



www.ichim.org

Les institutions culturelles et le numérique
Cultural institutions and digital technology

École du Louvre
8 - 12 septembre 2003

**NUMÉRISATION ET VALORISATION DES
COLLECTIONS D'OUVRAGES ANCIENS DU
LABORATOIRE ATILF, CNRS, UMR 7118, NANCY II.**

Isabelle Turcan, Université de Lyon, France

« Acte publié avec le soutien de la Mission de la Recherche et
de la Technologie du Ministère de la Culture et de la Communication »

humaines

Résumé

La richesse des ouvrages anciens conservés au laboratoire ATILF invite à définir rigoureusement la notion de *collection* pour présenter les critères conférant au fonds la valeur de *collection patrimoniale*.

L'identification de ses composantes conduit à déterminer des typologies de collections intéressant à des titres différents le patrimoine, le **patrimoine international** (ouvrages contribuant à mieux connaître l'histoire du livre ancien, de sa diffusion en Europe et dans le monde entier ; ouvrages inscrits dans l'histoire du développement de la francophonie), le **patrimoine national** (textes concernant l'histoire de la langue et de la littérature françaises), le **patrimoine régional** (rôle des ouvrages imprimés en Lorraine par rapport à la place historique de cette région au siècle des Lumières). (115)

L'identité des collections détermine alors un choix de **numérisation** ne se limitant pas à constituer un corpus d'images enregistrées dans une « numémathèque » statique : la gestion des documents électroniques et la qualité de transmission de leurs contenus contribueront à valoriser autant leur dimension patrimoniale auprès d'un public différencié que leur fonction de ressources documentaires ouvertes à de nouvelles orientations de recherche en sciences humaines ; le laboratoire apporte une indéniable valeur ajoutée aux documents avec, outre l'efficacité et la convivialité des outils d'exploitation automatique des matériaux numérisés et des outils pédagogiques associés permettant une hyper-navigation d'un corpus textuel à l'autre, la compétence des chercheurs impliqués dans ce programme qui associe numérisation et valorisation pour un meilleur partage des ressources. (121)

Mots-clés ☐ collection patrimoniale, ouvrages anciens, dictionnaires, numérisation, valorisation, hypernavigation, sciences humaines.

Abstract

The richness of the old works preserved at laboratory ATILF invites to rigorously define the concept of collection in order to present the criteria giving to the funds the value of patrimonial collection.

The identification of its components results in determining typologies of collections concerning for various reasons the inheritance, the international heritage (works contributing to

better knowing the history of the old book, its diffusion in Europe and in the whole world ; works registered in the history of the development of French language), national heritage (texts concerning the history of French language and French literature), the regional heritage (role of the works printed in Lorraine compared to the historical place of this area at the century of the Lights).

The identity of the collections then determines a choice of digitalization which is not limited to constitute a corpus of images recorded in a static "numérithèque" : the management of the electronic documents and the quality of transmission of their contents will contribute to develop for a varied public as much their patrimonial dimension that their function from documentary resources opened with new orientations of research in social sciences ; in addition to the effectiveness and the user-friendliness of the tools of automatic exploitation of digitized materials and the associated teaching aids which will allow an hypernavigation of a textual corpus to the other, the laboratory brings an undeniable value to the documents with the competence of the researchers involved in this program associating digitalization and valorization.

Keywords : patrimonial collection, old works, dictionaries, digitizing, enhanced-value (digitalization, valorization), hypernavigation, social sciences.

J'interviens en tant que responsable du programme de numérisation des ouvrages anciens conservés au laboratoire ATILF [Analyse et Traitement Informatique de la Langue Française, <http://www.inalf.fr/atilf>, successeur de la composante nancéenne de l'INaLF, Institut National de la Langue Française] le Centre de Documentation dirigé par Madame Viviane Berthelier a catalogué le fonds en harmonie avec les bibliothèques de Lorraine et a commencé à travailler à un catalogage enrichi répondant aux attentes du public spécialisé en ouvrages imprimés anciens et en lexicographie. J'inscrirai mon propos dans la continuité de la contribution que nous avons proposée, M. J. - M. Pierrel, directeur du laboratoire, et moi-même au numéro de la revue *Document Numérique* consacré à la *Numérisation du patrimoine* [Turcan - Pierrel (2003)].

Après avoir décrit les spécificités de notre fonds au sein de l'ensemble des ressources documentaires du laboratoire (comme les bases littéraires et lexicographiques informatisées, *FRANTEXT*, <http://www.atilf.fr/frantext> et le *TLFi*, <http://www.atilf.fr/tlfi>), j'attirerai l'attention sur les pièces remarquables de notre collection de dictionnaires anciens et sur les critères d'identification de collections particulières d'ouvrages sériels appartenant à l'histoire de la lexicographie française. Dans un second temps, je présenterai nos objectifs de valorisation et nos choix afférents de numérisation en mode image et en mode texte, d'indexation des pages numérisées et de gestion électronique des contenus typographiques et textuels.

Identité du fonds ancien du laboratoire ATILF

Présentation générale du fonds

Les imprimés anciens conservés dans notre laboratoire constituent une véritable richesse patrimoniale pour les SHS (Sciences Humaines et Sociales) et concernent l'histoire de la langue française et de ses textes fondateurs, avec, sur cinq siècles, du XVI^e au XX^e siècle, un ensemble de textes représentatifs des évolutions notre langue, de la diversité de ses réalisations – ainsi, par exemple, des modifications de ses principes normatifs au cours des siècles, de la reconnaissance de ses variétés linguistiques et de ses richesses dans les

domaines d'activités humaines intéressant les grandes composantes des SHS, les principaux domaines étant ceux de l'histoire de la langue et la littérature françaises, de l'histoire des idées et des théories grammaticales et linguistiques, de l'histoire des techniques via les vocabulaires spécialisés, de l'histoire des terminologies...

