » adaptés aux différents médias : les icônes classiques, des avertisseurs audio ou audiovisuels, mais aussi des informations spécialement adaptées aux technologies informatiques : marqueurs pour site Web, instructions pour podcast, metadonnées intégrées aux fichiers audio ou audiovisuels et lisibles par la machine. Le projet 'Liblicense' vise à proposer aux développeurs d'applications des plug-ins dédiés à la reconnaissance automatique de la licence sous laquelle est distribuée la ressource exploitée.

réel intérêt dans la perspective de favoriser la diffusion du principe de réutilisabilité conjointement avec des ressources linguistiques empruntant à tous types de medias.

3.5.1.2 Adaptation des licences libres au droit français.

Développées en référence étroite au droit anglo-saxon du copyright et, à l'origine, pour une application spécifique au domaine du logiciel, les licences libres soulèvent une question de compatibilité avec une utilisation, en droit français, pour la distribution de ressources linguistiques. Si peu d'études ont été consacrées à la question de la possibilité d'appliquer les licences libres à des distributions en France, la problématique a été traitée au travers d'efforts de développement de licences libres spécifiquement adaptées aux exigences du droit français : l'ensemble de licences libres CeCILL développées par le CNRS, l'INRIA et le Commissariat à l'Energie Atomique, licences applicables à la distribution de logiciels.

Considéré en lui-même, le principe de copyleft se présente comme transposable en droit français : extensive, la notion fondatrice de copyright équivaut, en droit français, à un champ couvrant l'ensemble des droits de la catégorie Propriété Littéraire et Artistique : le droit d'auteur, les droits voisins au droit d'auteur et le droit des producteurs de bases de données. Les libertés concédées en application du copyleft trouvent des équivalents en droit français dans les droits d'utilisation, modification, représentation, reproduction et communication des œuvres. Sur le fondement de la liberté contractuelle reconnue en France, il est possible, conformément au principe de copyleft, de soumettre la concession d'une licence à une restriction des droits de redistribution du licencié, de façon à subordonner l'exercice du droit de redistribution à l'utilisation de la même licence.

Les incompatibilités des licences libres d'origine américaine avec le droit français relèvent, classiquement, en premier lieu, de la profonde différence d'approche de la protection de l'auteur dans les deux régimes.

La généralité et l'extensivité, dans les licences américaines, de la désignation des droits concédés est contraire aux exigences de l'article L. 131-3 du Code français de la Propriété Intellectuelle, aux termes duquel « la transmission des droits de l'auteur est subordonnée à la condition que chacun des droits cédés fasse l'objet d'une mention distincte dans l'acte de cession et que le domaine d'exploitation des droits cédés soit délimité quant à son étendue et à sa destination, quant au lieu et quant à la durée » ; les licences CeCILL se distinguent à cet égard par un effort de détail de l'étendue des droits concédés.

Au regard des droits moraux que le droit français, spécifiquement, reconnaît à l'auteur, on note que le principe de l'association systématique à une œuvre ou une partie d'œuvre de la mention de son auteur (droit de paternité de l'auteur sur son œuvre) est fréquemment[1], mais pas toujours, respecté dans les licences libres anglo-saxonnes : ainsi, notamment, le caractère optionnel de l'obligation d'attribution de l'oeuvre à son auteur dans le système de licences Creative Commons (option 'By') a dû faire place, dans l'adaptation française de ces mêmes licences, à une obligation systématique. Conformément à l'esprit – sinon au droit – de licences libres dont un élément fondamental est l'ascendant moral de l'auteur sur la destination de son œuvre, on constate par ailleurs au plan international un travail d'harmonisation vers le haut des dispositions nationales relatives au droit moral : ainsi, les pays de *Common Law* ne renoncent pas aux droits moraux comme le permettraient leurs législations nationales, en vue de respecter les normes des autres pays et la protection des auteurs et interprètes.

La question de la conformité au droit français des clauses des licences libres relatives aux garanties et responsabilités est traitée infra, au titre des problématiques relatives aux contrôles, garanties et responsabilités dans le cadre de l'application de licences libres.

3.5.1.3 Adaptation des licences libres au ressources linguistiques

Si les licences libres se présentent globalement, mutatis mutandis, comme applicables en France, sont-elles applicables dans le domaine spécifique des ressources linguistiques ?

L'Institut d'électronique et d'informatique Gaspard-Monge, organisme de recherche et d'enseignement de l'Université de Marne-la-Vallée dans les domaines de l'informatique, l'électronique, les télécommunications et réseaux, a implicitement mais nécessairement opté pour une réponse positive en développant une licence libre ayant spécifiquement vocation à s'appliquer à la distribution de ressources linguistiques : la licence Lesser General Public License for Linguistic Resources (LGPLLR).

[1] En dehors des considérations de droit moral de l'auteur sur son œuvre, l'application large du principe d'attribution dans les licences libres américaines relève d'une volonté de suivi des modifications et identification des versions distribuées.

L'applicabilité du principe de copyleft aux ressources linguistiques soulève cependant une question essentielle : comme vu précédemment, l'application du principe de copyleft est conditionnée par la reconnaissance, en premier lieu, de droits protégeant l'oeuvre considérée, à l'image du copyright en droit anglo-saxon.

Or il ressort de nos développements que les ressources linguistiques primaires telles que les corpus non annotés sont fréquemment développées et distribuées en tant que produits corporels ou digitaux mais étrangers à la propriété intellectuelle, sur la base de l'idée d'une appartenance au domaine public des documents incorporés – documents à caractère banal (enregistrements ou reproduction de communications correspondant à des situations usuelles) ou purement informatif ; l'intégration dans le corpus d'une oeuvre de l'esprit protégée à raison de son caractère original apparaîtra dès lors comme accidentel. Dans un tel cas de ressources primaires relevant du domaine public, le principe de copyleft n'étant pas applicable, l'éventuel encadrement d'une exploitation conforme au principe de réutilisabilité devra être recherché non dans une fictive licence de droits inexistants mais dans le cadre du contrat de service de fourniture des ressources. Les ressources linguistiques primaires fournies dans le cadre de contrats de fourniture avec des éditeurs d'œuvres d'auteur, seront en revanche de nature à être protégées et donc à fonder l'application d'un principe de copyright. Plus encore, les ressources linguistiques dérivées obtenues par l'apport de commentaires ou annotations linguistiques ou par un travail d'organisation systématique des données sont de nature à être protégées, au moins pour la part correspondant à ces apports, par le droit d'auteur ou le droit des bases de données et fonder l'application du copyleft.

Une telle lecture semble en fait correspondre à l'approche des auteurs de la licence LGPLLR : aux termes de son préambule, la licence LGPLLR ne prétend pas s'appliquer à toutes les ressources linguistiques sans distinction mais « à certaines ressources linguistiques spécifiquement désignées – typiquement des lexiques, grammaires, thésaurus ou corpus textuels. » Son article 0 définit la ressource linguistique comme « une collection de données relatives à une langue, organisées de façon à être utilisées avec des programmes d'application » : à la lecture d'une telle formulation, il semble que les auteurs aient implicitement fondé l'applicabilité du copyleft sur une protection des ressources linguistiques par le droit des bases de données.

Le droit des producteurs de bases de données est, on l'a vu, probablement le seul fondement possible pour la protection des travaux des développeurs de ressources linguistiques, dès lors que ces travaux ne se concrétisent pas par des œuvres de l'esprit marquées par l'empreinte de la personnalité de leur auteur (commentaires originaux, architecture originale de la base de données). Le droit des producteurs des bases de données est cependant moins satisfaisant que le droit d'auteur en tant que fondement à l'application du copyleft : la possibilité de reprendre, sans porter atteinte aux droits du producteur, une partie « non substantielle » de la base de données ouvre des possibilités d'appropriation de parties de la ressource linguistique, en contournement du principe de copyleft.

3.5.2 Licences libres et exploitation commerciale de ressources linguistiques

3.5.2.1 Licences libres et distribution à titre commercial de ressources linguistiques

3.5.2.1.1 Question d'une gratuité de principe des ressources linguistiques

Contrairement à une perception largement répandue, la diffusion sous licence libre n'implique pas nécessairement une distribution à titre gratuit ; comme l'indique assez clairement la traduction française (« licence libre »), le terme 'free' dans l'appellation 'free licence' implique non la gratuité mais la liberté de distribution. Ainsi des licences de référence telles que GNU GPL, LGPL et – en France – les licences CeCILL ne comportent pas de restriction quant à la liberté de distribution à titre onéreux ; dans le système Creative Commons, la distribution à titre onéreux est libre par défaut et peut être exclue par sélection d'une option 'no commercial' (exploitation à titre commercial exclue).

A cet égard, la licence LGPLLR , dans le domaine spécifique des ressources linguistiques, se distingue notablement en ce qu'elle pose le principe d'une gratuité des ressources linguistiques distribuées sous la licence. Ce principe de gratuité ne s'applique cependant qu'aux ressources linguistiques prises en elles-mêmes : il ne s'oppose pas à ce qu'une rémunération soit perçue au titre de services qualifiés, dans une perspective contractuelle, d'"accessoires" mais recouvrant, outre des services effectivement accessoires comme l'éventuelle fourniture de garanties contractuelles, l'activité centrale de prestation de service de fourniture des ressources.

En ce sens, le principe de gratuité des ressources linguistiques dans la licence LGPLLR repose sans doute moins sur une volonté contraignante que sur la simple expression de l'idée que les documents constitutifs des ressources linguistiques sont dénués de valeur commerciale intrinsèque – soit par nature (œuvres non protégeables), soit par le caractère externe et purement formel de l'exploitation qui en est faite dans le domaine linguistique (donnant

lieu à rémunération forfaitaire et limitée) – et que leur valeur est en réalité celle du service de mise à disposition de ces ressources.

En logique, cette perspective ne vaut qu'au regard des ressources linguistiques au sens étroit (ressources primaires, corpus non annotés) ; protégées à titre d'œuvres d'auteur, des œuvres de l'esprit à objet directement linguistique (documents de dissertation et analyse linguistique, travaux de développement d'architectures originales de bases de données ...) présentent une valeur commerciale intrinsèque, se rattachent dans l'esprit aux services « accessoires » fournis par les développeurs et distributeurs de ressources linguistiques et ne relèveraient donc pas de l'application du principe de gratuité.

3.5.2.1.2 Licences libres et détermination du prix de distribution des ressources linguistiques

Au regard de la licence spécifique LGPLLR, l'application d'un principe de gratuité des ressources linguistiques prises en elles-mêmes soulève la question de contraintes dans la détermination du montant de la rémunération du service de fourniture des ressources : dans l'hypothèse où le principe de gratuité des ressources primaires serait entendu comme réellement contraignant, seule une tarification correspondant strictement aux coûts attestés de la prestation de service de fourniture serait de nature à en garantir le respect, dans la mesure où la recherche d'une marge commerciale serait indissociablement rattachable tant au service qu'aux ressources fournies.

Appliquée strictement, une telle logique de gratuité des ressources et de facturation des seuls coûts de mise à disposition écarte la prise en compte, dans la détermination du prix, de la rareté sur le Marché des ressources acquises et de l'influence de cette rareté sur la valeur commerciale de celles-là.

Dans les faits, un distributeur tel qu'ELDA se conformerait globalement à une telle logique de gratuité des ressources linguistiques en ce qu'il applique en principe une tarification à hauteur de ses coûts (rémunération des services de gestion et mise à disposition, rémunération des fournisseurs – en général par un système de royalties). Le prix de la ressource est ainsi déterminé, à titre essentiel, par le fournisseur initial ; si les négociations commerciales restent une réalité générale, la variété des types de fournisseurs initiaux et des circonstances de création de la ressource implique des logiques variées de détermination du prix, ne présentant de rattachement systématique ni avec le principe de gratuité, ni avec une application des prix du Marché : des entités académiques ayant constitué des ressources linguistiques sur la base de subventions peuvent ainsi suivre spontanément une logique de gratuité et de rémunération des seuls coûts éventuels ; des organismes ayant créé des ressources linguistiques dans le cadre de projets commerciaux tendront à prendre en compte des facteurs tels que l'exploitabilité des dites ressources pour les projets concurrents … La rareté et le prix sur le Marché de ressources d'une langue et d'un type donnés semblent ainsi moins conduire la détermination des prix pratiqués que conditionner la hauteur des investissements et le type de ressources créées.

Au plan juridique, la problématique de détermination du prix soulève la question de la liberté, dans le cadre d'une licence libre, de pratiquer des prix différenciés. Au-delà de considérations d'opportunité commerciale, la possibilité de pratiquer des prix différenciés conditionne la mise en place de systèmes tels qu'une tarification réservée à des cotisants ou déterminée à raison des capacités économiques des différents utilisateurs (entités académiques ou à but non lucratif ? organismes à but lucratif ?) ou qu'une péréquation tarifaire pour le soutien à certaines ressources.

Dans les faits, les distributeurs de licences linguistiques tels que LDC et ELDA ont largement recours à la pratique de prix réservés à des adhérents cotisants ; le site web du Linguistic Data Consortium[1] met ainsi tout spécialement en avant l'adhésion au groupe comme moyen normal d'obtenir les ressources distribuées[2] et de bénéficier de services spécifiques. La pratique d'ELDA comporte notamment la proposition de tarifs avantageux aux membres et aux organismes académiques.

Dans le cadre de licences libres autorisant une exploitation commerciale, rien ne semble restreindre la liberté de pratiquer des prix différenciés : la redistribution y étant garantie comme liberté et non comme obligation, la pratique de prix « prohibitifs » à l'égard de certaines personnes constitue a priori une simple restriction de fait, par le distributeur, de l'exercice de sa liberté de distribution ; on pourrait cependant considérer qu'un prix discriminatoire viole le principe de copyleft et la liberté de redistribution de l'acquéreur placé en situation

[1] http://www.ldc.upenn.edu/Obtaining

[2] La page « obtenir des ressources » ('obtaining') propose ainsi quatre accords d'adhésion (Membership Agreements mais, bien qu'il existe par ailleurs, l'accord par lequel un non adhérant peut obtenir des ressources n'est ni présenté sur cette page ni, de façon générale, mis en avant.

concurrentielle, dans la mesure où il place ce dernier dans une situation défavorable et restreint – en pratique – l'effectivité de sa liberté de redistribution[1].

Un principe contraignant de gratuité des ressources linguistiques et de rémunération des seuls coûts du service est-il de nature à conditionner la liberté d'appliquer des prix différenciés ? Deux interprétations du principe sont possibles : selon une première lecture, le prix de chaque prestation sera strictement déterminé par le relevé des coûts spécifiquement générés par la fourniture du service ; selon une seconde lecture, le principe de gratuité vise globalement l'activité du prestataire comme n'ayant pas vocation à dégager des recettes commerciales au-delà de la couverture de ses coûts : cette seconde lecture autorise l'application des différents systèmes tarifaires envisagés supra.

3.5.2.2 Licences libres et utilisation commerciale de la ressource

3.5.2.2.1 Licence de distribution et prise en compte des finalités commerciales de l'exploitation par l'acquéreur

Comme vu précédemment, la notion de licence libre et le principe de copyleft n'impliquent pas en eux-mêmes une distribution ou une exploitation à titre gratuit. L'option 'no commercial', dans le système de licences Creative Commons, permet de réserver l'exercice des droits concédés par la licence à des activités à but non lucratif.

Les schémas de distribution actuellement pratiqués reflètent ils une telle distinction entre activités commerciales et activités non commerciales ? La définition de licences de distribution sur la base d'une prise en compte explicite des finalités commerciales ou non de l'exploitation est-elle de nature à conforter ou préciser les pratiques de distribution actuelles ?

