Numérisation et valorisation des collections, AS-CNRS 96 Rapport d'activités

Abdel Belaïd, Hubert Emptoz, Georges Vignaux

CNRS-Département Sciences et Technologies de l'Information et de la Communication

Réseau thématique pluridisciplinaire

RTP 33 : « Document et contenu : création, indexation, navigation»

Action spécifique 96 : « Numérisation et valorisation des collections »

Animateurs :

Abdel Belaïd (LORIA, UMR 7503 du CNRS, Nancy)
Hubert Emptoz (RFV-ISDN, INSA Lyon)
Georges Vignaux (UPR 36 du CNRS, Paris)

Février 2004

Sommaire

II – Georges Vignaux
La notion de « collection » : genèses, développements, valorisations ……..... 27

Abdel BELAÏD
(LORIA, UMR 7503 du CNRS, Nancy)

Techniques de numérisation

1. Introduction

Ce rapport présente l'état des réflexions faites dans le cadre de l'Action Spécifique « Numérisation et Valorisation des Collections » du réseau thématique pluridisciplinaire 33 du département STIC du CNRS. Il tend à préciser en premier lieu le vocabulaire employé qui n'est pas toujours communément partagé par les différentes communautés du document, et à décrire en second lieu l'apport des travaux de recherche en numérisation dans la vaste opération lancée depuis peu dans la mise en ligne des documents patrimoniaux.

L'AS a jugé d'emblée les trois termes : numérisation, valorisation et collection problématiques et qu'ils devaient faire l'objet d'un travail de définition rigoureux. Aussi, trois voies de réflexion ont alors été proposées :

La problématique de la constitution des collections est centrale dans notre approche, mais il importe de cerner la notion de collection ? On sait que les collections anciennes constituent aujourd'hui un patrimoine considérable, d'où, plusieurs types de questions essentielles : qu'appelle-t-on « préservation » ? et qu'appelle-t-on « conservation » ?

Comment établir un partage des collections ? Quelles méthodes de consultation promouvoir (production des annotations, partage des annotations entre lecteurs) ? Il importe donc de reformuler le problème de la valorisation des collections et notamment de reconsidérer les interactions entre :
modalités de codage et de structuration,
pratiques et dispositifs de lecture.

Il importe enfin de faire un recensement et un bilan des actions de numérisation entreprises dans un certain nombre de lieux et de circonstances : quelles plate-formes ? et pour quels usages ? Quels impacts en retour sur les méta-données et les DTD ?

2. Actions de numérisation

2.1 Généralités

2.1.1 Les étapes de la chaîne de numérisation

Selon la nature de l'information, on peut distinguer trois niveaux de traitement du document :

Un niveau image : au cours duquel le document est «translaté » de son support papier vers un support numérique plus apte à la manipulation informatique. Ce niveau est en lien direct avec l'action de constitution des collections et va influer sur la notion de préservation et de conservation des documents. En effet, c'est au cours de cette action que des choix importants vont être faits sur le passage de l'analogique au numérique : en termes de résolution, de définition, de précision, de compression, etc.

Un niveau information de base : conduisant à l'extraction du contenu textuel. Cette information de base est obtenue par les opérations d'OCR[1], d'ICR[2] et de rétro-conversion des structures. C'est à ce niveau qu'interviennent les plate-formes, que l'on s'aperçoit de leurs capacités, et que l'on découvre les vraies avancées de la recherche. Selon que le document est imprimé ou manuscrit, les difficultés sont différentes. On enregistre actuellement des progrès substantiels en reconnaissance de l'imprimé et la recherche demeure active sur le manuscrit.

[1] Optical Character Recognition.
[2] Intelligent Character Recognition.

Un niveau méta-informations : plus lié à l'action de compréhension du contenu. Ce niveau met en avant la notion de méta-données. C'est sur ce niveau qu'il sera alors possible de bâtir des ontologies et des processus de navigation et de lecture de documents.