Or, cette composante de notre patrimoine documentaire mérite à la fois d'être mieux connue des chercheurs français ou étrangers intéressés par l'histoire de la langue française, par l'histoire des dictionnaires l'ayant accompagnée et enrichie de mots nouveaux au gré des évolutions de notre culture dans une Europe francophile et francophone, et à la fois de vivre, de rayonner grâce aux consultations d'un public diversifié ouvert à l'exploration et à l'exploitation automatique de documents numérisés constitués en bases de données. C'est ainsi que nous ne dissociions pas numérisation et valorisation par le partage de ressources documentaires ainsi mieux préservées et mieux diffusées.

La constitution du fonds est liée aux principales phases des développements des disciplines scientifiques enracinées dans l'histoire des SHS, notamment celle de la langue française dont le laboratoire a été un des principaux foyers de recherche dans le monde ces dernières décennies – ainsi, pour la lexicographie (l'art de composer et rédiger des dictionnaires), pour le traitement automatique de la langue (exploitation des nouvelles technologies pour la constitution de vastes corpus analysables via des bases de données), et pour la métalexigraphie fondée sur la critique des dictionnaires, discipline qui s'est imposée au cours de la seconde moitié du XX^e siècle grâce à la richesse et à la diversité de ressources linguistiques exploitables via les nouvelles technologies notamment développées à l'ATILF. Au cœur du programme de numérisation, la lexicographie est représentée par la réalisation du *Trésor de la langue française*, qui a nécessité la consultation permanente de nombreux dictionnaires anciens représentatifs de la variété des richesses lexicales de notre langue et d'ouvrages de grammaire représentatifs de l'évolution des usages – elle a suscité la constitution de corpus associés de textes littéraires pour affiner la représentation de la langue via la diversité des usages. La mise en œuvre de vastes corpus constitués en bases de données exploitables automatiquement a permis de valoriser les ressources documentaires littéraires par la confrontation avec les ressources d'ouvrages linguistiques anciens. Le développement de la critique textuelle, de plus en

plus appliquée aux textes des dictionnaires, a généré l'exigence scientifique conduisant à la maîtrise des contenus et des outils susceptibles d'en approfondir la connaissance.

La diversité et la richesse de représentativité de notre fonds, qu'il s'agisse de dictionnaires généraux et spécialisés, d'encyclopédies ou de vocabulaires spécialisés et techniques, de traités linguistiques, d'ouvrages de grammaire ou de textes critiques et didactiques sur des aspects particuliers de la langue française, invite à préciser la notion de *collection* / *collections* pour en présenter les critères qui confèrent à l'ensemble des ouvrages conservés la valeur de *collection patrimoniale*, de collection appartenant au patrimoine documentaire, littéraire et linguistique de la langue française.

Une collection, des collections □ une collection de collections

Le fonds ancien réunit plus de 500 ouvrages, soit 1500 volumes imprimés en France et, plus largement, en Europe parfois enrichis d'annotations manuscrites et d'indications intéressantes, au-delà de l'histoire des textes et de leurs contenus, l'histoire du livre □ sur l'ensemble du centre de documentation (un million de pages), on estime à 700 000 pages les dictionnaires anciens, ouvrages reliés, de taille diverses, avec essentiellement du texte et peu d'illustrations (annexe 5.1.).

Une collection □ un objet de collection

Si la notion de *collection* désigne depuis le milieu du XVIII^e siècle, de façon commune, l'association, la réunion d'un ensemble cohérent d'objets ayant quelque rapport, on n'oubliera pas que le terme ne s'est pas toujours appliqué à l'ensemble d'un trésor collationné au fil du temps par des collectionneurs que motivait le choix de beaux objets, de pièces rares, etc... Le mot de *collection* a d'abord désigné le recueil des passages les plus significatifs ou jugés les plus beaux trouvés dans divers ouvrages par les lettrés soucieux de conserver une trace de leurs lectures, en guise de notes, ce qui a progressivement conduit aux anthologies des meilleurs passages des grandes productions littéraires. Dans cette logique le livre qui se prête le mieux à la logique de la *collection* est bien d'abord le dictionnaire, comme en témoigne d'ailleurs, par exemple, la formule

proposée dans le *Discours préliminaire* de l'*Encyclopédie* (1751) «[...] nous envisagerons cette collection comme dictionnaire des sciences et des arts».

Au-delà de la conception du dictionnaire comme livre-bibliothèque par excellence [Turcan, 2003c], toute bibliothèque constituée par définition est une collection de livres.

Les pièces de collection

Parmi nos pièces remarquables, on signalera une des concordances des grands textes sacrés de la Bible (1540), un exemplaire d'un des dictionnaires de Calepin (1568) et de son dictionnaire octolingue (1620), un exemplaire du *Thresor de la langue française tant ancienne que moderne* de J. Nicot (1606) enrichi d'annotations manuscrites de deux des maîtres de l'histoire de la langue française au XIX^e s., A. Thomas et son disciple A. Darmesteter à qui Thomas a dédié son exemplaire, légué ensuite au laboratoire.