Au plan commercial, la considération des finalités lucratives ou non de l'exploitation des ressources linguistiques est significative dans les rapports commerciaux entre distributeur et exploitant, en ce qu'elle conditionne tant les capacités économiques des intéressés et la portée sur le Marché de l'exploitation qu'ils font des ressources acquises. La prise en compte explicite, par les accords d'adhésion du LDC, de la finalité commerciale ou non de l'exploitation répond notamment à de telles considérations ; quoique les licences ELDA ne soient pas définies sur la base de cette opposition, ELDA, en pratique, prend en compte cette considération notamment en proposant aux organismes académiques des prix significativement inférieurs à ceux appliqués aux personnes physiques ou morales à finalité commerciale.

Les intitulés des accords d'adhésion du Linguistic Data Consortium manifestent nettement l'application de régimes différents selon que l'acquéreur des ressources entend ou non les exploiter dans le cadre d'un projet commercial : se distinguent ainsi, à titre essentiel, un accord d'adhésion pour les organisations à but lucratif (For Profit Membership Agreement), un accord d'adhésion pour les organisations à but non lucratif (Not For Profit Membership) et un accord d'adhésion pour les Entités gouvernementales.

Au vu de leur intitulé, les licences ELDA reposent en revanche sur une différenciation à raison des modalités techniques ou fonctionnelles d'exploitation des ressources – selon que l'exploitation comporte ou non l'intégration de la ressource à la production propre de l'exploitant. Dans les faits, l'exploitation d'une ressource linguistique à titre commercial impliquera ordinairement son intégration, de sorte que la distinction licence intégrateur – licence utilisateur final correspond très largement, en pratique, à la distinction exploitant à titre commercial – exploitant à but non lucratif (académies essentiellement). Néanmoins, tant la licence utilisateur que la licence intégrateur envisagent les hypothèses de distribution à titre gratuit et à titre onéreux[2], ce qui illustre que l'exploitation à titre commercial ou non n'est pas, en droit, le critère d'application de l'une ou l'autre des licences.

3.5.2.2.2 Exploitation à titre gratuit et atteinte aux droits de propriété intellectuelle.

[1] Ces observations concernent les seules exigences de la ressource libre éventuellement utilisée ; la pratique de prix différenciés est susceptible d'être par ailleurs sanctionnée sur le fondement du droit commun, notamment du droit de la concurrence (abus de position dominante)

[2] Ainsi, la licence utilisateur final prévoit en son point 5 que « L'utilisateur n'est pas autorisé à reproduire les ressources linguistiques, ni à distribuer des produits dérivés ou services incluant tout ou partie des données dans un but commercial ou non (distribution gracieuse) sous toute forme ou par tout moyen. » ; la licence intégrateur (point 5) précise, comme la licence utilisateur final (point 6) que l'exclusion de tout droit de redistribution de la ressource s'étend à « toute forme de distribution, y compris des copies gratuites ou open source » [nous soulignons]

Au plan juridique, la prise en compte des finalités commerciales ou non de l'exploitation pourrait notamment relever de la recherche de garanties contractuelles et de l'idée qu'une exploitation à titre gratuit constituerait une garantie vis-à-vis de l'atteinte, au travers de l'exploitation, aux droits de propriété intellectuelle portant potentiellement sur les œuvres intégrées aux ressources exploitées.

Prise isolément, la circonstance d'une exploitation à titre non commercial n'est cependant pas de nature à écarter de tels risques.

Ainsi, l'article L.122-4 du CPI vise, sans restriction, comme illicite « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit » et précise qu' « Il en est de même pour la traduction, l'adaptation ou la transformation, l'arrangement ou la reproduction par un art ou un procédé quelconque. »

Si l'article L.122-5 CPI, prévoyant des limites à ces droits patrimoniaux de l'auteur de l'œuvre (après divulgation autorisée par l'auteur), envisage en son point 1° la question d'une représentation à titre gratuit, ce critère de gratuité est indissociablement lié à celui du caractère privé de la représentation. Le critère du caractère privé de l'exercice des droits est seul mentionné, à l'exclusion de celui de gratuité, au regard de l'exercice du droit de reproduction. L'exploitation de ressources linguistiques semble impliquer, à raison de leur nature, l'exercice de droits de représentation dans la plupart des cas – vis-à-vis de tiers ou au sein même de l'organisme ; en ce sens, la recherche de garanties contractuelles, dans la distribution des ressources linguistiques, vis-à-vis d'atteintes par l'exploitant aux droits d'auteur portant potentiellement sur les œuvres incorporées impliquera de poser systématiquement les deux exigences d'une exploitation à titre strictement privé et d'une exploitation à titre gratuit.

En fait, le seul cas de légitimation d'une exploitation publique par son caractère non commercial est visé à l'article L.122-5-3° e) du CPI qui, comme on l'a vu, prévoit, au bénéfice d'activités pédagogiques et scientifiques, un droit de représentation ou reproduction sans autorisation préalable de l'auteur[1] mais encadré par de strictes conditions, dont celle de l'absence de toute exploitation commerciale.

3.5.2.3 Le Copyleft et le principe d'un terme de la chaîne de distribution (notions d'utilisateur final et d'intégrateur)

Le système de licences actuellement utilisé par ELRA repose sur l'application de contrats de licences différents entre fournisseur et distributeur (Distribution Agreement), entre distributeur et intégrateur (Value-Added Reseller (VAR) Agreement) et entre distributeur et utilisateur final (End-User Agreement). On trouve au cœur des licences intégrateur et utilisateur final une volonté d'exclusion ou restriction des droits de redistribution à des tiers des ressources linguistiques distribuées par ELDA. Utilisateur final comme intégrateur sont conçus comme constituant à la fois la finalité et le terme de la chaîne de distribution/modification de ressources linguistiques stricto sensu : l'utilisateur final ayant uniquement vocation à utiliser la ressource linguistique, est expressément exclu de la licence qui lui est accordée tout droit tendant à la redistribution de la ressource à des tiers ; l'intégrateur n'ayant vocation à distribuer la ressource linguistique acquise que via l'intégration à un produit ou un service, est exclu de la licence intégrateur le droit de redistribution directe ou indirecte[2] de la ressource linguistique. En ce qu'elle prévoit le recours par le distributeur aux licences utilisateur et intégrateur évoquées, la licence distributeur exclut en principe le droit d'ELDA de transmettre les ressources à des distributeurs subséquents[3].

Malgré des intitulés différents, les accords d'adhésion proposés par le Linguistic Data Consortium suivent globalement une logique identique au système de licences ELDA. Tous les accords LDC prévoient que « sauf autorisation expressément prévue par le présent acte, il n'est reconnu au Membre aucun droit de copie, de redistribution, de transmission, de publication ou autre utilisation des bases de données LDC à toute fin autre » que celle expressément et spécifiquement visée par l'accord. La distinction établie par LDC entre les Membres selon qu'ils poursuivent ou non des finalités commerciales (For Profit Member Agreement / Not For Profit Member Agreement) rejoint la distinction opérée par ELDA entre Utilisateur final et Intégrateur : les clauses

[1] Etant entendu que cette limite apportée ponctuellement aux droits de l'auteur ne concerne que les aspects patrimoniaux de ces derniers, les droits moraux ne pouvant être mis en cause en aucune circonstance.

[2] Licence ELDA - Intégrateur, article 6 : « L'intégrateur n'est pas autorisé à reproduire les ressources linguistiques, ni à distribuer des produits dérivés ou services incluant tout ou partie des données dans un but commercial ou non (distribution gracieuse), sous toute forme ou par tout moyen, qui permettraient de reconstituer les ressources linguistiques ou une partie des ressources linguistiques. » (nous soulignons)

[3] Une version longue de la licence Distributeur, rarement utilisée, prévoit explicitement que la transmission par ELDA des ressources à un distributeur subséquent implique l'obtention préalable d'une autorisation spécifique, par écrit, du fournisseur : le droit de redistribution n'est donc en tout état de cause pas intégré au système de licences utilisé par ELDA.

tendent à exclure la redistribution directe ou indirecte, à des tiers[1], des ressources linguistiques acquises ; comme la Licence Intégrateur d'ELDA, l'accord d'adhésion pour les Membres poursuivant un but commercial accorde, y compris à des fins commerciales, un droit de reproduction et distribution de parties des bases de données du LDC dans la mesure où ces données sont intégrées à la propre production du Membre.[2]

De tels schémas de distribution reposant sur une exclusion générale du droit de redistribution semblent incompatibles avec l'application de licences libres.

Un engagement contractuel par lequel le licencié s'interdirait d'exercer le droit de redistribution tiré de la licence libre impliquerait la participation du donneur de licence à la détermination de conditions de distribution divergeant de façon critique de celles prévues par la licence libre : dès lors, soit le donneur de licence est le titulaire original des droits et les conditions nouvelles sont applicables quoique non qualifiables de licence libre ; soit le donneur de licence a lui-même acquis ses droits en application de la licence libre considérée, et l'application d'une restriction, même conventionnelle, du droit de distribution du licencié devrait constituer une atteinte nette au principe essentiel de copyleft qui s'impose à lui.

Que penser par ailleurs de la possibilité que le licencié s'interdise unilatéralement d'exercer le droit de redistribution qu'il tire par principe de la licence libre ? La redistribution étant, aux termes de la licence, un droit et non une obligation, un tel engagement n'apparaîtrait pas comme une violation frontale de la licence libre ; néanmoins, la réalité du caractère purement unilatéral d'un tel engagement paraîtrait, en premier lieu, pour le moins douteuse dans la mesure où l'engagement se présenterait, en fait sinon en droit, comme la condition de la concession de licence ; de plus, un engagement purement unilatéral en ce sens serait probablement sans portée dans la mesure où l'engagement unilatéral n'est reconnu que de façon restrictive en droit français, uniquement dans un esprit de sanction de promesses abusives – hypothèse ne correspondant pas au cas présent.

Si l'application de licences libres se présente comme de nature à remettre en cause les schémas actuels de distribution de ressources linguistiques, apprécier la mesure de cette mise en cause implique de détailler les aspects du principe de copyleft.

3.5.2.4 Copyleft fort, copyleft faible et ressources linguistiques

3.5.2.4.1 Principe de la distinction

Comme mentionné supra, les licences libres régissent l'exploitation et la redistribution des produits acquis, mais reconnaissent également au licencié un droit d'accès, d'analyse et de modification vis-à-vis du produit, ainsi que, fréquemment, un droit de distribution des modifications ainsi apportées – l'idée originale des licences libres étant d'assurer la mise en commun non seulement des oeuvres de base mais également des modifications, améliorations et contributions personnelles apportées à ces oeuvres par les licenciés.

La problématique du régime juridique applicable à la distribution des modifications a été formalisée, dans le cadre de la Licence GNU GPL, par l'opposition entre les principes dits de copyleft faible ('weak copyleft') et de copyleft fort ('strong copyleft'). Le 'copyleft faible' se limite à l'exigence, vue précédemment, d'une libre redistribution du produit acquis moyennant que cette redistribution soit opérée sur la base de la même licence libre : ainsi, dans le cas d'apport de modifications, ces dernières pourront être distribuées par le licencié sous la licence de son choix. Le principe de 'copyleft fort' se distingue du 'copyleft faible' par une expansion du principe de « contamination » : les règles posées par la licence libre considérée seront applicables non seulement à la redistribution du produit acquis sous copyleft fort, mais également à la distribution des œuvres dérivées de ce dernier et aux modifications qui lui ont été apportées ; le principe de contamination du copyleft fort ne s'étend pas en revanche aux simples « agrégats », produits qui, bien que distribués conjointement avec l'œuvre sous copyleft, par exemple sur un même support, présentent vis-à-vis de cette dernière un caractère indépendant et autonome.

1 La notion de tiers dans les accords LDC ressort d'une lecture a contrario de la disposition selon laquelle : « Afin de recevoir certaines bases de données de LDC, le Membre a l'obligation de restreindre l'utilisation de ces bases de données à ses seuls employés et consultants placés sous son contrôle et qui ont signé, préalablement à l'obtention de l'accès aux bases de données spécifiées, tout accord Utilisateur additionnel nécessaire. »

2 L'accord LDC est en lui même moins précis et restrictif sur ce point que la licence Intégrateur en ce qu'il ne traite pas, contrairement à la seconde, de la question du degré d'intégration, de l'accessibilité des ressources intégrées et de la possibilité de les extraire et reconstituer ; des garanties indirectes peuvent apparemment être apportées sur ce point par la signature prévue, en tant que de besoin, d'accords Utilisateur complémentaires.

Au regard des pratiques de distribution, cette distinction entre copyleft fort et copyleft faible relativise l'incompatibilité de principe entre application, d'une part, d'une licence libre et, d'autre part, de schémas de distribution limitée reposant sur les notions d'utilisateur final et intégrateur : en application d'un copyleft faible, le licencié aurait la liberté de distribuer sous la licence de son choix – donc éventuellement sans droit de modification ou redistribution – les modifications et apports dont il serait l'auteur : ainsi, dans le domaine de ressources linguistiques composites, le corpus sous copyleft faible resterait obligatoirement soumis à la liberté de redistribution tandis que les modifications et commentaires linguistiques pourraient être librement soumis aux règles décidées par leur auteur.

3.5.2.4.2 Le Copyleft fort appliqué aux ressources linguistiques

Dans le domaine des ressources linguistiques, deux types d'œuvres sont susceptibles d'être soumises au principe de contamination, à raison du fait qu'elles dérivent d'une oeuvre soumise au copyleft fort.

En premier lieu, la modification d'une ressource linguistique peut conduire à la création d'une nouvelle ressource linguistique, au sens strict : tel sera le cas pour l'annotation linguistique d'un corpus ou la compilation et la réorganisation de bases de données. En application du principe de contamination du copyleft fort, une ressource dérivée d'une telle nature sera normalement soumise dans son ensemble à la licence applicable à la ressource de base. Dans l'optique de poursuite d'un objectif de réutilisabilité des ressources linguistiques, le copyleft fort présenterait l'intérêt considérable d'introduire une spirale de propagation rapide du principe d'une distribution des ressources linguistiques sous licence libre. Si l'application du principe de copyleft fort encadre les conditions de distribution des modifications œuvres dérivées, il ne peut faire de cette distribution une obligation ; la distribution de tels apports étant d'intérêt majeur dans l'optique du principe de « réutilisabilité », il conviendra dès lors de s'assurer que le développeur de modifications et ressources linguistiques dérivées soit porté à exercer son droit de distribution par des motivations commerciales : le principe de copyleft fort ne saurait ainsi se doubler d'un principe de gratuité de la ressource linguistique primaire, ce principe de gratuité devant se communiquer en second lieu aux modifications et œuvres dérivées et restreindre les perspectives commerciales de la distribution de ces apports à la simple compensation des coûts de développement et fourniture.

Les limites du champ d'application du copyleft fort soulèvent par ailleurs la question des rapports entre les différentes ressources organisées en un ensemble élargi : la licence GNU GPL prévoit en effet que le copyleft fort applicable à un produit déterminé ne s'applique pas aux produits qui, quoique distribués conjointement avec le produit considéré, ne constituent, vis-à-vis de ce dernier que des « agrégats » : ces derniers sont définis comme des produits séparés et indépendants ne constituant pas des extensions de l'œuvre sous copyleft fort et qui ne sont pas distribués conjointement avec cette dernière de façon à former un produit fonctionnel[1] plus large. L'application de cette définition aux cas de distribution conjointe de ressources linguistiques semble bien délicate et de nature à conduire à des appréciations largement subjectives ; au regard de la dernière partie de la définition, on peut penser des rapports logiques (tels que des comparaisons statistiques) pourront facilement être établis entre des ressources d'origines indépendantes et de natures variées, si bien que la notion d'agrégat sera d'applicabilité exceptionnelle et la « contamination » d'applicabilité générale.

Un second type d'oeuvre dérivée des ressources linguistiques est constitué par les logiciels intégrant ces ressources.