Collection ou séries ? rappels

On restera conscient que si le terme de *collection* implique d'abord un rapport externe entre différents objets rassemblés selon des motifs variables d'un collectionneur à l'autre, la reconnaissance de certaines collections se fonde sur des critères relevant de la logique interne déterminée par les propriétés intrinsèques aux objets concernés. Ainsi, dans l'histoire de la lexicographie française faut-il isoler des ensembles de lexicographie sérielle qu'il s'agisse

- de séries fermées et connues, comme c'est le cas des éditions du *Dictionarium* de Calepin recensées par A. Labarre [Labarre 1975], des éditions du *Grand Dictionnaire François-Latin* d'Estienne étudiées par T. R. Wooldridge [Wooldridge, 1992], des éditions du *Dictionnaire de Trévoux* avec une première édition datée de 1704 et une dernière de 1771, que dont nous avons réalisé une première étude de bibliographie matérielle [Turcan, 1999].
- de séries ouvertes sur le futur, comme c'est le cas des éditions officielles du *Dictionnaire de l'Académie française* dont les huit premières sont achevées (1694-1935), mais dont la neuvième est en cours.

- de séries ouvertes sur l'inconnu, comme c'est le cas des éditions non officielles du *Dictionnaire de l'Académie Française* dont ni le recensement ni l'identification ne sont actuellement terminées.

- des séries non encore reconnues comme les différentes éditions du *Dictionnaire François* de C. P. Richelet dont le nombre exact (éditions, tirages, retirages, contrefaçons) est encore en discussion à l'heure actuelle, ce à quoi travaille M. G. Pétrequin qui prépare une thèse sous ma direction, et a été rattaché au laboratoire pour participer à notre programme. Il en est de même pour bien d'autres dictionnaires imprimés sous l'Ancien régime et qui ont connu un succès suffisant pour avoir suscité la concurrence éditoriale (cf. les dictionnaires de Moreri, Furetière, Basnage, Boyer, Bayle, Kramer, etc. ...), mais dont l'identité absolue n'émergera que progressivement au gré des recherches menées par des spécialistes de métalexigraphie ne rechignant pas à l'étude de bibliographie matérielle, longue et ardue.

Parmi les collections de dictionnaires anciens de notre fonds, les plus représentatives sont celles qui correspondent à des séries inscrites dans des périodes diachroniques variées [Turcan, 2001] ☐ ces séries sont définissables selon des critères rigoureux répondant aux exigences scientifiques de la critique bibliographique et des analyses métalexigraphiques.

Les différentes logiques d'appréciation des collections ☐

Il peut paraître aisé pour un spécialiste de lexicographie de discerner la spécificité historique de certaines collections telles les huit éditions officielles du *Dictionnaire de l'Académie française* (1694-1935) par opposition aux éditions non-officielles produites dès la fin du XVII^e s., tout au long du XVIII^e et jusqu'à la première moitié du XIX^e siècle, éditions qui ne furent pas reconnues par l'Académie mais, qui, se recommandant de l'institution en récupéraient le prestige, ont joué un rôle important à la fois dans l'histoire de l'imprimé et dans l'histoire de l'introduction de graphies nouvelles et surtout de mots nouveaux, comme les mots de la Révolution, dans les colonnes de dictionnaires concurrents. Mais, il est en revanche souvent très délicat de décrire de façon idéale, précise, technique, des ouvrages dont le contenu peut être aléatoire et conditionner ainsi l'appréciation d'une impression (tirage, tirage avec ou sans carton, contrefaçon) au sein

d'une série ce qui conditionne la reconnaissance d'une collection. Si une série complète constitue une collection, un ensemble d'exemplaires différents d'une même édition peut aussi constituer une collection spécifique, sans pour autant relever d'une série officielle □ ainsi, doit-on admettre les séries virtuelles en cours d'étude pour les éditions non officielles du *Dictionnaire de l'Académie française*, domaine qui, à ma connaissance n'a pas encore été étudié de façon systématique tant il est complexe (j'y travaille déjà depuis plusieurs années mais il est difficile en la matière d'avoir la certitude d'un recensement de l'ensemble des contrefaçons).

La collection des éditions du *Dictionnaire universel françois & latin de Trévoux* produites tout au long du XVIII^e siècle (1704-1771) offre un exemple similaire dans l'appréciation de collections patrimoniales, avec la réserve du flou concernant l'identification de certaines éditions comme officielles ou non, dès les premiers tirages de 1704 (dont notre fonds conserve trois exemplaires différents), c'est-à-dire attribuables aux instances parisiennes ou trévoltiennes, sans négliger, à partir de 1734, le rôle du libraire Nancéen Antoine qui a produit une « □ contrefaçon □ » en 1734 de l'édition officielle de 1732 (Trévoux / Paris), ce qui renforce la pertinence du critère historique, éditions réalisée dans le Royaume ou hors du Royaume de France : concurrence ou complémentarité □ □ L'étude de la reconnaissance de ces éditions repose sur le critère historique plus ou moins objectif de diffusion et de réception, d'interdiction et de destruction au pilon, pour les témoignages dont on dispose, mais surtout sur le critère métalexigraphique d'analyse des contenus qui permet de dégager les particularités textuelles susceptibles de démontrer d'autres fonctionnements d'ordre hiérarchiques que ceux de la seule chronologie éditoriale. Ainsi, pour les éditions du *Dictionnaire de Trévoux* imprimées réalisées à Nancy par le libraire Antoine, en 1734 puis de 1738 à 1742 □ si la contrefaçon de 1734 a été avérée par l'histoire de la librairie frauduleuse (contrefaçon de l'édition de 1732 considérée comme la troisième officielle, imprimée à Paris et diffusée à Trévoux, mais ne portant sur sa page de titre que la mention de Trévoux), en revanche l'histoire de l'édition de 1738-1742 est plus complexe et encore mal connue □ habituellement rejetée de la série reconnue comme émanant des libraires du royaume, cette édition mériterait d'être prise en compte en vertu du principe historique de la cohérence éditoriale telle que nous pouvons, plus de deux siècles après, l'analyser désormais au vu d'un ensemble lexicographique répondant lui-même à différents principes de cohérences, qu'il s'agisse de la cohérence de

la nomenclature et des enrichissements progressifs apportés d'une édition à l'autre, de l'enrichissement progressif des discours afférents à cette nomenclature, de l'exploitation progressive de sources nouvelles.

Or, seule une procédure de numérisation enrichie d'une OCRisation qui convertirait une partie des images en bases textuelles permettrait de constituer des bases de données différenciées et complémentaires (d'images et de textes, sachant qu'il devient pertinent de considérer la notion d'« image textuelle » [Turcan - Pierrel, 2003]), non seulement des différents tirages dont nous disposons pour une même édition afin d'en dégager les différences ou au contraire d'en garantir la similitude absolue, mais surtout des différents exemplaires d'un ensemble sériel au regard de la chronologie, qui serait susceptible de constituer soit une collection officielle (la collection des éditions trévoltiennes puis parisiennes) par opposition à des impressions parallèles réalisées hors du royaume de France au regard de l'histoire du livre (contrefaçons, retirages considérés comme illégitimes mais finalement intégrés), soit une seule collection hétérogène au regard de la cohérence sérielle des contenus telle que la métalexigraphie peut l'étudier.

Nos collections au sein de la lexicographie sérielle

Parmi les ensembles relevant de la lexicographie sérielle, le laboratoire possède la collection complète des éditions officielles du *Dictionnaire Universel françois & latin dit de Trévoux* (1704-1771) avec laquelle seule la bibliothèque de Trévoux pourrait rivaliser en France et en Europe (cf. notre présentation de la bibliothèque de Trévoux sur le site de l'association ASTRID [ASsociationTRévouxImprimerieDictionnaire]) une collection assez riche des éditions du *Dictionnaire de l'Académie française*, associant les éditions officielles achevées (1694-1935) et éditions non officielles, des contrefaçons réalisées dès le XVII^e s. en Hollande (1695) et d'autres produites tout au long de la seconde moitié du XVIII^e s. des contrefaçons pré-révolutionnaires imprimées notamment à Nîmes et Avignon. En outre, le fonds dispose d'une collection riche de dictionnaires du XIX^e siècle oubliés ou négligés par l'étude linguistique, telles les éditions méconnues des dictionnaires de Boiste et de Laveaux.

Enfin, pour ce qui concerne la représentativité des genres de dictionnaires, un des intérêts du fonds est de posséder à la fois bon nombre d'ouvrages répondant au genre du

«Dictionnaire universel» dans son évolution historique, et plusieurs dictionnaires de spécialités ayant contribué à servir de source à de nombreux dictionnaires oscillant entre le genre du dictionnaire universel et celui du dictionnaire encyclopédique...

Intérêt scientifique et culturel du fonds

Mes premières analyses de l'identité des collections du fonds me permettent donc d'en définir les grands pôles d'intérêt pour le laboratoire et la communauté scientifique des chercheurs susceptibles de consulter nos ressources documentaires. Le premier travail d'envergure concerne les dictionnaires anciens, du XVI^e au XIX^e s., dans des formats très différents - du grand et majestueux in folio au petit in 12° - avec une grande variété de caractères et d'ornements typographiques [Turcan 2003b].

L'ensemble s'analyse selon la logique des collections identifiables en vertu de critères bibliophiliques, bibliographiques ou «méta-bibliographiques» et métalexicographiques.

L'intérêt scientifique et culturel du fonds est indéniable, qu'il s'agisse de contenus restés jusqu'à ce jour ignorés telles les annotations manuscrites signalées ci-dessus, et, pour l'imprimé, l'exemple significatif des mots du vocabulaire révolutionnaire enregistré par certains des éditeurs ayant contrefait la quatrième édition du *Dictionnaire de l'Académie française* (1762) avant les dates officielles de la révolution française, donc bien avant la publication d'éditions de dictionnaires données à la fin du siècle qu'il s'agisse encore de la masse diversifiée des savoirs réunis dans un vaste ensemble de ressources documentaires exploitables par un public diversifié dès lors que les principes d'une transmission efficace et fiable sur le WEB sont respectés.

L'appréciation de l'identité du fonds passe surtout par la notion de **représentativité** des ressources imprimées, représentativité historique, textuelle et linguistique

- représentativité **historique** par l'épaisseur de la diachronie ouvrages allant du XVI^e au XX^e siècle, les plus anciens imprimés du fond étant des dictionnaires, tel le Calepin de 1568 dont le format *in folio* ne ressemble pas à nos petits carnets désignés du même nom...

- représentativité **textuelle** des genres d'écrits par la diversité des textes concernés, dictionnaires généraux, spécialisés, universels, encyclopédiques, monolingues, plurilingues, français, étrangers ainsi le célèbre dictionnaire italien imprimé à Florence,

le *Vocabulario della Crusca* qui a servi de modèle à la première édition du *Dictionnaire de l'Académie Française* (1694), textes théoriques et didactiques, traités et grammaires dont l'intérêt est de compléter le discours des dictionnaires ou d'avoir contribué à le nourrir et à l'enrichir, textes critiques relevant de la critique littéraire et linguistique et de la métalexicographie historique, avant même l'officialisation de cette discipline...