L'expérience montre clairement la difficulté de l'application du principe de copyleft fort dans des domaines comportant la distribution à titre commercial de logiciels. L'application du copyleft fort implique l'obligation pour les personnes entendant commercialiser des logiciels intégrant tout ou partie des produits soumis au principe d'appliquer ce dernier au produit d'intégration même, à savoir le logiciel commercialisé ; si la fourniture à titre gratuit n'est pas systématiquement requise par l'application d'une licence libre, l'obligation pour le distributeur du logiciel de concéder la libre modification et redistribution de ce dernier et d'en fournir le code source est ressenti en pratique comme s'opposant à une exploitation à titre commercial. Ainsi, les revendications des distributeurs commerciaux de logiciels ont conduit la Free Software Foundation, créateur de la licence GNU GPL, à proposer à titre alternatif une version revenant à l'ancien principe de copyleft faible, la licence LGPL (notamment pour la distribution de bibliothèques logicielles ayant vocation à être intégrées notamment à des logiciels commerciaux).

Dans le domaine de la modification et redistribution de ressources linguistiques au sens strict, l'application d'un principe de copyleft fort ne semble pas devoir entraver l'exploitation, commerciale ou non, des ressources dérivées constituant elles mêmes des ressources linguistiques ; en revanche, dans la mesure où une finalité

[1] Nous proposons cette notion de « produit fonctionnel plus large » en forme d'adaptation générique de la notion spécifique de « programme plus large » utilisée par une licence GNU GPL conçue pour des logiciels.

essentielle, en pratique, de la production et distribution de ressources linguistiques est l'exploitation et intégration de ces dernières par des logiciels commerciaux, l'application d'une licence libre aux ressources linguistiques soulèverait dans le domaine une problématique identique à celle ayant conduit au repli sur la licence LGPL. Il est significatif à cet égard que dans son effort de développement d'une licence libre spécialement adaptée à la distribution des ressources linguistiques, la licence LGPLLR ait retenu comme modèle de base la licence LGPL.

L'application pragmatique du principe de copyleft dans le domaine des ressources linguistiques devrait ainsi probablement conduire à établir une distinction entre produits dérivés selon qu'ils restent qualifiables de ressource linguistique au sens étroit ou qu'il s'agit de logiciels intégrant la ressource – distinction conduisant à appliquer un copyleft fort dans le premier cas, un copyleft faible dans le second[1].

3.5.2.5 Le copyleft comme levier de mise en place de standards techniques en adéquation avec le principe de réutilisabilité

Comme on l'a vu, l'application d'un principe de copyleft est de nature, notamment dans le domaine des logiciels, à compromettre l'exploitation commerciale des modifications et produits dérivés de l'œuvre dont la distribution est soumise au principe.

L'intérêt des développeurs et distributeurs, à titre commercial, de logiciels étant globalement de se soustraire à l'application du principe de contamination, la définition du champ d'application de se dernier est susceptible de faire pression dans le sens de l'exploitation de l'œuvre distribuée sous certaines formes ou selon certaines modalités (selon une logique comparable, par exemple, aux mesures fiscales incitatives et dissuasives) ; ce champ d'application peut notamment être délimité par la frontière que traceront les définitions de la licence entre, d'une part, produits dérivés de l'œuvre sous licence libre (soumis à la licence) et, d'autre part, œuvres présentant vis-à-vis de l'œuvre sous licence libre une autonomie justifiant qu'elles échappent à une obligation d'appliquer la même licence.

Une originalité et un apport essentiels de la licence LGPLLR sont à ce titre la distinction qu'elle introduit, relativement à l'application du copyleft, entre œuvre fondée sur une ressource linguistique ('work based on the Linguistic Resource') et œuvre utilisant la ressource linguistique ('work that uses the Linguistic Resource').

La licence définit le premier type comme couvrant « tant la ressource linguistique elle-même que toute œuvre dérivée soumise au droit du Copyright » ; l' « œuvre dérivée » est elle-même définie comme « l'œuvre contenant tout ou partie de la Ressource Linguistique, soit telle quelle (verbatim), soit avec des modifications et/ou la traduction fidèle dans une autre langue »[2]. Selon la section 3 de la licence LGPLLR, est au contraire appelée « œuvre qui utilise la Ressource Linguistique » « un programme qui n'intègre aucun élément dérivé de quelque partie de la Ressource Linguistique, mais est conçu pour fonctionner avec la Ressource Linguistique (ou une version encryptée de la Ressource Linguistique) en la lisant, en étant compilé ou lié avec elle.»

Les conséquences juridiques attachées par la licence LGPLLR à la qualification en « œuvre dérivée » ou en « œuvre qui utilise la ressource linguistique » conduisent à faire de la seconde le standard impératif de distribution d'un logiciel exploitant une ressource linguistique sous licence LGPLLR.

En premier lieu, selon la section 2 de la licence, le droit de distribution de modifications qualifiables d' "oeuvre fondée sur la ressource linguistique" est subordonné notamment (point a) à la condition que l'œuvre dérivée considérée soit elle même une ressource linguistique. La licence définit la ressource linguistique comme « une collection de données relatives à la langue, organisée de façon à être utilisée avec des programmes d'application" ; la notion de ressource linguistique est donc entendue dans un sens étroit n'incluant pas les logiciels d'exploitation : la condition du point a de la section 2 signifie donc finalement que n'est pas concédé de droit de distribution d'un logiciel intégrant tout ou partie de la ressource.

Au contraire, la section 3 de la licence énonce, immédiatement après avoir défini l' « œuvre qui utilise la ressource linguistique » qu' « une telle œuvre, prise isolément, n'est pas une oeuvre dérivée de la Ressource Linguistique, et n'entre dès lors pas dans le champ d'application de cette Licence » ; il s'ensuit que la distribution

[1] Même dans le second cas, l'application d'un copyleft fort ne serait pas théoriquement une entrave irrémédiable à l'intégration de la ressource linguistique à un logiciel propriétaire, dans la mesure où une dispense d'application de la licence et du copyleft fort peut naturellement être recherchée auprès du titulaire des droits concédés. S'annonçant systématique, le recours à une telle solution réduirait cependant considérablement la portée effective de la licence.

[2] Traduit de la version anglaise.

séparée d'un logiciel conçu pour exploiter une ressource linguistique mais ne l'intégrant ni totalement ni en partie pourra être librement soumise à une licence au choix du distributeur – donc notamment une licence conçue pour une distribution à titre commercial. La licence tend donc à imposer le standard technique de logiciels non intégrateurs, non seulement en prohibant la distribution de logiciels intégrateurs mais également en ménageant, d'autre part, les conditions d'une motivation commerciale pour la distribution de logiciels non intégrateurs.

Une telle démarche en faveur du développement, dans le domaine de l'exploitation des ressources linguistiques, du standard technique du logiciel non intégrateur, est de nature à servir fortement les intérêts du principe de réutilisabilité des ressources linguistiques en ce qu'elle tend à écarter les standards d'intégration qui, constituent, au plan technique, l'entrave essentielle au principe de réutilisabilité. Dans une telle perspective de réutilisabilité des ressources linguistiques, deux limites apparaissent néanmoins dans l'hypothèse présente d'une distribution séparée du logiciel non intégrateur et de la ressource : d'abord, rien ne s'oppose à ce que l'intéressé diffuse la ressource sous forme cryptée uniquement, aucune obligation de fournir une version non cryptée n'étant prévue dans ce cas ; ensuite, le fait que le logiciel n'intègre pas la ressource linguistique et soit distribuable séparément n'exclue pas que les deux éléments présentent un caractère mutuellement spécifique, réduisant ou écartant les possibilités d'utilisation du logiciel avec d'autres ressources linguistiques, et de la ressource linguistique avec d'autres logiciels d'exploitation.

De telles questions sont cependant traitées par la licence LGPLLR dans le cas précis où le logiciel non intégrateur est distribué conjointement avec une ressource linguistique qu'il utilise. Il en résulte, dans le cas visé, des solutions présentant un réel intérêt dans la poursuite d'un objectif de réutilisabilité des ressources linguistiques.

D'après la section 4 de la licence LGPLLR, le droit de distribution de tels packs est notamment soumis à deux[1] conditions alternatives, visées aux points a) et b)[2] :

« Accompagner le pack de l'intégralité de la ressource linguistique dans sa forme non cryptée ('legible') et lisible par la machine, en incluant dans le pack toute modification apportée dans le cadre de l'utilisation (modifications qui devront être distribuées conformément aux sections 1 et 2 [de la licence LGPLLR]) et, si le pack comprend une forme cryptée de la ressource linguistique, de l'intégralité, dans une version lisible par la machine, de l' « oeuvre qui utilise la ressource linguistique », comme code objet et/ou code source, de façon à ce que l'utilisateur puisse modifier la ressource linguistique puis l'encrypter pour produire un pack modifié intégrant la ressource linguistique modifiée. »

« Utiliser, pour la combinaison avec la ressource linguistique, un mécanisme pertinent ('suitable'). Est pertinent le mécanisme qui fonctionnera correctement avec une version modifiée de la ressource linguistique, si l'utilisateur en installe une, dès lors que la version modifiée présentera une compatibilité d'interface avec la version avec laquelle le pack a été créé. »

Les conditions du point a) reprennent globalement la logique, suivie par les licences libres applicables aux logiciels, tendant à assurer l'effectivité du droit d'analyse et de modification par une obligation de fournir systématiquement ou mettre à disposition le code source. La définition donnée du terme 'legible', désignant la forme non cryptée et intelligible de la ressource linguistique reprend d'ailleurs mot pour mot la définition que donne les licences GPL et LGPL du code source.

Le point b) présente un intérêt tout particulier, en ce qu'il exige en substance que les développeurs d' « œuvres utilisant la ressource linguistique » se conforment à des standards techniques de nature à garantir une interopérabilité du logiciel avec diverses ressources linguistiques utilisant la même interface. La mise en place de conditions techniques d'interopérabilité se présente comme une condition matérielle essentielle pour assurer la réutilisabilité des ressources linguistiques ;

Le titulaire des droits éventuels sur la ressource linguistique aurait il les moyens d'étendre de telles exigences dans le cas de la distribution séparée de la ressource et d'un logiciel non intégrateur ? La clause d'inapplicabilité de la licence LGPLLR dans ce cas n'est-elle pas un simple constat d'une impossibilité d'étendre les droits sur la ressource linguistique à un contrôle de l'exploitation et distribution d'un logiciel distribué séparément ? De fait, l'absence d'intégration de tout ou partie de la ressource linguistique dans le logiciel implique que la distribution de ce dernier ne comporte pas en soi l'exercice d'un droit de distribution de la ressource. Néanmoins, les droits du producteur d'une base de données sur cette dernière comportent non seulement la reproduction de tout ou

[1] Les conditions alternatives de la section 4, visées aux points a) à e), sont au nombre de 5, mais les points c) à e) constituent globalement des modalités alternatives de satisfaire aux conditions matérielles du point a)

[2] traduit de l'anglais

partie mais également l'utilisation de la base : dans la mesure où le logiciel utilise la base de données constitutive d'une ressource linguistique, il implique l'exercice du droit d'utilisation de la base, de sorte que le titulaire des droits sur la base peut, indirectement, contrôler l'exercice du droit d'utilisation du logiciel et le soumettre à des conditions telles que celles vues aux paragraphes précédents.

Dans les faits, compte tenu des possibilités de modifier un logiciel pour l'adapter à des ressources différentes, exploiter les droits sur une ressource linguistique pour contrôler indirectement l'exercice des droits d'utilisation du logiciel impliquera de constituer, par concentration, des ressources linguistiques qui, par des apports quantitatifs et qualitatifs déterminants, sauront se rendre incontournables.

3.5.3 Garanties, responsabilités et contrôles dans l'exploitation de ressources linguistiques sous licence libre.

3.5.3.1 Clauses limitatives ou exclusives de garantie et responsabilité dans les licences libres

Les larges libertés accordées par les licences libres ont pour contrepartie, dans le cadre du modèle type de la licence libre considérée, la stipulation systématique de clauses aussi restrictives que possibles au regard des garanties apportées et de l'engagement de la responsabilité du concédant.

Certaines des clauses limitatives ou exclusives de responsabilité, déterminées en considération du droit anglo-saxon, paraissent excessives au regard du droit français en ce qu'elles visent à exclure la responsabilité du fournisseur de l'œuvre exploitée dans le cas de l'atteinte aux droits de tiers par contrefaçon. Une telle clause est contraire au principe appliqué par les pays de droit civil, selon lequel le donneur de licence garantit une jouissance paisible au licencié : il garantit qu'il est le titulaire de l'ensemble des droits en jeu, qu'il n'a pas noué de relation contractuelle l'empêchant d'en disposer, et que l'œuvre offerte ne porte pas atteinte aux droits de tiers : diffamation, vie privée, droit à l'image...

Selon les articles 1626 et 1627 du code civil :
Art. 1626 : Quoique lors de la vente il n'ait été fait aucune stipulation sur la garantie, le vendeur est obligé de droit à garantir l'acquéreur de l'éviction qu'il souffre dans la totalité ou partie de l'objet vendu, ou des charges prétendues sur cet objet, et non déclarées lors de la vente.
Art. 1627 : Les parties peuvent, par des conventions particulières, ajouter à cette obligation de droit ou en diminuer l'effet ; elles peuvent même convenir que le vendeur ne sera soumis à aucune garantie.

L'article X du contrat régissant les rapports d'ELDA à ses fournisseurs stipule ainsi que :
« Le fournisseur garantit le droit d'exploiter la licence de distribution faisant l'objet du présent contrat. »
« Le fournisseur garantit au distributeur l'exercice paisible des droits cédés. Il le garantit contre tous troubles, revendications et évictions quelconques. Il le défend contre toutes les atteintes qui seraient portées à jouissance paisible des droits cédés. »
« Le contrat de licence CeCILL écarte expressément toute garantie du concédant en cas d'action en contrefaçon qui serait intentée par des tiers à l'encontre du licencié (article 9.3). Cette clause ne pose pas de difficulté dans la mesure où la validité des clauses écartant la garantie d'éviction du fait des tiers est reconnue par les tribunaux, à condition que le concédant soit de bonne foi, c'est-à-dire qu'il n'ait pas eu connaissance du risque d'éviction, ce que la jurisprudence admet rarement lorsqu'il s'agit d'un professionnel tel que le sont souvent les auteurs de contributions réalisées dans des programmes libres. »

De même généralement définies en considération du droit anglo-saxon, les clauses limitatives ou exclusives de garantie constituent le rejet exprès de garanties qui sont, en application du Uniform Commercial Code, présumées être implicitement accordées – mais peuvent être contractuellement écartées entre professionnels : la garantie de conformité ou de bon fonctionnement (« Implied Warranty ») est ainsi limitée au minimum d'une conformité à la documentation ; toute garantie de performance du logiciel ou d'adéquation à un usage déterminé (« Fitness for a Particular Purpose ») est par ailleurs exclue.

Est ainsi commune aux licences libres de distribution de logiciels la clause selon laquelle le produit est fourni « en l'état », sans garantie de fonctionnement non préjudiciable ni d'adéquation à un usage déterminé. Une telle clause se retrouve dans la pratique actuelle des contrats de fourniture et distribution utilisés par ELDA : ainsi, selon l'article 10 du contrat « distributeur », « Le distributeur reconnaît accepter les RL "telles quelles" c'est-à-dire avec tous les défauts qui peuvent subsister, et sans aucune garantie de l'adéquation de telles ressources pour un usage particulier. »

L'application de licences libres ne remettrait donc pas en cause, à cet égard, les pratiques actuelles. De plus, quoiqu'elle constitue une contrepartie logique des libertés accordées et de la multiplicité potentielle des versions

diffusées, cette exclusion de garantie et de responsabilité commune aux modèles types de licences libres ne correspond pas, pour le donneur de licence, à un principe contraignant : il lui est loisible – notamment à titre commercial et ce même dans le cas d'une licence prévoyant une distribution à titre gratuit – de fournir une garantie dans le cadre d'une convention accessoire, en sus des conditions de base prévues par la licence libre. De même, peut être conclu à titre accessoire un contrat de maintenance et de mise à jour.

Une clause de fourniture « en l'état » aura normalement pour contrepartie la reconnaissance au profit de l'acquéreur d'un droit d'accès et évaluation. Par ailleurs, une telle clause n'écarte pas l'obligation de conseil qui peut s'imposer fournisseur vis-à-vis de l'acquéreur.