- représentativité **linguistique** de la diversité des usages langagiers de la langue française à l'image de la richesse de ses composantes, sur le plan historique (textes littéraires du vieux et moyen français), sur le plan des variétés dialectales (textes relatifs aux parlers régionaux), sur le plan des niveaux de langue, avec, face aux textes littéraires, l'argot et les ouvrages du XIX^e siècle concernant les formes vicieuses, etc...

Un fonds à faire connaître pour le faire vivre

Après une révision systématique du catalogue de façon à en permettre à tout public l'interrogation en ligne, le laboratoire a mis en œuvre un travail de catalogage plus détaillé, tenant compte des caractéristiques propres à chaque ouvrage comme les *ex libris* et *ex dono*, indices bibliophiliques des différents possesseurs, tels les anciens fonds de collectionneurs, de bibliothèques religieuses ou encore de bibliothèque de châteaux qui répondent au principe de «traçabilité» des ouvrages : la présence de tampons anciens, d'annotations manuscrites marginales ou sur paperolles, de gravures signées, avec précision de leur emplacement : telle la typologie des ornements typographiques.

Tous ces indices, indispensables pour mieux connaître et identifier les imprimés anciens, contribuer au repérage des exemplaires rares et à l'identification des éditions particulières, retirages d'impressions, contrefaçons, etc..., imposent l'enrichissement du catalogue pour faire connaître les richesses du fonds, autant au grand public qu'à la communauté nationale et internationale des chercheurs intéressés par l'histoire de la langue française et par l'histoire du livre imprimé, notamment sous l'Ancien Régime.

Impératif de préservation et de valorisation du fonds par la numérisation

La priorité de traitement numérique répond à l'impératif de préservation des fonds, avec les dictionnaires anciens dont les reliures sont endommagées et/ou dont le papier se délite, les ouvrages porteurs de notes manuscrites dont l'encre risque de s'effacer avec le temps et d'*ex libris* manuscrits ou étiquetés, enrichissements matériels précieux pour l'histoire puisqu'ils confèrent aux ouvrages conservés une identité particulière susceptible de retracer l'itinéraire de ses anciens lecteurs et, éventuellement, de mieux les connaître par les lectures qu'ils en ont faites, les ouvrages constituant des séries officielles ou parallèles ayant marqué, au-delà de l'histoire de la langue française et de la lexicographie, l'histoire du livre imprimé, les conditions particulières de sa diffusion et donc les succès de sa réception.

Ce fonds intéressant directement les recherches en sciences humaines tout en étant susceptible d'interpeller de façon large un public diversifié, mérite donc bien d'être mieux connu du grand public, français et étranger, d'être partagé par une communauté culturelle tout en étant préservé et valorisé en vertu des exigences de sauvegarde des richesses du patrimoine écrit.

Dès lors que l'identification des composantes de notre fonds confirme la réalité de collections intéressant à des titres différents le patrimoine, le patrimoine international (ouvrages contribuant à une meilleure connaissance de l'histoire du livre ancien, de sa diffusion en Europe et dans le monde entier) ouvrages inscrits dans l'histoire du développement de la francophonie), le patrimoine national (textes concernant l'histoire de la langue et de la littérature françaises), et le patrimoine régional (rôle des ouvrages imprimés en Lorraine par rapport à la place historique de cette région au siècle des Lumières) ouvrages concernant les parlers régionaux), on est en mesure de définir les priorités de traitement numérique selon des grandes étapes. C'est aussi l'identité des collections qui détermine un choix de *numérisation* ne se limitant pas à un corpus d'images enregistrées dans une «numérithèque» statique) la gestion des documents électroniques et de transmission de leurs contenus doit contribuer à valoriser autant leur dimension patrimoniale auprès d'un public différencié que leur fonction de ressources documentaires ouvertes à des orientations natives de recherche en sciences humaines.

Notre laboratoire a la mission d'apporter une valeur ajoutée scientifique aux documents avec, outre l'efficacité et la convivialité des outils d'exploitation automatique des matériaux numérisés et des outils pédagogiques associés, la compétence des chercheurs impliqués dans ce programme de numérisation et de valorisation.

La numérisation ☐ principe d'une valorisation ouverte sur la recherche

Notre objectif est de valoriser notre patrimoine documentaire pour mieux le partager, ce qui implique de le préserver pour mieux en transmettre la mémoire aux générations futures.

Préservation, fiabilité documentaire et partage

Conservation des objets les plus fragiles ou précieux du fonds

Nous devons sécuriser et préserver cette composante du patrimoine documentaire européen que représente notre fonds d'imprimés anciens (qui doit être manipulée et feuilletée le moins possible) pour en assurer la conservation et une réelle valorisation par le partage avec la communauté des chercheurs et le public intéressé, en choisissant des sauvegardes fiables dans le temps et en garantissant une diffusion de qualité sur le WEB qui respecte le livre, ses contenus et les consultants. Or, la mission du laboratoire ATILF est de permettre le développement de recherches diversifiées, en diachronie comme en synchronie, sur l'ensemble des composantes de la langue française en assurant la gestion d'une plate-forme nationale de ressources linguistiques informatisées constituant un authentique patrimoine littéraire et linguistique. L'objectif principal de la numérisation du fonds ancien est d'offrir progressivement, à l'instar de ce qui existe pour la période moderne, une interconnexion la plus diversifiée, la plus riche possible, entre bases textuelles littéraires et lexicographiques pour faciliter à tout utilisateur une hyper navigation efficace entre un texte et un dictionnaire de référence de sa période d'écriture,

condition nécessaire pour mieux comprendre les contenus en maîtrisant les contextes d'usage précis.

Or pour préserver les collections en limitant la dégradation matérielle des volumes fragiles, la numérisation en mode image offre des garanties irréfutables si les travaux sont réalisés dans des conditions adaptées aux spécificités des ouvrages concernés, les ressources du laboratoire étant destinées à des chercheurs soucieux de disposer de documents fiables d'un point de vue scientifique. D'autre part, pour faciliter la publication en ligne de l'ensemble des masses de données ainsi produites et pour rendre plus efficace l'hyper-navigation, les équipes de recherche spécialisées dans l'édition électronique de grands corpus lexicographiques mettent en œuvre des outils et des procédures d'indexation fondés sur la connaissance approfondie de chaque ouvrage. [Wooldridge, 1997] et [Turcan, 1998] et les possibilités d'une exploitation automatique des collections d'images alors constituées. Le programme de numérisation répond donc à l'objectif de valorisation des collections.

Le principe de fiabilité

Le principe absolu de transmission numérique des masses documentaires de textes anciens constituées en bases de données consultables sur ordinateurs, quel que soit le support de transmission (WEB) ou de diffusion (cédérom ou DVD), et quel que soit le public visé, est celui de la fiabilité absolue des images numériques par rapport aux objets d'origine : les documents soumis à la numérisation étant des textes anciens, il est capital qu'en soient conservées absolument toutes les caractéristiques concrètes, formelles (indications mise en page, polices et tailles de caractères, graphies - qu'elles paraissent erronées ou non -, marques typographiques). La réalité de l'impression ancienne, du fait des modalités artisanales de sa réalisation, notamment des particularités de la composition typographique, fait de chaque exemplaire une pièce quasiment unique, constituant au regard de l'histoire du livre un objet précieux pour les études de bibliographie matérielles par exemple pour l'étude des erreurs purement matérielles de typographie. Cet objectif de fiabilité impose une qualité de définition des images (format TIFF 400 DPI pour une analyse fine des détails des ornements typographiques □ 300 DPI en niveaux de gris pour le texte destiné à l'OCR) pour permettre non seulement de lire dans d'excellentes

conditions les documents quels qu'ils soient, mais surtout d'en explorer puis d'en exploiter automatiquement la diversité des contenus. Il est donc exclu de modifier de quelque façon que ce soit, manuellement ou automatiquement, les contenus des images numériques qui correspondent à la spécificité des objets numérisés□ ainsi, pour une uniformisation des lettres *u / v* et *i / j* dans les textes antérieurs au XVIII^e siècle ou pour une transcription des *s* longs, les ligatures et perluettes constituant un ensemble relevant davantage de l'identité typographique que des spécificités linguistiques.

Le partage

Le partage des ressources documentaires correspondant à des imprimés anciens implique une politique de conservation - rénovation des reliures et entretien des ouvrages les plus endommagés - et de préservation des objets précieux, à quelque titre que ce soit, chaque ouvrage de linguistique constituant, outre les spécificités formelles, un témoignage particulier au regard de l'histoire de la langue française, surtout s'il est enrichi d'annotations manuscrites intéressant l'historiographie, témoignages uniques dont la fragilité est extrême.

Préservation [sauvegarde et entretien] et rénovation sont donc indissociables pour garantir un partage des collections avec la communauté des lecteurs diversifiés : de récentes études de bibliographie matérielle sur deux ensembles de lexicographie sérielle, le *Dictionnaire Universel françois & latin dit de Trévoux* (1704-1771) et le *Dictionnaire de l'Académie française* (1694-1935) [Turcan 1999] et [Turcan 2000] ont à cet égard permis de démontrer en toute rigueur l'importance d'une approche complémentaire concernant les contenants et les contenus, ce qui implique d'associer aux compétences techniques des services de numérisation et de documentation, les compétences scientifiques des chercheurs spécialistes du livre ancien, soucieux de répondre aux exigences d'une éthique numérique. Nous avons d'ailleurs déjà commenté [Turcan, 2003a] les décalages référentiels constatés hélas sur le site de la BNF entre des images tronquées et des références déracinées de la réalité des textes concernés, écueils dans lesquels il est hors de question que nous nous égarions, conscients des conséquences d'une telle défaillance éthique, hélas aussi constatée ailleurs, ce qui a suscité une première réaction en 2000 [Turcan – Saint-Gérard, 2000].

Nos premières évaluations à partir d'échantillons représentatifs

Pour mieux être en mesure d'évaluer les charges de travail relatives au corpus propre à chaque ouvrage ou à chaque série d'ouvrages constituant une collection particulière au sein du fonds conservé au laboratoire, nous avons d'abord choisi, de façon raisonnée et aléatoire à la fois, des extraits des ouvrages considérés comme représentatifs de l'ensemble destiné à être numérisé. De façon raisonnée a été soumise au service de numérisation du laboratoire pour des essais de numérisation en mode image avec OCRisation une première série d'échantillons rendant compte des spécificités de chaque composante des collections, des difficultés propres à un ouvrage relevant de la lexicographie sérielle représenté dans le fonds, des difficultés liées à la taille des grands in folio ou aux polices de caractère, aux déficiences dans la qualité d'impression, à l'état du papier trop jauni ou maculé de tâches d'humidité (pages extraites des grands dictionnaires de chaque grande période de l'histoire de la langue française □ Calepin, Nicot, Richelet, Furetière, Ménage, Académie 1, Trévoux, Académie 5, Boiste) □ s'ajoutaient à cela les modalités d'apprentissage des OCR pour la reconnaissance des caractéristiques typographiques telles que les ligatures, les s longs et la perluette (fin 2002-début 2003).