« Fondée sur une idée de loyauté et de justice, l'obligation de conseil permet de rétablir entre les parties l'égalité trop souvent rompue par la supériorité technique ou économique de l'une des parties. L'obligation de conseil se distingue de l'obligation d'information et de mise en garde. Plus qu'une indication, le conseil implique une incitation, une recommandation, une orientation de choix, une préconisation de la solution la plus adaptée aux besoins exprimés par le client. » « Ainsi, un arrêt récent de la Cour de Cassation du 3 avril 2002 a retenu un manquement à l'obligation de conseil de la part d'un fournisseur qui a livré du matériel informatique incompatible avec le logiciel de traitement de texte utilisé dans l'entreprise. »[1]

3.5.3.2 Obligations de mention des modifications et de leurs auteurs

Les licences libres prévoient fréquemment, sous le nom 'attribution', l'obligation de mentionner tant les auteurs originaux que les auteurs d'apports ultérieurs et d'associer ces mentions, dans l'œuvre et/ou dans les documents devant être fournis avec, aux contributions respectives de ces auteurs. Outre un respect des droits moraux reconnus à l'auteur par le droit français (droit de paternité), une telle obligation d'attribution présente une importance majeure en ce qu'elle constitue, vis-à-vis de produits librement modifiés et redistribués, un fondement essentiel de l'identification des versions rencontrées et de leur contenu. Une certaine garantie de fait est ainsi apportée aux utilisateurs par la possibilité de communiquer, notamment entre utilisateurs, sur les expériences d'utilisation, les apports et limites des différentes versions.

Compte tenu de l'étendue de la diffusion à laquelle elles conduisent normalement, la question d'un contrôle de l'exploitation des ressources linguistiques sous licence libre soulève une problématique de traçabilité des modifications et distributions. La mention de l'identité des auteurs de contributions sera-t-elle ainsi suffisante pour identifier, le cas échéant, l'auteur d'une contribution portant atteinte aux droits de tiers ?

[1] Mascré Heguy Associés, « l'obligation de conseil dans les contrats », http://www.mascre-heguy.com/htm/fr/conseils/conseil_obligation_conseil_contrats.htm

4. Aspects stratégiques

Il est important de faire prendre conscience aux acteurs du domaine de l'importance de partager leur expertise, pour faire avancer l'ensemble de la communauté vers l'innovation technologique. Par ailleurs, il est d'autant plus important d'éviter de répéter les mêmes efforts sur des ressources similaires et optimiser la productivité des organismes du domaine sur un même chemin d'évolution. En particulier, nous souhaitons mettre ici en évidence le concept BLARK (Basic Language Resource Kit), né d'une initiative conjointe entre ELSNET (European Network of Excellence in Language and Speech) et ELRA/ELDA, et qui a comme but de définir un ensemble minimum de ressources linguistiques nécessaires au développement des technologies de la langue et de combler les manques identifiés dans ce domaine. Par ailleurs, le travail de groupe au détriment du travail individuel semble de plus en plus une option se présentant aux différents acteurs comme une option de choix. En parallèle avec ce concept de BLARK, le LDC a développé le concept de « Less Commonly Taught Languages ». L'objectif de ce projet est de créer et de partager des ressources pour soutenir la recherche fondamentale et les premiers développements de la technologie dans ce que l'on a appelé « les langues minoritaires » (langues appelées à faible densité, pas pour la population de locuteurs natifs, mais plutôt pour la rareté des ressources).

Le travail de réalisation jusqu'à la mise à disposition effective d'une ressource linguistique, comme nous l'avons précisé dans la section sur les aspects techniques requiert bon nombre d'étapes, pouvant être résumées par le schéma suivant :

A chacune de ces étapes, il est bon de noter que de nombreuses informations sont dores et déjà existantes et, combinées aux expertises diverses, celles-ci peuvent être employées afin de réduire les coûts nécessaires à la bonne marche de cette chaîne.

En termes de spécifications, de production et de validation de ressources, nous avons déjà vu dans la section sur les aspects techniques que nombreux sont les standards existants et que pour toute nouvelle ressources l'on peut faire appel à des groupes de travaux bien spécifiques.

En ce qui concerne la distribution et la maintenance, nous verrons ci-après que des expertises existent également.

4.1 Synergie et partage des ressources linguistiques

L'un des pères fondateurs de l'association ELRA, Antonio Zampolli, a été l'un des premiers à mettre en exergue le besoin d'associer les différentes branches de l'ingénierie linguistique que sont à la fois les ressources linguistiques, les technologies linguistiques et les projets applicatifs (Maegaard et al., 2005). Antonio Zampolli a notamment introduit le besoin de bénéficier d'une solide infrastructure autour des ressources linguistiques pour permettre une meilleure synergie et coordination des travaux dans le domaine des technologies de la langue.

Par ailleurs, les organismes de financements ont eux-mêmes constaté ce besoin de synergie dans les projets traitant des technologies de la langue. Par exemple, la Commission européenne a pris conscience que des financements étaient demandés régulièrement pour produire les mêmes ressources linguistiques que dans d'autres projets, par le simple fait que les précédents projets étaient terminés et que les ressources linguistiques produites restaient inexploitées, voire inexploitables de par les formats proposés.

Ainsi, les organismes de financements insistent de façon de plus en plus appuyée pour obtenir des justifications très fortes sur ce qui ressortira du projet une fois les financements terminés, ce que la Commission européenne appelle « Exit Strategy ». Cette demande de justification a été reprise au niveau national. En particulier, il est intéressant de mentionner le programme Technolangue[1], action commune aux trois réseaux de recherche et d'innovation technologique (RNRT, RNTL, RIAM), financée dans un cadre interministériel (Ministère chargé de la Recherche et des Nouvelles Technologies, le Ministère délégué à l'Industrie et le Ministère de la Culture et de la Communication). L'objectif principal de ce programme était en effet de mettre en place de manière pérenne une infrastructure de production et diffusion de ressources linguistiques, d'évaluation des technologies de la langue écrite et orale, de participation aux instances nationales et internationales de normalisation et de

[1] Portail des technologies de la langue : http://www.technolangue.net

standardisation et de veille informationnelle sur le domaine. Les ministères financeurs avaient exigé de la part des différents partenaires de projets de :

1) proposer des projets visant à combler les manques dans le domaine des technologies de la langue
2) rendre possible l'exploitation des résultats au-delà du projet financé
3) travailler en synergie pour apporter la meilleure expertise possible : partage des savoir-faire, partage des ressources, ou du personnel entre organismes publics et privés, nationaux et internationaux.

Plus concrètement nous pouvons mentionner les travaux réalisés dans les différentes campagnes d'évaluation EVALDA financées dans le cadre du programme Technolangue. De nombreuses synergies se sont traduites à l'intérieur d'EVALDA, où des ressources créées pour une campagne ont pu être réutilisées ou adaptées pour d'autres campagnes. Des synergies ont été également rendues possibles à l'extérieur d'EVALDA, notamment avec des projets européens, et la réutilisation d'outils logiciels.

4.1.1 Expérience du marché et des acteurs des ressources linguistiques

S'engager sur le marché des ressources linguistiques nécessite une expertise du marché. Des organismes spécialisés dans les ressources linguistiques sont nés en Europe et dans le monde pour pallier aux différents besoins d'identification, de production et d'échanges de ressources linguistiques.

Deux centres sont désormais incontournables dans ce domaine et reconnus internationalement :

- ELRA (*European Association for Language Resources* - Association européenne pour les ressources linguistiques)
- LDC (*Linguistic Data Consortium*)

ELRA et LDC, de par leur implication dans le monde des technologies de la langue, détiennent une vision d'ensemble du marché des ressources linguistiques. Leurs actions autour de la mise à disposition des ressources linguistiques sont présentées ci-après.

En Asie, le GSK[1] (*Gengo Shigen Kyouyuukikou* – Consortium pour les ressources linguistiques), au Japon, et le SITEC[2] (*Speech Information Technology & Industry Promotion Center* – Centre pour la technologie de l'information parlée et la promotion industrielle) en Corée, sont en train de prendre le pas de l'occident et de constituer de nouveaux centres pour couvrir les besoins du territoire asiatique.

4.1.1.1 ELRA / ELDA[3]

ELRA (Association européenne pour les ressources linguistiques), et son organisme opérationnel ELDA (Agence pour l'évaluation et la distribution des ressources linguistiques) ont été impliqués dans différents travaux d'identification à la fois des acteurs du domaine des technologies de la langue et des ressources linguistiques existantes. ELRA a pu mettre ces différentes informations à disposition de la communauté par divers biais et notamment :

- Un Catalogue de Ressources Linguistiques[4] : ce catalogue présente des ressources « packagées » dont les droits de distribution ont été négociés par ELDA pour le compte d'ELRA et à destination de la communauté des technologies de la langue. Ces ressources proviennent à la fois de producteurs individuels que de consortiums ou sont le résultat de projets financés par des organes tels que la Commission européenne ou les organes de financements nationaux. Ce catalogue comprend également des ressources intégralement produites par ELDA dans le cadre de divers projets nationaux ou internationaux.
- Un « Catalogue Universel » : ELRA propose à ses membres un « Catalogue universel » qui a pour but de recenser les ressources linguistiques existant partout dans le monde (qu'elles soient disponibles ou non).

[1] GSK (site en construction) : http://www.gsk.or.jp/index_e.html

[2] SITEC : http://www.sitec.or.kr

[3] ELRA : www.elra.info et ELDA : www.elda.org

[4] Catalogue ELRA : http://catalogue.elra.info

4.1.1.2 LDC[1]

Depuis sa création en 1992, le LDC (Linguistic Database Consortium) soutient le domaine de la recherche et du développement de technologies et le partage de ressources linguistiques (données, outils et standards). Le LDC a tout d'abord travaillé de manière étroite avec le gouvernement américain pour produire des ressources linguistiques répondant aux besoins des acteurs du domaine. Tout comme ELRA, le LDC dispose d'un catalogue de ressources linguistiques[2] qui recense à la fois les ressources produites sous un financement gouvernemental, ainsi que toute autre ressource produite par un acteur externe souhaitant la partager avec le reste de la communauté.

4.1.2 Le concept BLARK

Le concept BLARK (pour *Basic LAnguage Resource Kit* – Kit de base de ressources linguistiques) a d'abord été défini en Hollande. L'une des premières références au BLARK a été faite notamment dans un article rédigé par Steven Krauwer (Krauwer, 1998), faisant suite à une proposition de coopération commune entre ELSNET (European Network of Excellence in Language and Speech) et ELRA dans le cadre du 5e Programme-Cadre de la Commission européenne. L'action proposée, qui n'a cependant pas pu être soumise dans les temps au 5e PCRD, était présentée suivant trois étapes :
1) Définir le BLARK, c'est-à-dire définir pour chaque langue les instruments de base (ressources linguistiques, outils de manipulation de ressources et compétences) nécessaires à la bonne mise en œuvre d'un travail de recherche pré-concurrentiel sur la langue.
2) Identifier les collections de données existantes, les outils, voire les cours dispensés pour chaque langue (y compris les aspects multilingues et interlingues).
3) Lancer des actions coordonnées pour combler les manques identifiés.

Ce concept a pu être mis en œuvre pour la première fois dans le cadre de l'initiative « Dutch Human Language Technologies Platform » (plate-forme pour les technologies de la langue hollandaise), lancée en avril 1999 par le Dutch Language Union, un organisme inter-gouvernemental en charge de renforcer la situation de la langue hollandaise (Cucchiarini et al., 2001a) et (Cucchiarini et al., 2001b).

Cette initiative a permis de définir les actions nécessaires pour la langue hollandaise :
- définir une liste des ressources devant être développées de façon prioritaire pour le hollandais,
- définir les critères que devront remplir ces ressources de base,
- rédiger un livre blanc pour la gestion, la maintenance, la mise à disposition et la distribution des ressources de base pouvant servir à l'usage de l'éducation et de la recherche, ainsi que pour le développement d'outils et applications des technologies de la langue.

Plus récemment, la notion de BLARK a été adaptée à la langue arabe, dans le cadre du projet NEMLAR (Network for Euro-Mediterranean LAnguage Resources). NEMLAR a été soutenu par le programme INCO-MED (février 2003 à juillet 2005). Il avait pour objectif de mettre en place un réseau de partenaires qualifiés dans la zone euro-méditerranéenne pour lancer et soutenir le développement de ressources linguistiques de premier ordre pour la langue arabe et d'autres langues méditerranéennes. Le projet s'est attaché à dresser l'état de l'art des ressources linguistiques dans cette région, d'une part en identifiant, en collaboration avec les acteurs du domaines et les industries, quelles peuvent être les priorités en la matière, et d'autre part en établissant un protocole pour la production des ressources de base pour les langues principales parlées dans cette région (Krauwer et al., 2006). Depuis le 1er février 2008, et ce jusqu'au 1er août 2010 , un nouveau projet, MEDAR (Mediterranean Arabic Language and Speech Technology) a pour mission de prolonger le travail réalisé dans NEMLAR et d'encourager une collaboration internationale pour répondre aux besoins des systèmes traitant de la langue arabe.

4.1.2.1 Les matrices BLARK

ELDA a développé un service interactif BLARK[3] permettant d'identifier les différents besoins en termes de ressources linguistiques vis-à-vis d'applications spécifiques et ce pour autant de langues possibles. Suite à sa propre expérience et aux différents rapports issus des autres initiatives telles que celle hollandaise, ELDA a implémenté et augmenté sa matrice d'origine qui consistait en un croisement entre les langues et les différents types de ressources qu'elle avait pu identifier.

---

[1] LDC : www.ldc.upenn.edu

[2] Catalogue LDC : http://www.ldc.upenn.edu/Catalog

[3] BLARK : http://www.elda.org/blark

Afin de comprendre les besoins plus clairement et de manière plus exhaustive, ELDA a étendu cette matrice à une liste d'applications et de modules potentiels pouvant être mis en relation avec les ressources linguistiques et langues nécessaires. Les matrices qui en résultent viennent en partie des travaux réalisés dans le cadre du projet NEMLAR. Deux matrices (Applications / Modules et Langues / Modules) ont été développées et sont disponibles et modifiables directement depuis les pages web.

En pratique, les matrices BLARK sont divisées en deux tableaux :

Le tableau "Applications/Modules" montre le niveau d'importance des modules nécessaires (ou non) à une application donnée, tant pour les technologies de l'écrit que celles de l'oral, et ce pour une langue donnée : important (✚), très important (✚✚), essentiel (✚✚✚) ou sans importance (🚫).

Figure 1 : Extrait du tableau "Applications/Modules" pour la langue arabe, domaine des technologies orales:

Spoken applications vs **spoken modules** for **Arabic** language

close window	Customization to Different	Dialect/Language	Dictation	Embedded Speech	Emotion Identification	Emotion/Prosody Output	Generation Lips Movement	Lips Movement Reading
Acoustic Models	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚
Dialect/language Identification		✚	✚	✚	✚			✚
Emotion Identification		✚	✚	✚		✚✚		✚
Language Models		✚✚	✚✚✚	✚✚		✚✚		
Lexicon Adaptation			✚	✚				
Lips Movement Reading		✚✚						✚✚✚
Phoneme Alignment			✚	✚				
Pronunciation Lexicon			✚✚✚	✚✚✚				
Prosody Prediction						✚✚✚		
Prosody Recognition		✚	✚	✚	✚✚✚			
Segmenter Speech/silence		✚✚	✚✚	✚✚	✚✚	✚		✚
Sentence Boundary Detection		✚	✚	✚	✚✚	✚✚		✚
Speaker Adaptation		✚	✚✚	✚✚	✚			✚
Speaker Recognition/identification		✚	✚	✚	✚			✚
Speech Units Selection						✚✚✚		
Speech/non-speech Music Detection		✚	✚	✚	✚✚			✚
Word Boundary Identification		✚	✚	✚	✚	✚✚		✚

Le tableau "Ressources/Modules" montre le niveau de nécessité en terme de ressources linguistiques à inclure ou non dans des modules spécifiques, tant pour les technologies de l'écrit que celles de l'oral, et ce pour une langue donnée : important (✚), très important (✚✚), essentiel (✚✚✚) ou sans importance (🚫).

Figure 2 : Extrait du tableau "Langues/Modules" pour la langue arabe, domaine des technologies orales:

Spoken resources vs **spoken modules** for **Arabic** language

close window	Annotated Written Corpus	Audio Data with Prosodic Markers and other	BNSC	Desktop/Microphone & High Quality	Non Vowelised Corpus	Onomastica (proper names)	Phonetic Lexicon	Telephony
Acoustic Models		✣✣✣	✣✣✣	✣✣✣				✣✣✣
Dialect/language Identification		✣	✣✣	✣✣		✣	✣	✣✣
Emotion Identification		✣	✣	✣		✣	✣	✣
Language Models	✣✣				✣✣			
Lexicon Adaptation	✣				✣	✣✣✣	✣✣✣	
Lips Movement Reading								
Phoneme Alignment	✣✣	✣✣	✣✣	✣✣		✣✣✣	✣✣✣	✣✣
Pronunciation Lexicon	✣					✣✣✣	✣✣✣	
Prosody Prediction	✣✣	✣✣				✣✣	✣✣	
Prosody Recognition	✣✣	✣✣✣		✣		✣✣	✣✣	✣
Segmenter Speech/silence		✣✣	✣✣	✣✣				✣✣
Sentence Boundary Detection		✣✣	✣✣	✣✣		✣	✣	✣✣
Speaker Adaptation		✣	✣✣	✣✣				✣✣
Speaker Recognition/identification		✣	✣	✣				✣
Speech Units Selection	✣✣	✣✣✣		✣		✣	✣	✣
Speech/non-speech Music Detection		✣✣	✣✣	✣				✣
Word Boundary Identification		✣	✣	✣		✣	✣	✣

4.1.3 Coopération et groupes de travail pour la production de ressources

De plus en plus, les organismes producteurs de ressources prennent conscience de la nécessité de collaborer avec d'autres organismes afin d'optimiser la qualité et la quantité de ressources linguistiques à produire.

Parmi les collaborations possibles, il a été très rapidement exigé de la part de la commission européenne dans le cadre de ses financements de projets de rapprocher ces deux types d'institutions que sont les institutions publiques et les institutions privées.

Cette motivation s'appuie sur la nécessité de faire avancer en parallèle les travaux de recherche avec les travaux à objectifs purement commerciaux.

Au-delà des projets de financement classique, on a constaté l'augmentation du nombre de groupes de travail constitués autour du thème des ressources linguistiques.

Il est ici difficile de présenter une liste exhaustive des groupes de travail impliqués dans la production de ressources linguistiques. Cependant, il est primordial de noter la constitution, voire la longévité de certains consortiums et groupes de travail ou autres réseaux d'excellence au niveau européen et mondial. Nous présenterons dans cette section les axes d'intérêts récemment visés par ces différents consortiums/groupes.

Depuis plusieurs années déjà, un groupement de spécialistes des ressources orales (avec un intérêt plus récent pour les ressources multimédia/multimodales) a été créé afin d'encourager et promouvoir l'interaction internationale autour du traitement de la langue orale : **COCOSDA**[1] (International **Co**mmittee for the **Co**-ordination and **S**tandardisation of Speech **D**atabases and **A**ssessment Techniques – *Comité international pour la coordination et la standardisation de bases de données de parole et des techniques de validation*). COCOSDA a pour stratégie de promouvoir le développement de corpus de parole ayant pour but de permettre la production ou l'évaluation des technologies de la parole tant actuelles que futures. Sa formation permet également d'insister sur la coordination des projets au niveau international et de la mutualisation des efforts de recherche pour une meilleure efficacité. COCOSDA est distribué au niveau mondial grâce à un découpage en six régions : Afrique, Asie, Europe, Amérique latine, Amérique du nord et Océanie.

COCOSDA se distingue notamment par l'organisation régulière d'ateliers de travail permettant d'établir un état des lieux régulier des dernières actualités du domaine, le dernier en date s'étant tenu lors de la conférence LREC 2008 (Cocosda/WRITE Workshop, Marrakech, 1er juin 2008) en conjonction avec le groupement spécialisé dans les ressources écrites WRITE (voir plus bas).

Au-delà de COCOSDA, plusieurs groupements se sont formés par spécialités. On peut noter notamment le consortium LILA qui, grâce à la constitution d'un consortium de différentes sociétés privées impliquées dans la technologie de la reconnaissance automatique de la parole, a pu produire une série de ressources linguistiques pour les langues d'Asie. Le consortium **LILA** n'aurait pu être réalisé sans la constitution d'un réseau d'institutions intéressées par la reconnaissance automatique de la parole autour de la série de projets **SpeechDat** (voir la section Ressources Orales dans la partie Aspects Techniques).

On peut également citer le projet **NEMLAR**[2] (Network for Euro-Mediterranean LAnguage Resources – *Réseau pour les ressources linguistiques d'Europe et de Méditerranée*) dont l'objectif a été de constituer un réseau de partenaires qualifiés en Europe et autour de la Méditerranée, afin de réfléchir et soutenir le développement de ressources hautement prioritaires pour la langue arabe et autres langues locales dans un contexte systématique, standardisé et collaboratif. Ce projet a notamment permis de produire des ressources tant orales qu'écrites. Ce projet s'est terminé en 2005 mais a pu être prolongé via un nouveau projet, **MEDAR** (Mediterranean Arabic Language and Speech Technology – *Langue arabe et technologie de la parole en Méditerranée*) jusqu'en 2010.

Sous l'influence du groupe de travail COCOSDA pour la parole, plusieurs lancements de groupes de travail sur le thème des ressources linguistiques écrites ont été réalisés. Dans le cadre du projet ENABLER, financé par la Commission européenne (2001-2002), une première proposition de groupement de spécialistes des ressources écrites a été lancée sous le nom de **ICCWLR** (International Coordination Committee for Written Language Resources and Evaluation). L'objectif premier d'ICCWLR était tout comme COCOSDA d'offrir un forum international pour l'encouragement et la soutien de travaux de coordination et de coopération dans le domaine des technologies de l'écrit. ICCWLR est rapidement devenu **WRITE**[3] (Written Resources Infrastructure Technology and Evaluation). L'une des premières réalisations de WRITE a notamment été la rédaction d'une feuille de route présentant les besoins en termes de ressources linguistiques, lors d'un atelier de la conférence LREC2004 organisé conjointement avec COCOSDA.

C'est également sous l'influx de WRITE et COCOSDA qu'a été proposé le projet **FlaReNet**[4] (Fostering Language Resources Network – *Réseau pour la promotion des ressources linguistiques*) qui démarrera le 1er septembre 2008. Les activités principales de FlaReNet visent à élaborer des études et analyses sur les ressources linguistiques et standards correspondants, ainsi que leurs modèles organisationnels et économiques, et de lancer des discussions avec les acteurs économiques du domaine pour un meilleur déploiement et une meilleure utilisation des ressources dans des produits du monde réel.

Du côté des normes et standards, il est important d'observer les travaux du comité technique **TC37 d'ISO**, dans le cadre du sous-comité **SC4**[5] dédié à la gestion des ressources linguistiques. L'objectif de ce sous-comité, lancé officiellement en 2002, est focalisé sur les ressources linguistiques, et vise à mettre en place des standards et méthodologies pour la création, le codage, le traitement et la gestion des ressources linguistiques. L'intervention d'ISO se fait dans un cadre industriel international afin d'y assurer un développement des standards et une production des ressources linguistiques adaptées aux applications réelles, mais également de permettre le développement d'applications basées sur des méthodes et outils standardisés. Le SC4 est distribué en cinq

[1] COCOSDA : http://www.cocosda.org

[2] NEMLAR/MEDAR : http://www.nemlar.org

[3] WRITE: http://www.ilc.cnr.it/write

[4] FlaReNet: http://www.ilc.cnr.it/flarenet

[5] ISO TC37 SC4: http://www.tc37sc4.org

groupes de travail (WG – Working Groups) :

- WG1 Descripteurs et mécanismes de base pour les ressources linguistiques
- WG2 Schémas de représentation
- WG3 Représentation des textes multilingues
- WG4 Base de données lexicale
- WG5 Workflow de la gestion des ressources linguistiques

Enfin, pour ce qui est de la diffusion des ressources linguistiques, le consortium **CLARIN**[1] (Common Language Resources and Technology INfrastructure) a lancé ses activités le 1er janvier 2008 pour une durée de 3 ans. Constitué dans le cadre du programme FP7 de la Commission européenne (Capacities Specific Programme - Research Infrastructures), CLARIN a pour objectif de créer, coordonner et rendre les ressources linguistiques et technologies affiliées disponibles et réutilisables sur la base d'une communauté collaborative paneuropéenne. Les actions de CLARIN sont principalement adressées au domaine des sciences humaines et sociales afin d'offrir aux scientifiques et à leurs étudiants des outils permettant un traitement assisté de la langue (conduits de contenu culturel et de connaissance, instruments de communication, etc.).

4.2 Diffusion/Capitalisation des ressources linguistiques

Même si l'objectif d'un organisme (organisme commercial ou laboratoire universitaire) n'est pas de diffuser les ressources linguistiques, il est intéressant de pouvoir les rendre disponibles à l'ensemble de la communauté pour diverses raisons:

- Participation à l'évolution du marché
- Echanges d'information
- Retour sur investissement
- Valorisation des travaux

Dans cette partie, nous présenterons les différents avantages pour chaque type d'organisme que peuvent apporter la diffusion de ses ressources et travaux affiliés à l'extérieur d'un cadre strict de projet interne.

L'activité de diffusion de ressources linguistiques requiert l'organisation de différents services. Il est important pour tout organisme souhaitant s'occuper de la diffusion de ressources linguistiques de connaître l'ampleur des tâches à réaliser. Nous présenterons ci-dessous les tâches à prendre en compte pour une bonne diffusion de ressources linguistiques. Face à l'ampleur de la tâche, il est souvent difficile pour un organisme de passer à cette étape, principalement par le simple fait que son activité principale n'est pas en premier lieu la diffusion des ressources. C'est pourquoi il existe des organismes spécialisés dans ces activités. Nous présenterons plus loin la base de leurs modus operandi.

4.2.1 Etapes nécessaires à la diffusion de ressources linguistiques

- **Archivage des ressources :**

Cette première étape est d'une grande importance. En effet, le risque zéro de perdre la ressource linguistique n'existe pas et seul un archivage systématique des données peut éviter leur perte définitive. Au-delà du risque de perdre des données, il est important de savoir à tout moment à quel endroit physique les retrouver. En effet, entre deux diffusions, il se peut que beaucoup de temps se passe, ainsi que la connaissance de celui qui les a archivées la première fois.
Par ailleurs, nous savons que l'évolution technologique peut amener à changer les supports de diffusion. Il est donc important non seulement de conserver les données mais également de conserver le matériel permettant de les réutiliser.

- **Information détaillée sur les ressources à fournir sur demande :**

Comme nous l'avons vu dans la section 2.1 sur les formats de description de ressources, il est en effet nécessaire de fournir à tout utilisateur potentiel d'une ressource donnée, l'ensemble des informations nécessaires à l'utilisateur pour se faire une idée la plus précise possible du contenu de la ressource. Dans la section 2.2, nous avons vu un certain nombre de bonnes pratiques pour la description des ressources, pouvant être utilisées par tout fournisseur/producteur pour décrire ses ressources au mieux.

- **Etablissement d'accords entre le fournisseur et l'utilisateur :**

[1] CLARIN: http://www.clarin.eu

Avant de pouvoir fournir des données entre un producteur ou fournisseur de ressources et un utilisateur, les aspects juridiques, tels que définis dans la section 3, doivent être pris en compte. En effet, pour chaque diffusion, le fournisseur de ressources devra se mettre d'accord, de façon légale, avec l'utilisateur pour la transmission de la ressource.

- **Fixation des prix :**

Il est souvent difficile pour un producteur de ressources de fixer le « juste » prix d'une ressource linguistique. En effet, la diversité des ressources en termes de contenu et de qualité complexifie la fixation des prix. Souvent, le choix du prix par un producteur se basera sur les coûts de production ou encore en fonction de ses relations privilégiées ou non avec l'utilisateur. On remarquera également la différence de considération vis-à-vis des aspects financiers entre un producteur privé et un organisme de recherche, le premier cherchant plutôt à obtenir un bénéfice financier, le second optant plutôt pour le bénéfice technique.

- **Préparation des ressources :**

Pour toute diffusion, la phase de préparation des ressources n'est pas non plus aisée. En effet, il est nécessaire d'utiliser un matériel technologique adéquat pour dupliquer les ressources archivées et les fournir sur un support correspondant à l'attente de l'utilisateur. Actuellement, les supports électroniques utilisés restent principalement le CD-ROM, DVD-ROM ou le disque dur, sachant que chacun des 3 supports requiert un matériel de préparation correspondant (matériel de gravure ou de copiage, étiquetage, …).

- **Fourniture des ressources :**

Enfin, la fourniture des ressources doit être également assurée. Ici, il faut considérer les coûts de livraison, le matériel d'empaquetage, les délais de livraison, les garanties/assurances, etc.

4.2.2 Diffusion des ressources linguistiques par des organismes spécialisés

Les étapes mentionnées précédemment permettant d'assurer une bonne diffusion de ressources linguistiques sont autant de contraintes pour un organisme dont l'objectif principal n'est pas centré sur la distribution de ressources linguistiques. C'est pourquoi des organismes spécialisés ont été créés et prennent en compte toutes ces considérations.

Encore une fois, deux organismes majeurs sont à mentionner : ELRA et LDC. Même si les modes de fonctionnement d'ELRA et LDC restent différents, les options proposées aux producteurs pour la diffusion de leurs ressources linguistiques peuvent être résumées comme suit.

Les catalogues ELRA et LDC sont ouverts à tous types de fournisseurs de ressources linguistiques, que ce soit des institutions de recherche ou organismes commerciaux, ou encore tout fournisseur individuel.

Les principaux requis pour les ressources fournies concernent la réutilisabilité des données. Cela implique la fourniture des ressources sur des supports standards, avec une documentation minimale permettant à tout utilisateur externe de réutiliser les données sans support externe et avec un minimum de post-traitement. Par exemple, LDC demande à chaque fournisseur, lors de sa livraison de ressources un détail du contenu des ressources sous un format spécifique:

1. Un répertoire *doc*, contenant la documentation accompagnant les données
2. Un répertoire *data*, contenant les données, qu'elles soient constituées de données orales ou textuelles
3. Lorsque nécessaire, des répertoires supplémentaires, tels qu'un repertoire *dtd* contenant des schémas de fichiers spécifiques aux données

Pour ELRA, comme déjà développé en section 2.2, une description minimale doit être fournie, cette description permettant à ELRA et tout utilisateur de se faire une idée a priori du contenu des données. Des échantillons complémentaires peuvent ensuite être fournis avant toute acquisition par un tiers. Par ailleurs, ELRA demande à ses fournisseurs de compléter un formulaire, basé sur le système de méta-données expliqué en Section 2.2, et permettant de décrire de la façon la plus détaillée possible le contenu des ressources. Par exemple, pour un corpus écrit, cela comprend :

1. Des informations générales, telles que le type de corpus, la date de création, les applications visées par l'utilisation de la ressource
2. Les sources utilisées, telles que l'origine des documents électroniques utilisés pour l'annotation (par exemple, les références d'un magazine, journal, livre, etc.), le domaine, la taille en nombre de mots.
3. Le type d'annotation réalisé sur le corpus
4. Des informations techniques, telles que les formats de fichiers, les standards utilisés, les jeux de caractères.

En termes de qualité, ELRA propose un service de validation à la demande permettant d'apporter une certification supplémentaire sur les ressources distribuées (voir section 2.1 sur les tâches de validation).