Une seconde série de travaux sur corpus fermé a été définie pour sa double représentativité nationale (histoire de la lexicographie et de la langue française au XVIII^e s.) et régionale (portée d'une impression lorraine à vocation internationale dans l'Europe francophone des Lumières), l'édition nancéienne du *Dictionnaire de Trévoux*, imprimée de 1738 à 1742 sur les presses du libraire Antoine □ l'objectif de gestion des ornements typographiques et celui d'essais d'OCRisation des vedettes pour une indexation automatique des articles marqués nous a conduit à choisir, faute d'outils adéquats au sein du laboratoire, de faire numériser en sous-traitance un de nos deux exemplaires de cette édition, pour pouvoir disposer dès le mois d'août d'images numériques de qualité, prétraitées par un logiciel de redressement ou restauration des images (bookrestorer). C'est grâce à une subvention du CPER [Contrat Plan Etat Région] de Lorraine que cette tranche de travaux concernant 10000 pages in folio a pu être commencée, ce qui nous permettra en outre de répondre à l'engagement du laboratoire pour sa participation aux manifestations culturelles sur le thème « □ Nancy au siècle des Lumières □ », en 2005.

Nous envisageons, ensuite, au gré des subventions dont nous espérons bénéficier, de traiter selon le même principe l'ensemble du fonds, en définissant toutefois, là encore, des étapes, par exemple, en isolant différentes catégories de volumes selon leur genre ou leur typologie, différentes composantes d'un même ensemble sériel comme les paratextes - textes de préfaces, avertissements, épîtres - , les listes de noms d'auteurs ou d'étymons, notes marginales imprimées ou manuscrites, notes et additions, tables...), sans négliger la constitution de fichiers limités aux marques et ornements typographiques d'ouvrages pour lesquels une étude thématique systématique s'impose (par exemple, les gravures, bandeaux et frises, culs de lampe et lettrines d'ouvrage imprimés à Nîmes ou à Avignon au XVIIIe siècle) [Turcan 2003b].

Mais l'ampleur des moyens nécessaires pour faire numériser notre fonds, ne nous empêche pas de progresser dans notre démarche fondamentale qui est épistémologique : nous avons soumis un dossier d'aide à la numérisation auprès du ministère de la Culture, dans le cadre des appels à proposition 2003, et nous restons encore dans l'expectative...

Une finalité épistémologique

Il est évident que notre fonction, en tant que laboratoire de recherche, n'est pas de nous limiter à produire de belles images, à stocker des données pour ne constituer qu'une simple « numérisation », c'est-à-dire une base cumulative d'images, qui ne serait alors qu'un substitut de bibliothèque sans le charme de la consultation de nos vieux in folio. Notre programme de numérisation est d'abord sous-tendu par une démarche épistémologique alliant plusieurs axes de recherche : une recherche préalable en matière de traitement informatique de fichiers images destinés à différentes possibilités d'exploration et d'exploitation automatique et de procédures de diffusion électronique ; une recherche de type bibliographique ou metabibliographique, pour que chaque ouvrage numérisé soit accompagné d'une notice présentant son identité, non pas par un simple coupé-collé de notices déjà faites sur d'autres exemplaires, mais une fiche signalétique individualisée.

D'autre part, en vertu de la connaissance intime que nous avons des fonctionnements textuels de nos vieux dictionnaires, nous sommes en mesure de préparer tout un travail de réflexion sur les différentes modalités d'indexation non pas seulement limité aux vedettes

et sous-vedettes des articles, mais étendue à d'autres champs informationnels, pour proposer des procédures de balisage qui donnent accès, sinon à l'ensemble des contenus, du moins à une partie qui soit encodable. Par exemple, nous avons commencé à réfléchir aux modalités de traitement automatique de la synonymie dans l'optique de travaux portant sur la reformulation des données récurrentes, identifiables par des critères formels, donc balisables de façon semi-automatique. D'autre part, il nous importe de constituer, indépendamment des bases textuelles lexicographiques, des bases réunissant l'ensemble des ornements typographiques (annexe 5. 2.) [Turcan, 2003b], ce qui implique, là encore toute une réflexion méthodologique sur les principes choisis d'indexation des ornements et implique un choix de GED (gestion de documents électroniques) qui n'est pas le même que pour les images numériques de textes.

Il va de soi que, à l'instar de ce que nous pourrions faire sur la base des premiers corpus d'images numérisées, il sera progressivement possible de mener à la fois des requêtes relevant de l'analyse des marques typographiques permettant de mener à terme les recherches de bibliographie matérielle qui nous sont indispensables pour mieux caractériser une partie du fonds et des requêtes relevant de l'analyse des corpus par la mise en œuvre de bases de données de nomenclatures (vedettes et sous-vedettes, reprises de vedettes par la double prise en compte des positions) □ d'autre part, si le processus de reconnaissance des caractères typographiques et de gestion d'unités balisées le permet, nous envisageons diverses formes d'exploitations assorties de critères textuels minimaux et d'indices de contextualisation que favorise l'hyper-navigation d'une base à l'autre □ enfin, notre objectif est de mener des requêtes relevant du traitement automatique d'une langue telle qu'elle est représentée dans un dictionnaire ancien.