Une fois toutes ces informations obtenues de la part du fournisseur, un accord officiel du fournisseur doit être obtenu. Ainsi, c'est la phase contractuelle entre le distributeur et le fournisseur qui doit être mise en œuvre. ELRA et LDC utilisent tous deux des contrats de distribution rédigés tout spécialement pour la diffusion de ressources linguistiques. Ceux-ci peuvent être obtenus directement depuis leurs sites web respectifs.

Lorsque tous les détails techniques et contractuels sont finalisés, l'un des grands intérêts d'employer un organisme spécialisé comme ELRA ou LDC en tant qu'intermédiaire pour la diffusion de ressources linguistiques est que le suivi de la diffusion est réalisé par un organisme externe et permet donc au fournisseur de s'investir dans d'autres affaires plus propres à son objectif professionnel.

5. Annexes

5.1 Lesser General Public License for Linguistic Resources

Preamble

The licenses for most data are designed to take away your freedom to share and change it. By contrast, this License is intended to guarantee your freedom to share and change free data--to make sure the data are free for all their users.

This license, the Lesser General Public License for Linguistic Resources, applies to some specially designated linguistic resources -- typically lexicons, grammars, thesauri and textual corpora.

TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License Agreement applies to any Linguistic Resource which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Lesser General Public License for Linguistic Resources (also called "this License"). Each licensee is addressed as "you".

A "linguistic resource" means a collection of data about language prepared so as to be used with application programs.

The "Linguistic Resource", below, refers to any such work which has been distributed under these terms. A "work based on the Linguistic Resource" means either the Linguistic Resource or any derivative work under copyright law: that is to say, a work containing the Linguistic Resource or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Legible form" for a linguistic resource means the preferred form of the resource for making modifications to it.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Linguistic Resource is not restricted, and output from such a program is covered only if its contents constitute a work based on the Linguistic Resource (independent of the use of the Linguistic Resource in a tool for writing it). Whether that is true depends on what the program that uses the Linguistic Resource does.

1. You may copy and distribute verbatim copies of the Linguistic Resource as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Linguistic Resource.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Linguistic Resource or any portion of it, thus forming a work based on the Linguistic Resource, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- **a)** The modified work must itself be a linguistic resource.
- **b)** You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
- **c)** You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.

  These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Linguistic Resource, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Linguistic Resource, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Linguistic Resource.

In addition, mere aggregation of another work not based on the Linguistic Resource with the Linguistic Resource (or with a work based on the Linguistic Resource) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. A program that contains no derivative of any portion of the Linguistic Resource, but is designed to work with the Linguistic Resource (or an encrypted form of the Linguistic Resource) by reading it or being compiled or linked with it, is called a "work that uses the Linguistic Resource". Such a work, in isolation, is not a derivative work of the Linguistic Resource, and therefore falls outside the scope of this License.

However, combining a "work that uses the Linguistic Resource" with the Linguistic Resource (or an encrypted form of the Linguistic Resource) creates a package that is a derivative of the Linguistic Resource (because it contains portions of the Linguistic Resource), rather than a "work that uses the Linguistic Resource". If the package is a derivative of the Linguistic Resource, you may distribute the package under the terms of Section 4. Any works containing that package also fall under Section 4.

4. As an exception to the Sections above, you may also combine a "work that uses the Linguistic Resource" with the Linguistic Resource (or an encrypted form of the Linguistic Resource) to produce a package containing portions of the Linguistic Resource, and distribute that package under terms of your choice, provided that the terms permit modification of the package for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the package that the Linguistic Resource is used in it and that the Linguistic Resource and its use are covered by this License. You must supply a copy of this License. If the package during execution displays copyright notices, you must include the copyright notice for the Linguistic Resource among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- **a)** Accompany the package with the complete corresponding machine-readable legible form of the Linguistic Resource including whatever changes were used in the package (which must be distributed under Sections 1 and 2 above); and, if the package contains an encrypted form of the Linguistic Resource, with the complete machine-readable "work that uses the Linguistic Resource", as object code and/or source code, so that the user can modify the Linguistic Resource and then encrypt it to produce a modified package containing the modified Linguistic Resource.

- **b)** Use a suitable mechanism for combining with the Linguistic Resource. A suitable mechanism is one that will operate properly with a modified version of the Linguistic Resource, if the user installs one, as long as the modified version is interface-compatible with the version that the package was made with.

- **c)** Accompany the package with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 4a, above, for a charge no more than the cost of performing this distribution.

- **d)** If distribution of the package is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.

- **e)** Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

If the package includes an encrypted form of the Linguistic Resource, the required form of the "work that uses the Linguistic Resource" must include any data and utility programs needed for reproducing the package from it. However, as a special exception, the materials to be distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Linguistic Resource together in a package that you distribute.

5. You may not copy, modify, sublicense, link with, or distribute the Linguistic Resource except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Linguistic Resource is void, and will automatically terminate your rights under this License. However, parties

who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

6. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Linguistic Resource or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Linguistic Resource (or any work based on the Linguistic Resource), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Linguistic Resource or works based on it.

7. Each time you redistribute the Linguistic Resource (or any work based on the Linguistic Resource), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Linguistic Resource subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties with this License.

8. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Linguistic Resource at all. For example, if a patent license would not permit royalty-free redistribution of the Linguistic Resource by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Linguistic Resource.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free resource distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of data distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute resources through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

9. If the distribution and/or use of the Linguistic Resource is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Linguistic Resource under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

10. The Free Software Foundation may publish revised and/or new versions of the Lesser General Public License for Linguistic Resources from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Linguistic Resource specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Linguistic Resource does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

11. If you wish to incorporate parts of the Linguistic Resource into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission.

NO WARRANTY

12. BECAUSE THE LINGUISTIC RESOURCE IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LINGUISTIC RESOURCE, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LINGUISTIC RESOURCE "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LINGUISTIC RESOURCE IS WITH YOU.

SHOULD THE LINGUISTIC RESOURCE PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

13. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LINGUISTIC RESOURCE AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LINGUISTIC RESOURCE (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LINGUISTIC RESOURCE TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

5.2 GNU General Public License

Version 3, 29 June 2007

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program--to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the GNU GPL protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the GPL clearly explains that there is no warranty for this free software. For both users' and authors' sake, the GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the GPL to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the GPL, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the GPL assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

TERMS AND CONDITIONS

0. Definitions.

“This License” refers to version 3 of the GNU General Public License.

“Copyright” also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

“The Program” refers to any copyrightable work licensed under this License. Each licensee is addressed as “you”. “Licensees” and “recipients” may be individuals or organizations.

To “modify” a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a “modified version” of the earlier work or a work “based on” the earlier work.

A “covered work” means either the unmodified Program or a work based on the Program.

To “propagate” a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To “convey” a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays “Appropriate Legal Notices” to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

1. Source Code.

The “source code” for a work means the preferred form of the work for making modifications to it. “Object code” means any non-source form of a work.

A “Standard Interface” means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The “System Libraries” of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and (b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A “Major Component”, in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The “Corresponding Source” for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work's System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding Source.

The Corresponding Source for a work in source code form is that same work.

2. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works for you must do so exclusively on your behalf, under your

direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

3. Protecting Users' Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

4. Conveying Verbatim Copies.

You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

5. Conveying Modified Source Versions.

You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

- a) The work must carry prominent notices stating that you modified it, and giving a relevant date.
- b) The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to “keep intact all notices”.
- c) You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.
- d) If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so.

A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an “aggregate” if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's users beyond what the individual works permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

6. Conveying Non-Source Forms.

You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

- a) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.

- b) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.

- c) Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.

- d) Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.

- e) Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d.

A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

A “User Product” is either (1) a “consumer product”, which means any tangible personal property which is normally used for personal, family, or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, “normally used” refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

“Installation Information” for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in ROM).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or installed. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

7. Additional Terms.

“Additional permissions” are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though

they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

- a) Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or
- b) Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or
- c) Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or
- d) Limiting the use for publicity purposes of names of licensors or authors of the material; or
- e) Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or
- f) Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors.

All other non-permissive additional terms are considered “further restrictions” within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

8. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License (including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

9. Acceptance Not Required for Having Copies.

You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

10. Automatic Licensing of Downstream Recipients.

Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

An “entity transaction” is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

11. Patents.

A “contributor” is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor's “contributor version”.

A contributor's “essential patent claims” are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, “control” includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a “patent license” is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To “grant” such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party.

If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2) arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. “Knowingly relying” means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient's use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is “discriminatory” if it does not include within the scope of its coverage, prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business of distributing software, under which you make payment to the third party based on the extent of your

activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

12. No Surrender of Others' Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

13. Use with the GNU Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the GNU Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the GNU Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

14. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the GNU General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the GNU General Public License “or any later version” applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the GNU General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the GNU General Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

15. Disclaimer of Warranty.

THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A

FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

17. Interpretation of Sections 15 and 16.

If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively state the exclusion of warranty; and each file should have at least the “copyright” line and a pointer to where the full notice is found.

```
    <one line to give the program's name and a brief idea of what it does.>
    Copyright (C) <year>  <name of author>

    This program is free software: you can redistribute it and/or modify
    it under the terms of the GNU General Public License as published by
    the Free Software Foundation, either version 3 of the License, or
    (at your option) any later version.

    This program is distributed in the hope that it will be useful,
    but WITHOUT ANY WARRANTY; without even the implied warranty of
    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
    GNU General Public License for more details.

    You should have received a copy of the GNU General Public License
    along with this program.  If not, see <http://www.gnu.org/licenses/>.
```

Also add information on how to contact you by electronic and paper mail.

If the program does terminal interaction, make it output a short notice like this when it starts in an interactive mode:

```
    <program>  Copyright (C) <year>  <name of author>
    This program comes with ABSOLUTELY NO WARRANTY; for details type `show
w'.
    This is free software, and you are welcome to redistribute it
    under certain conditions; type `show c' for details.
```

The hypothetical commands `show w' and `show c' should show the appropriate parts of the General Public License. Of course, your program's commands might be different; for a GUI interface, you would use an “about box”.

You should also get your employer (if you work as a programmer) or school, if any, to sign a “copyright disclaimer” for the program, if necessary. For more information on this, and how to apply and follow the GNU GPL, see <http://www.gnu.org/licenses/>.
The GNU General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License instead of this License. But first, please read <http://www.gnu.org/philosophy/why-not-lgpl.html>.

5.3 Système de licence Creative Commons

Les options et les contrats disponibles

Voici les 6 licences disponibles à partir de l'interface "Choisissez votre licence"

Elles sont désignées par leur nom et les icônes représentant les différentes options choisies par l'auteur qui souhaite accorder plus de libertés que le régime minimum du droit d'auteur en informant le public que certaines utilisations sont autorisées à l'avance.

Attribution (by) 

Attribution Share Alike (by-sa) 

Attribution No Derivatives (by-nd) 

Attribution Non-commercial (by-nc) 

Attribution Non-commercial Share Alike (by-nc-sa) 

Attribution Non-commercial No Derivatives (by-nc-nd) 

Traduction des options :

-  **Paternité** : l'oeuvre peut être librement utilisée, à la condition de l'attribuer à son l'auteur en citant son nom.
-  **Pas d'Utilisation Commerciale** : le titulaire de droits peut autoriser tous les types d'utilisation ou au contraire restreindre aux utilisations non commerciales (les utilisations commerciales restant soumises à son autorisation).
-  **Pas de Modification** : le titulaire de droits peut continuer à réserver la faculté de réaliser des oeuvres de type dérivées ou au contraire autoriser à l'avance les modifications, traductions...

 **Partage à l'Identique des Conditions Initiales** : à la possibilité d'autoriser à l'avance les modifications peut se superposer l'obligation pour les oeuvres dites dérivées d'être proposées au public avec les mêmes libertés (sous les mêmes options Creative Commons) que l'oeuvre originaire.

5.4 Contrat de distribution ELRA

ELRA-Agence de Distribution (ELDA)
55-57 rue Brillat Savarin,
F-75013 PARIS, FRANCE
Tel. +33 1 43 13 33 33
Fax: +33 1 43 13 33 30
Email: choukri@elda.org
Web : http://www.elra.info or http://www.elda.org

CONTRAT DE DISTRIBUTION DE RESSOURCES LINGUISTIQUES

ENTRE

"……."

ET

ELRA

(Association Européenne pour les Ressources Linguistiques)

(Référence LC/ELDA/DISTR-S/FR/2007/)

Ce CONTRAT est conclu entre :

"……..", (ci-après dénommé le **FOURNISSEUR**), dont le siège social est situé : ……

ET

ELRA, dont le siège social est situé : 46, Grand-Rue, Luxembourg 1660, LUXEMBOURG (ci-après dénommée le **DISTRIBUTEUR**)

Préambule:

ELRA a désigné **ELDA S.A.**, son « organe opérationnel », comme étant en charge de toutes ses activités liées à la Distribution de Ressources Linguistiques, et dont le siège social est situé : 55-57 rue Brillat Savarin - 75013 Paris, FRANCE
Enregistrée par le Tribunal de commerce de Paris : RCS Paris B 402 781 876 (95b147 95).

IL A ETE CONVENU ET ARRETE CE QUI SUIT :

« Les parties » désignent le DISTRIBUTEUR et le FOURNISSEUR.

1. Le FOURNISSEUR certifie qu'il est le propriétaire légitime des **Ressources Linguistiques** décrites en Annexe A.

2. Le FOURNISSEUR accorde au DISTRIBUTEUR, qui l'accepte, le droit non-exclusif de distribuer les **Ressources Linguistiques** décrites en Annexe A. "Distribution" signifie que le FOURNISSEUR cède au DISTRIBUTEUR le droit de reproduction, de représentation, d'adaptation, de traduction, d'utilisation et donc de commercialisation, de tout ou partie, des **Ressources Linguistiques**.

3. Le FOURNISSEUR autorise le DISTRIBUTEUR à concéder des Licences UTILISATEUR pour l'utilisation de **Ressources Linguistiques** à toute entité juridique. Le DISTRIBUTEUR devra imposer les conditions pertinentes et nécessaires de ce CONTRAT à cette susdite entité.

4. Le FOURNISSEUR accorde au DISTRIBUTEUR le droit de procéder, directement ou par l'intermédiaire de ses sous-traitants désignés, à un Contrôle de Qualité rapide du Contenu (QQC, Quick Quality Check) des **Ressources Linguistiques** en fonction de leur documentation, à tout moment jugé opportun par le DISTRIBUTEUR. Le DISTRIBUTEUR fournira des comptes rendus tels que des rapports d'incidents ou des réponses directes au service de Contrôle de Qualité rapide du Contenu et devra établir une Liste Formelles d'Erreurs (FEL, Formal Error List). Cette liste sera rendue publique et liée à la description des **Ressources Linguistiques** au sein du catalogue. Suite à ces actions, les ressources linguistiques pourront se voir attribuer un Label Qualité. Le DISTRIBUTEUR pourra être amené à effectuer des travaux de corrections sur les incidents rapportés. Les détails sur la procédure seront à discuter directement avec le FOURNISSEUR.

   Le DISTRIBUTEUR pourra distribuer un patch de correction, qui permettra de supprimer les incidents rapportés et permettra ainsi à l'utilisateur de générer une nouvelle version des **Ressources Linguistiques**.

   Le DISTRIBUTEUR accorde au FOURNISSEUR les droits permanents, irrévocables, non transférables, et gratuits d'utiliser le patch de correction des **Ressources Linguistiques** pour son bénéfice exclusif.

5. Les **Ressources Linguistiques** pourront être dupliquées par le DISTRIBUTEUR comme indiqué en Annexe B. Le DISTRIBUTEUR est également autorisé à reproduire, dans son ensemble ou en partie, et de modifier les **Ressources Linguistiques**, aussi bien que de joindre de la DOCUMENTATION et/ou des MANUELS, aux seules fins de leur distribution.

6. Le DISTRIBUTEUR accepte de payer au FOURNISSEUR une compensation. Les conditions financières sont précisées en Annexe C.

7. Le DISTRIBUTEUR s'engage à faire référence au FOURNISSEUR, ainsi qu'au nom et référence des **ressources linguistiques** dans ses publications qui mentionnent les **ressources linguistiques**. Le DISTRIBUTEUR s'engage à ne pas utiliser le nom du FOURNISSEUR pour la promotion de ses produits ou services. Il s'engage à éviter toute publication qui laisserait sous-entendre une approbation ou l'aval du distributeur pour lesdits produits ou services..

8. Le FOURNISSEUR garantit au DISTRIBUTEUR l'exercice paisible des droits cédés. Il le garantit contre tous troubles, revendications et évictions quelconques. Il le défend contre toutes les atteintes qui seraient portées à jouissance paisible des droits cédés.

9. Le DISTRIBUTEUR reconnaît accepter les **Ressources Linguistiques** « telles quelles », c'est-à-dire avec tous les défauts qui peuvent subsister, et sans aucune garantie de l'adéquation de telles ressources pour un usage particulier.

10. Les deux parties excluent toute responsabilité de quelque nature que ce soit concernant la perte ou les dommages directs, consécutifs ou indirects, tels que la perte de bénéfices, les pertes de commandes, préjudices financiers ou commerciaux, que pourraient subir l'une ou l'autre des parties en ce qui concerne la distribution des **Ressources Linguistiques**.

11. Les parties ne peuvent être tenues pour responsables de l'inexécution de leurs obligations prévues dans le présent contrat si celle-ci est due à un cas de « force majeure » ou bien à des circonstances qui échappent raisonnablement à leur contrôle. Dans le cas où un tel événement se déroulait, toutes les conditions de ce CONTRAT devront être suspendues pendant la durée de cet événement.

12. Le FOURNISSEUR ne mettra pas en vente directement ou indirectement les **Ressources Linguistiques** décrites en Annexe A, à des conditions et à une politique tarifaires qui diffèrent de celles indiquées publiquement par le DISTRIBUTEUR. Si le FOURNISSEUR décide d'avoir d'autres canaux de distribution directs ou indirects, le FOURNISSEUR offrira au DISTRIBUTEUR des conditions avantageuses, qui seront non discriminatoires et qui seront au moins aussi avantageuses que celles offertes à d'autres. Le prix minimum à respecter par les autres distributeurs, y compris des distributions directes réalisées par le FOURNISSEUR, ne devra pas être inférieur au prix du catalogue du distributeur. Néanmoins, le FOURNISSEUR informera le DISTRIBUTEUR lorsqu'il décidera d'offrir à des tiers des conditions et une politique tarifaire différentes de celles convenues en Annexe C.

13. Les parties déclarent leur intention de chercher une solution amiable à toutes difficultés qui pourraient surgir à propos du présent CONTRAT. Si l'une des clauses du contrat est nulle au regard d'une règle de droit ou d'une loi en vigueur, elle sera réputée non écrite mais n'entraînera pas la nullité de l'ensemble du contrat.

14. Le présent CONTRAT est régi par le Droit français. Le Tribunal de commerce de Paris sera le seul compétent en cas de litige.

Ce CONTRAT se compose de 14 articles ainsi que des Annexes A, B et C ci-après.

Les parties reconnaissent avoir pris connaissance de ce CONTRAT, et acceptent de le parapher, de le signer et de le renvoyer à ELDA, l'organe opérationnel d'ELRA, en deux exemplaires :

SIGNATURES DES DEUX PARTIES :

_____	_____
Représentant du **FOURNISSEUR**	Représentant d'**ELRA**
Nom :	Nom : Khalid CHOUKRI
Titre :	Titre : PDG
Date :	Date :

ANNEXES

ANNEXE A: DESCRIPTION DES RESSOURCES LINGUISTIQUES :

ANNEXE B: MOYENS DE DISTRIBUTION :

ANNEXE C : DESCRIPTION DES CONDITIONS FINANCIERES :

Les aspects financiers, incluant le paiement des « Royalties » au FOURNISSEUR, sont gérés par ELDA, agissant en qualité d'organe opérationnel d'ELRA.

La compensation financière est répartie sur la base suivante :
ELDA= % PROVIDER= %

Le prix par exemplaire (en €) s'élève à :

Usage de Recherche :
Membres ELRA :
Non-membres ELRA :

Usage Commercial :
Membres ELRA :
Non-membres ELRA :

Les ventes seront rapportées au FOURNISSEUR chaque semestre (fin décembre et fin juin) par écrit.
Les paiements, tels que définis ci-dessus, seront réglés sous trente jours à réception des factures, par virement du montant sur le numéro du compte bancaire indiqué sur les factures respectives.
Lesdits montants sont indiqués hors taxe.

5.5 Contrat intégrateur ELDA

CONTRAT INTEGRATEUR DE RESSOURCES LINGUISTIQUES

(Contrat n° LC/ELDA/VAR/2007/000/NAME)

Entre
".....", dont le siège social est situé :
Ci-dessous dénommée "**l'intégrateur**" et représentée par ....., agissant en sa qualité de .....,

et

ELDA (Agence pour l'Evaluation et la Distribution des Ressources Linguistiques), S.A. au capital de 250 000 EURO, dont le siège social est situé : 55-57 rue Brillat Savarin - 75013 Paris, France, immatriculée au Registre du Commerce et des Sociétés de Paris : RCS Paris B 402 781 876 (95b147 95)
Ci-dessous dénommée "**le distributeur**", et représentée par Monsieur Khalid CHOUKRI agissant en sa qualité de gérant,

IL A ETE CONVENU ET ARRETE CE QUI SUIT :

"Les parties" désigne le distributeur et l'intégrateur.

1. Les **ressources linguistiques**, objet de ce contrat, pour lesquelles le distributeur dispose d'un droit de distribution cédé par l'ayant-droit, sont définies en annexe A.

2. Le lieu d'utilisation des **ressources linguistiques** est indiqué en annexe B.

3. Le distributeur accorde à l'intégrateur le droit de reproduction des **ressources linguistiques,** de façon temporaire ou permanente, ainsi que la traduction, l'adaptation, la modification par tout moyen des **ressources linguistiques**, si l'une ou l'autre de ces actions, qui requiert habituellement une autorisation de la part de l'ayant-droit des **ressources linguistiques**, s'avère nécessaire pour **accéder** aux **ressources linguistiques** et rendre possible leur utilisation.

4. Le distributeur accorde à l'intégrateur le droit non exclusif de créer et de développer des produits ou des services dérivés, basés sur tout ou partie des **ressources linguistiques**, et qui peuvent être commercialisés selon la politique commerciale de l'intégrateur.

5. L'intégrateur n'est pas autorisé à mettre à la disposition des tiers (en particulier du public) tout ou partie des **ressources linguistiques**, évalué quantitativement et/ou qualitativement, que ce soit par la redistribution de copies, par la location, le leasing ou toute autre forme de distribution, y compris des copies gratuites ou open source. Tous les droits mentionnés dans ce contrat son entendus perpétuels, mondiaux et libres de toute redevance.

6. L'intégrateur n'est pas autorisé à reproduire les **ressources linguistiques**, ni à distribuer des produits dérivés ou services incluant tout ou partie des données dans un but commercial ou non (distribution gracieuse), sous toute forme ou par tout moyen, qui permettraient de reconstituer les **ressources linguistiques** ou une partie des **ressources linguistiques**.

7. La concession de droit d'utilisation dont bénéficie l'intégrateur n'entraîne le transfert d'aucun droit de propriété sur tout ou partie des **ressources linguistiques**.

8. Sans préjudice des autres dispositions du présent contrat, l'intégrateur n'est pas autorisé à transférer à des tiers les droits mentionnés dans le présent contrat. Les **ressources linguistiques** ne pourront pas être transférées vers un lieu d'utilisation différent de celui indiqué dans le présent contrat. L'accès aux **ressources linguistiques** ne pourra non plus s'effectuer à partir d'un autre lieu d'utilisation.

9. Le distributeur et l'ayant-droit ne peuvent être tenus pour responsables de la qualité des données ou pour toutes conséquences résultant de leur utilisation. Le distributeur et l'ayant-droit ne garantissent pas l'adéquation des **ressources linguistiques** pour un usage particulier. L'intégrateur reconnaît accepter les **ressources linguistiques** "telles quelles" c'est-à-dire avec tous les défauts qui peuvent subsister, et sans aucune garantie de l'adéquation de telles ressources pour un usage particulier.

10. L'intégrateur et le distributeur sont des entités légales indépendantes. Aucune des dispositions du présent contrat ne saurait être interprétée comme créant une relation employeur/employé, un partenariat ou une joint venture entre l'intégrateur et le distributeur.

11. L'intégrateur n'est pas autorisé à engager ni à assumer, par écrit ou par tout autre moyen, la responsabilité ni aucune obligation, de quelque nature que ce soit, expresse ou implicite, au nom ou pour le compte du distributeur.

12. Les parties ne peuvent être tenues pour responsables vis-à-vis l'une de l'autre des dommages, directs ou indirects, tels que la perte de bénéfices, les pertes de commandes, préjudices financier ou commercial qui résulteraient de l'exécution de ce contrat.

13. L'intégrateur s'engage à faire référence au distributeur, ainsi qu'au nom et à la référence des **ressources linguistiques** dans ses publications qui mentionnent les **ressources linguistiques**. La mention suivante doit être utilisée : « catalogue ELRA (http://catalog.elra.info), NOM DE LA RESSOURCE LINGUISTIQUE, référence catalogue : ELRA-XXXX »

14. L'intégrateur s'engage à ne pas utiliser le nom du distributeur pour la promotion de ses produits ou services. Il s'engage à éviter toute publication qui laisserait sous-entendre une approbation ou l'aval du distributeur pour lesdits produits ou services.

15. L'intégrateur s'engage à payer au distributeur une compensation. Le mode de paiement et l'échelonnement des paiements sont définis en annexe C.

16. Le présent contrat est régi par le droit français. Les parties déclarent leur intention de chercher une solution amiable à toute difficulté qui pourrait surgir à propos du contrat. En cas d'impossibilité et à défaut d'accord entre les parties, le Tribunal de commerce de Paris sera le seul compétent pour connaître du litige.

Signatures des parties

Fait à Paris, en deux exemplaires

_____	_____
Représentant de l'intégrateur	Représentant d'ELDA
Nom :	Nom : Khalid CHOUKRI
Titre :	Titre : Gérant
Date :	Date :

ANNEXES

ANNEXE A : DESCRIPTION DES RESSOURCES LINGUISTIQUES

ANNEXE B : LIEU D'UTILISATION

ANNEXE C : DESCRIPTION DES CONDITIONS FINANCIERES

5.6 Contrat utilisateur final ELDA

CONTRAT UTILISATEUR FINAL DE RESSOURCES LINGUISTIQUES

(Contrat n° LC/ELDA/ENDUSER/2007/000/NAME)

Entre
"....." , dont le siège social est situé : ,
Ci-dessous dénommée "**l'utilisateur**" et représentée par ,
agissant en sa qualité de ,

et

ELDA (Agence pour l'Evaluation et la Distribution des Ressources Linguistiques), S.A. au capital de 250 000 EURO, dont le siège social est situé : 55-57 rue Brillat Savarin - 75013 Paris, France, immatriculée au Registre du Commerce et des Sociétés de Paris : RCS Paris B 402 781 876 (95b147 95)
Ci-dessous dénommée "**le distributeur**", et représentée par Monsieur Khalid CHOUKRI agissant en sa qualité de gérant,

IL A ETE CONVENU ET ARRETE CE QUI SUIT :

"Les parties" désigne le distributeur et l'utilisateur.

1. Les **ressources linguistiques**, objet de ce contrat, pour lesquelles le distributeur dispose d'un droit de distribution cédé par l'ayant-droit, sont définies en annexe A.

2. Le lieu d'utilisation des **ressources linguistiques** est indiqué en annexe B.

3. L'utilisateur s'engage à utiliser les **ressources linguistiques** dans le cadre de ses activités de recherche en ingénierie linguistique.

4. Le distributeur accorde à l'utilisateur le droit de reproduction des **ressources linguistiques,** de façon temporaire ou permanente, ainsi que la traduction, l'adaptation, la modification par tout moyen des **ressources linguistiques**, si l'une ou l'autre de ces actions, qui requiert habituellement une autorisation de la part de l'ayant-droit des **ressources linguistiques**, s'avère nécessaire pour **accéder** aux **ressources linguistiques** et rendre possible leur utilisation.

5. L'utilisateur n'est pas autorisé à reproduire les **ressources linguistiques**, ni à distribuer des produits dérivés ou services incluant tout ou partie des données dans un but commercial ou non (distribution gracieuse), sous toute forme ou par tout moyen. Tous les droits mentionnés dans ce contrat sont entendus perpétuels, mondiaux et libres de toute redevance.

6. L’utilisateur n'est pas autorisé à mettre à la disposition des tiers (en particulier du public) tout ou partie des **ressources linguistiques**, évalué quantitativement et/ou qualitativement, que ce soit par la redistribution de copies, par la location, le leasing ou toute autre forme de distribution, y compris des copies gratuites ou open source.

7. La concession de droit d'utilisation dont bénéficie l'utilisateur n'entraîne le transfert d'aucun droit de propriété sur tout ou partie des **ressources linguistiques**.

8. Sans préjudice des autres dispositions du présent contrat, l’utilisateur n’est pas autorisé à transférer à des tiers les droits mentionnés dans le présent contrat. Les **ressources linguistiques** ne pourront pas être transférées vers un lieu d’utilisation différent de celui indiqué dans le présent contrat. L’accès aux **ressources linguistiques** ne pourra non plus s’effectuer à partir d’un autre lieu d’utilisation.

9. Le distributeur et l’ayant-droit ne peuvent être tenus pour responsables de la qualité des données ou pour toutes conséquences résultant de leur utilisation. Le distributeur et l’ayant-droit ne garantissent pas l’adéquation des **ressources linguistiques** pour un usage particulier. L’utilisateur reconnaît accepter les **ressources linguistiques** "telles quelles" c'est-à-dire avec tous les défauts qui peuvent subsister, et sans aucune garantie de l'adéquation de telles ressources pour un usage particulier.

10. L’utilisateur et le distributeur sont des entités légales indépendantes. Aucune des dispositions du présent contrat ne saurait être interprétée comme créant une relation employeur/employé, un partenariat ou une joint venture entre l’utilisateur et le distributeur.

11. L’utilisateur n’est pas autorisé à engager ni à assumer, par écrit ou par tout autre moyen, la responsabilité ni aucune obligation, de quelque nature que ce soit, expresse ou implicite, au nom ou pour le compte du distributeur.

12. Les parties ne peuvent être tenues pour responsables vis-à-vis l'une de l'autre des dommages, directs ou indirects, tels que la perte de bénéfices, les pertes de commandes, préjudices financier ou commercial qui résulteraient de l'exécution de ce contrat.

13. L'utilisateur s'engage à faire référence au distributeur, ainsi qu’au nom et référence des **ressources linguistiques** dans ses publications qui mentionnent les **ressources linguistiques**. La mention suivante doit être utilisée : « catalogue ELRA (http://catalog.elra.info), NOM DE LA RESSOURCE LINGUISTIQUE, référence catalogue : ELRA-XXXX »

14. L'utilisateur s'engage à ne pas utiliser le nom du distributeur pour la promotion de ses produits ou services. Il s'engage à éviter toute publication qui laisserait sous-entendre une approbation ou l'aval du distributeur pour lesdits produits ou services.

15. L'utilisateur s'engage à payer au distributeur une compensation. Le mode de paiement et l'échelonnement des paiements sont définis en annexe C.

16. Le présent contrat est régi par le droit français. Les parties déclarent leur intention de chercher une solution amiable à toute difficulté qui pourrait surgir à propos du contrat. En cas d'impossibilité et à défaut d'accord entre les parties, le Tribunal de commerce de Paris sera le seul compétent pour connaître du litige.