Les premières phases de numérisation sont donc pour nous une étape nécessaire à une valorisation qui passe par la mise en œuvre d'un projet scientifique, ce qui relève véritablement des compétences d'un laboratoire de recherche en traitement automatique de la langue, tout en ouvrant sur la pluridisciplinarité des sciences humaines.

Références □

[Labarre 1975] A. Labarre (1975), *Bibliographie du Dictionarium d'Ambrogio Calepino (1502-1779)*, Valentin Koerner, Baden-Baden, 1975.

[Turcan, 1998] I. Turcan (1998), «Balisage formel ou balisage fin pour les dictionnaires anciens informatisés: objectifs et implications méthodologiques. L'exemple du *Dictionnaire de l'Académie*

Française (1694) et des bases échantillons des dictionnaires de Gilles Ménage (1694) et de Thomas Corneille (1694)», colloque international sur les problèmes de balisage de dictionnaires électroniques, Limoges, 1998, Actes du colloque sous forme électronique publiés en janvier 1999.

[Turcan, 1999] I. Turcan (1999), Présentation des différentes éditions du *Dictionnaire de Trévoux* recensées à Nancy (INALF), Lyon (BM), Paris (BNF), Bourg-en-Bresse (BM) et Trévoux : comparaison des éditions et principes méthodologiques pour une nouvelle approche de l'ordre canonique des éditions trévoltiennes, parisiennes et nancéiennes», Colloque international sur «Connaissance et rayonnement du *Dictionnaire Universel ... de Trévoux* (1704-1771)», publication des *Actes*, sur <http://www.univ-lyon3.fr/siehdaweb/Trévoux-accueil.htm>.

[Turcan, 2000] I. Turcan (2000), publication du cédérom réunissant les huit éditions officielles achevées du *Dictionnaire de l'Académie Française* (1694-1935), sous notre responsabilité scientifique, aux éditions Redon (diffusion Le Robert).

[Turcan, 2001] I. Turcan, (2001), texte de présentation du cédérom publié aux éditions Redon, le *Grand Atelier Historique de la langue française* (diffusion Le Robert) éd. électronique sur www.dictionnaires-france.com.

[Turcan, 2003a] I. Turcan (2003), «Édition électronique de textes imprimés anciens : connaissance de l'histoire du livre et respect numérique», contribution au colloque consacré à *l'Éthique Numérique*, Saint-Cyr-sur Mer, mai 2003 (édition électronique sur le site du colloque).

[Turcan, 2003b] I. Turcan (2003) «Édition scientifique d'ouvrages anciens sur support électronique : traitement numérique des ornements et marques typographiques dans le programme de numérisation des collections d'ouvrages anciens du laboratoire ATILF, UMR 7118, Nancy II», participation à la XIV^e Conférence Européenne TeX "Retour à la typographie", Brest, 24-27 juin 2003 (publication électronique des actes sur le site du colloque).

[Turcan, 2003c] I. Turcan (2003) *Les dictionnaires de l'Ancien Régime*, Editions Desjonquères (à paraître, fin 2003) cf. le chapitre 2 de notre ouvrage sur le dictionnaire, livre-bibliothèque

[Turcan-Pierrel, 2003] Isabelle Turcan et Jean-Marie Pierrel (2003), «Valorisation du patrimoine littéraire et linguistique. Analyse d'expériences de numérisation et projet d'avenir en ré-édition électronique de textes et de dictionnaires», *Document Numérique*, Vol. X, en sept. 2003 (à paraître).

[Turcan-Saint-Gérand, 2000] I. Turcan et Jacques-Philippe Saint-Gérand (2000), «Une bonne version électronique de dictionnaires anciens : Définitions, objectifs et risques» MANIFESTE POUR LE RESPECT DES CONSULTANTS, texte publié sur «Le Net des Etudes françaises» ACRE (Toronto) et en annexe au texte «Les acquis des premiers travaux effectués concernant les différentes formes d'édition électronique des dictionnaires anciens et les projets en cours.» (I. Turcan), in Actes du colloque NTIC-SHS organisé à Lyon, le 27 mars 2001 sur *Les nouvelles technologies de l'information et de la communication et les sciences humaines*, Lyon, Service commun de la Recherche, 2002, p. 25-36.

[Wooldridge, 1992] T. R. Wooldridge (1992), *Le Grand Dictionnaire François-Latin (1593-1628). Histoire, types et méthodes*, Paratexte, Toronto, 1992.

[Wooldridge, 1997], T. R. Wooldridge (1997), «Balisage un texte, c'est le penser» le cas du Dictionnaire de l'Académie Française, mai 1997 (éd. électronique).

Annexes

1. Tableau de synthèse du nombre d'ouvrages anciens du fonds (NB. ne sont pas ici pris en compte les ouvrages sans date, ni ceux de la première moitié du XXe siècle).

Siècle	Nombre d'ouvrages	Estimation du nombre de pages	Format le plus petit	Format le plus grand
16 ^e siècle	3	3600	39	39
17 ^e siècle	14	9000	13	37
18 ^e siècle	102	180000	14	40
19 ^e siècle	400	460000	12	34
Somme	519	766000		

2. Schématisation de notre conception d'une numérisation ayant pour finalité la construction de bases de données complémentaires.