Signatures des parties

Fait à Paris, en deux exemplaires

_____	_____
Représentant de l'utilisateur	Représentant d'ELDA
Nom :	Nom : Khalid CHOUKRI
Titre :	Titre : PDG
Date :	Date :

ANNEXES

ANNEXE A : DESCRIPTION DES RESSOURCES LINGUISTIQUES

ANNEXE B : LIEU D'UTILISATION

ANNEXE C : DESCRIPTION DES CONDITIONS FINANCIERES

5.7 Tableau comparatif entre les différentes licences libres

			GNU General Public License et dérivées				Creative Commons		CeCILL		
Développeur / exploitant de référence			*Free Software Foundation*		*dév. Marne la Vallée ; utilisée à l'INRIA*				*INRIA, CNRS, Commissariat à l'Energie Atomique*		
Variantes			GNU GPL	Lesser GPL (LGPL)	Lesser GPL for Linguistic Resources (LGPLLR)	GNU Free Documentation License (FDL)	Contrat type international (Common Law)	Adaptation au droit français	CeCILL	CeCILL-C	CeCILL-B
Domaine matériel d'application			logiciels	Bibliothèques logicielles	"S'applique à certaines ressources linguistiques spécialement désignées - généralement les lexiques, grammaires, thesaurus et corpus textuels"	Documents (tous media)	Toute œuvre de l'esprit protégeable		Conçu pour des logiciels		
Modification de licence			système de "permissions additionnelles" optionnelles ; restrictions additionnelles exclues		Sous-licence réputée accordée et déterminée uniquement par l'ayant droit	Pas de restrictions additionnelles					
Durée d'application / renouvellement			durée du copyright						Toute la durée de protection des droits portant sur le logiciel		
Statut des sous-licenciés en cas de résiliation			Maintien des sous-licences						Maintien des sous-licences		
Utilisation			Exécuter le logiciel pour n'importe quel usage				Exploitation non commerciale OK	idem + affirmation des droits moraux de l'auteur			
			applicabilité de la notion de 'fair use' ou équivalent		Non restreinte		Exploitation commerciale : option		Utilisation libre du logiciel, sans restriction quant au domaine d'application		
Accès et analyse / version intelligible			Obligation de fournir ou mettre à disposition le code source ; interdiction des mesures techniques de protection		Obligation de fournir une version non cryptée ('legible form'), notion transposant la définition du Code Source		Obligation optionnelle de fournir le code source (Licence FSF GNU GPL)		droit d'observer, analyser, tester le fonctionnement. Fournir le code source ou permettre d'y accéder		
Modification	dans le cadre de l'utilisation personnelle		OUI	l'œuvre modifiée doit être elle-même une bibliothèque logicielle			Oui		Oui		
	dans une perspective de redistribution						Option 'No modification'		Oui, Modification-distribution sujette à conditions		
Redistribution	Principe		OUI	OUI	OUI	notion de Copie en Quantité (plus de 100 exemplaires)	OUI		OUI		
	Distribution à titre onéreux		fréquemment à titre gratuit ; possible à titre onéreux		RL distribuées à titre gratuit ; rétribution possible du transfert physique de copie, d'une fourniture de garantie	OUI	Option 'Non commercial'		OUI		
Principe de "contamination"			Copyleft Fort	Copyleft Faible	Copyleft Faible	Copyleft Fort	Option 'Share-alike'		Copyleft Fort	Copyleft Faible	Pas de Copyleft ; forte obligation de citation
Licence applicable à la distribution	ressource telle qu'acquise (verbatim)		Même licence		Même licence	Même licence	Même licence		Même licence		
	Produit dérivé	type 1	Produits dérivés en général : même Licence	Version modifiée de la bibliothèque : même Licence	En principe assimilé à l' "œuvre dérivée" ; exception si : 1) fourniture d'une version décryptée de la RL [et] 2) compatibilité du logiciel avec des versions modifiées de la RL	Produits dérivés (document modifié, traduction) : même Licence			"module interne" : même Licence		
		type 2	"agrégat" de produits séparés et indépendants (pas des extensions de l'œuvre protégée, pas de distrib. conjointe formant un programme plus large) : Licence au choix pour ces derniers	combinaison de bibliothèques : licence au choix		Combinaison ou collection de documents soumis à la licence FDL : même Licence			"module externe" : exécution dans des espaces d'adressage différents		
		Type 3		Produit combiné (combined work) issu de liens établis entre application et bibliothèque : licence au choix, sous conditions	LGPLLR si les modifications sont des RL ; sinon, simple aggrégation ("mere aggregation"), sans portée / Productions (output) du logiciel d'exploitation couvertes si elles constituent un 'work based on the LR'	aggrégat avec des œuvres indépendantes : licence au choix pour ces dernières					
'Attribution' et suivi des modifications			Marquage très apparent des modifications, de leur(s) auteur(s), de leur date				Option 'By'	Non optionnel - Attribution obligatoire en droit français	Obligation pour l'auteur d'une modification d'indiquer son nom et la date de création ; obligation de conserver les mentions de Propriété Intellectuelle		
Suivi des distributions											
Garanties / Responsabilités / Obligations	Garantie du concédant qu'il est titulaire des droits					Mentions dans le texte de couverture	OUI, garantie de jouissance paisible		Simple affirmation, pas de garantie vis-à-vis de l'atteinte aux		
	Fonctionnement conforme / non préjudiciable		NON		NON	NON	NON		NON ; clauses de limitation de garantie et de responsabilité (conformes au droit français)		
	Adéquation à un usage déterminé		NON		NON	NON	NON		NON		
	Assistance technique, correctifs, mises à jour		NON		NON				Liberté du concédant ; Obligation expressément écartée		
	Information / avertissement				NON				obligation de fournir un avertissement quant aux restrictions de		
	Obligations du concédant								Responsabilité pour faute		
	Obligations du licencié	Obligations vis-à-vis du concédant	Sous-licence réputée accordée et déterminée par le donneur original : pas d'obligation du licencié d'assurer le respect de la licence par les sous-licenciés						Ne pas porter atteinte aux droits du titulaire ; prendre toute mesure pour éviter que le personnel n'y porte atteinte		
		Obligations vis-à-vis du sous-licencié									
	Clauses spécifiques sur la responsabilité	Préjudice commercial									
		autres									
Interprétation, Règlement des litiges	Loi d'autonomie		comporte des références à la Common Law				Non spécifié dans nombre de cas	Applicabilité du Droit Français	Applicabilité du Droit Français	Applicabilité du Droit Français	Applicabilité du Droit Français
	Compétence juridictionnelle										

6. Références bibliographiques

Aarts et Eggen, 2002	Aarts E.H.L., and Eggen B. (eds.) *Ambient Intelligence in HomeLab*. Eindhoven: Neroc, 2002.
Baude, 2006	Baude O., *Corpus Oraux : Les Bonnes Pratiques d'une Communauté Scientifique*, 2006.
Burger et al., 2002	Burger S., McLaren V., Yu H. *The ISL Meeting Corpus: The impact on meeting type on speech style*, Actes d'ICSLP, Denver, USA, 2002.
Cappelletti et al., 2008	Cappelletti A., Lepri B., Mana N., Pianesi F., and Zancanaro M. *A multimodal data collection of daily activities in a real instrumented apartment*, Actes du Workshop on Multimodal Corpora, LREC 2008, Marrakech, Maroc, mai 2008.
Carletta J. et al., 2005	Carletta J. *et al. The AMI Meetings Corpus*, Actes de the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior".
Casas et al., 2004	Casas J., Stiefelhagen R. *et al. Multi-camera/multi-microphone system design for continuous room monitoring*, CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL Consortium Deliverable D4.1, juillet 2004.
Chen et al., 2005	Chen L., Rose R.T., Parrill F., Han X., Tu J., Huang Z., Harper M., Quek F., McNeill D.,Tuttle R., Huang T. *VACE multimodal meeting corpus*. Actes de Multimodal Interaction and Related Machine Learning Algorithms, 2005.
Chitu et Rothkrantz, 2008	Chitu A. G., Rothkrantz L. J.M. *Dutch Multimodal Corpus for Speech Recognition*, Actes de Workshop on Multimodal Corpora, LREC 2008, Marrakech, Maroc, mai 2008.
Colletta et al., 2008	Colletta J.-M., Kunene R., Venouil A., Tcherkassof A. *Double level analysis of the Multimodal Expressions of Emotions in Human-Machine Interaction*, Actes du Workshop on Multimodal Corpora, LREC 2008, Marrakech, Maroc, mai 2008.
Cucchiarini et al., 2001a	Cucchiarini C., Daelemans W. et Strik H., *Strengthening the Dutch Human Language Technology Infrastructure*, ELRA Newsletter Vol. 6 N. 4. 2001a.
Cucchiarini et al., 2001b	Cucchiarini C., Daelemans W. et Strik H., *Strengthening the Dutch Language and Speech Technology Infrastructure*, Actes de la conférence COCOSDA 2001b.
Dejean et Gaussier, 2002	Hervé Dejean & Éric Gaussier, *une nouvelle approche à l'extraction de lexique bilingues à partir de corpus comparables*, *Lexicometrica*, No spécial 2002, pp. 22.
Draxler, 1995	Draxler C., *Advanced Distribution Means for Spoken Language Corpora*, 1995.
Francis et Kucera, 1982	Francis , W. Nelson, and Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar*, Boston: Houghton Mifflin Company, 1982.
Garofolo J. et al., 2006	Garofolo J. et al. *Performance Evaluation Protocol for Face Person and Vehicule Detection and Tracking in Video Analysis and Content Extraction (VACE-II)*. Public Document, ARDA, 2006.
Garofolo, 2004	Garofolo J. et al. *The NIST Meeting Room Pilot Corpus*, Actes de la 4ème conférence internationale sur les ressources linguistiques et l'évaluation (LREC), 2004.
Ide & Véronis, 1996	Ide, N., & Véronis, J. (1996). Application de la TEI aux industries de la langue: le Corpus Encoding Standard. *Cahiers GUTenberg, 24*, 166-169.
Intille et al., 2006	Intille S.S., Larson K., Munguia Tapia E., Beaudin J, Kaushik P., Nawyn J., and Rockinson R. (2006). *Using a live-in laboratory for ubiquitous computing research*. In K.P. Fishkin, B. Schiele, P. Nixon, and A. Quigley (eds.) Actes de PERVASIVE 2006, vol. LNCS 3968,. Berlin Heidelberg: Springer-Verlag, pp. 349-365, 2006.
Janin et Baron, 2003	Janin A., Baron D. *et al. The ICSI Meeting Corpus*, Actes de ICASSP'03, Hong Kong, China, avril 2003.
Kidd et al., 1999	Kidd C.D., Orr R., Abowd G.D., Atkeson C.G., Essa I.A., MacIntyre B., Mynatt E., Starner T.E., Newstetter W. *The Aware Home: A Living Laboratory for Ubiquitous Computing Research*. Actes de the Second International Workshop on Cooperative Buildings - CoBuild'99.
Krauwer, 1998	Krauwer S., *ELSNET and ELRA: A common past and a common future*, ELRA Newsletter Vol. 3 N. 2. 1998.
Krauwer et al., 2006	Krauwer S., Maegaard B., Choukri K., Damsgaard Jørgensen L., *Report on Basic Language Resource Kit (BLARK) for Arabic*, projet NEMLAR, 2006.

Lathoud et al., 2004	Lathoud G., Odobez J.-M., Gatica-Perez D. *AV16.3: an audio-visual corpus for speaker localization and tracking*, Actes du MLMI Workshop, 2004.
Lincoln, 2005	Lincoln M. *The Multi-channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments*, Actes de ASRU05.
Maegaard et al., 2005	Maegaard B., Choukri K., Calzolari N., Odijk J., *ELRA – European Language Resources Association-Background, Recent Developments and Future Perspectives*, LRE Journal, Volume 39, Number 1 / février, 2005.
Martens et al., 2008	Martens S., Becker J.H., Tuytelaars T., Moens M.-F. *Multimodal Data Collection in the AMASS++project*, Actes de Workshop on Multimodal Corpora, LREC 2008, Marrakech, Maroc, mai 2008.
Martial et Stanford, 2001	Martial M., Stanford V. *Synchronizing Multimodal Data Streams Acquired Using Commodity Hardware*, Actes de VSSN'01, 2001.
Martinez 2004	Martínez, J. M. *MPEG-7 Overview (version 10)*. Coding of moving picture and audio. International Organization for standardization (ISO/IEC JTC1/SC29/WG11 N6828), Palma de Majorque, 2004.
McCowan et al., 2003	McCowan, I, Gatica-Perez, D, Bengio, D, Moore, D and Bourlard, H. *Towards Computer Understanding of Human Interactions*. Proceedings of the European Symposium on Ambient Intelligence (EUSAI) (invited keynote paper), Eindhoven, Nov. 2003.
McEnery et al. 1998	McEnery T., Burnard L., Wilson A. and Baker, *Validation of Linguistic Corpora*, 28 avril 1998.
Moore, 2002	Moore D.C. *The IDIAP Smart Meeting Room*. IDIAP Specification Document, novembre 2002.
Moreau et al. 2008	Moreau N., Mostefa D., Stiefelhagen R. (2008). *Perceptual Component Evaluation and Data Collection*, in: Alex Waibel, Rainer Stiefelhagen (Eds.), "CHIL: Computers in the Human Interaction Loop", Springer, 2008.
Moreau et al., 2007a	Moreau N. *et al*. *Exploitation Material for the CHIL Evaluation Campaign 3*. CHIL Public Deliverable D7.14, 2007.
Mostefa et al., 2006	Mostefa D., Garcia M.-N., Choukri K. *Evaluation of Multimodal Components within CHIL*, Actes de la 5ème conférence internationale sur les ressources linguistiques et l'évaluation (LREC), Gênes, Italie, 2006.
Mostefa et al., 2007	Mostefa D., Moreau N., Choukri K. *et al.* The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms, *Journal on Language Resources and Evaluation*, Vol. 41, No. 3-4, , pp. 389-407, Springer, décembre 2007.
Pastra, 2006	Pastra K. *Beyond multimedia integration: corpora and annotations for cross-media decision mechanisms*, Actes de the 5th Language Resources and Evaluation Conference (LREC), 2006.
Raju et Prasad, 2006	Raju H., Prasad S. *Annotation Guidelines for Video Analysis and Content Extraction (VACE-II)*. Annotation Guidelines, Public Document, ARDA, 2006.
Sampson, 1995	Geoffrey Sampson (Ed.). *Susanne Corpus*, School of Cognitive & Computing Sciences, University of Sussex, 1995.
Smeaton et al., 2006	Smeaton, A., Over, P., & Kraaij, W. *Evaluation campaigns and TRECVid*, Actes de the 8th ACM International Workshop on Multimedia Information Retrieval MIR '06. pp. 321-330, 2006.
Spachos et al., 2008	Spachos D., Zlatintsi A., et al. *The MUSCLE movie database: A multimodal corpus with rich annotation for dialogue and saliency detection*, Actes du Workshop on Multimodal Corpora, LREC 2008, Marrakech, Maroc, mai 2008.
Stanford et al., 2003	Stanford V., Garofolo J., Galibert O., Michel M., Laprun C. *The NIST Smart Space and Meeting Room Projects: Signals, Acquisition Annotation and Metrics*, in Proceedings of ICASSP'03, vol.4, pp.736-9, avril 2003.
Stiefelhagen et al., 2007	Stiefelhagen R., Bernardin K., Bowers R., Garofolo J., Mostefa D., Soundararajan P. *The CLEAR 2006 Evaluation*. In R. Stiefelhagen and J. Garofolo, editors, Multimodal Technologies for Perception of Humans, Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006, number 4122 in Springer Lecture Notes in Computer Science, pages 1–45, 2007.
Zancanaro et al., 2004	Zancanaro M., Pianesi F. *et al. Initial Specification of the CHIL Scenario*, CHIL-WP3-Scenarios-V2.2-2004-02-13, CHIL Consortium Document, février 2004.