2.1.2 Les types de documents

Selon que le document est moderne ou ancien, les objectifs de la numérisation sont différents. En effet :

Les documents modernes sont typographiés avec des typographies actuelles, compréhensibles par les OCR. Bien que quelques difficultés puissent subsister ça et là dans la reconnaissance de certains formats ou modes dues à l'emploi de polices particulières (comme l'italique ou le script) ou d'arrangements spéciaux (tableau, formulaire, équation, etc.), on peut aujourd'hui considérer que la thématique est assez mûre et qu'il existe un marché pour une grande classe de documents imprimés. En d'autres termes, il existe un savoir faire et des logiciels industriels capables d'atteindre des performances très élevées. L'investissement de la recherche dans ce domaine se situe davantage dans l'adaptation des logiciels existants et dans leur combinaison au sein de plate-formes d'accueil.

Les documents anciens contiennent en plus des difficultés habituelles d'autres plus spécifiques qui rendent les traitements encore plus difficiles. L'écriture se trouve parfois plus proche du manuscrit et la composition n'est pas conventionnelle et souvent pas normalisée, empêchant tout traitement générique. Il n'existe donc pas d'outils génériques tels que les OCR pour les traiter et la recherche se trouve dans son rôle le plus complet pour proposer des solutions innovantes. Par ailleurs, n'étant jamais sûr de pouvoir fidéliser l'opération de reconnaissance, l'original (facsimilé) reste plus que jamais le document de référence : il faut le conserver, le montrer, le rendre accessible. Dans ce sens, la reconnaissance sera vue comme une aide et un moyen d'accès au document original.

2.2 Pour les documents modernes, un marché, une industrie

2.2.1 Un matériel dédié et varié

Il existe aujourd'hui une panoplie de matériels adaptés à la numérisation de documents dans toutes ses formes avec des possibilités variées de résolution et de définition. Le site du Ministère de la culture[3] fait un état détaillé de cette technologie dont nous rappelons l'essentiel dans la suite.

Concernant le matériel, le scanner est aujourd'hui l'appareil dédié à la saisie de documents. Muni de capteurs CCD[4] sensibles à la lumière, le scanner restitue la couleur diffusée par le document en tout point. La restitution est mesurée à l'aide de deux quantités : la résolution (ou précision exprimée en points par pouce : ppp) et la dynamique (étendue de la gamme de couleurs ou de niveaux de gris que peuvent prendre les points). Selon le type de scanner, les modes de fonctionnement sont différents : par balayage (ponctuel ou linéaire) ou par analyse matricielle. On distingue plusieurs types de scanner adaptés aux différents usages : a) bureautique, dédiés à la saisie à plat, offrant des résolutions importantes mais sont souvent limités à des formats réduits de documents, b) de livre, dédiés à la numérisation de livres ouverts ayant des formats importants et pouvant donc s'adapter à l'inclinaison et au bombage des feuilles, c) de microfilm ou microfiche, proposant une numérisation en mode bitonal, plus rarement en niveaux de gris, s'échelonnant entre 200 et 400 ppp, d) de

[3] <http://www.culture.gouv.fr/culture/mrt/numerisation/fr/dll/techn.htm#04>.
[4] Charge Coupled Device.

diapositive ou de transparent. Une bibliographie complémentaire sur le sujet peut être consultée[5][6][7][8].

2.2.2 Des techniques de stockage adaptées

L'ouverture du marché du document sur la GED[9] a conduit tout naturellement à réfléchir aux moyens de stockage et d'archivage de grandes quantités de documents. Même si le stockage électronique n'est pas sans poser de problèmes législatifs, les entreprises veulent alléger la gestion des documents papier et des documents multimédia qui sont trop encombrants. Aussi, des solutions matérielles variées ont été proposées pour différents types de documents.

Concernant le support, on distingue aujourd'hui trois types de support différents : magnétique, optique et chimique offrant des caractéristiques relatives à la vitesse d'accès, à la sécurité de l'information et la pérennité très différentes. Tous ces supports autorisent des formats particuliers avec des capacités de rangements différents. On peut trouver sur le site de l'ENSSIB[10] une description détaillée de ces solutions.

Concernant les techniques de compression d'images, la panoplie est là-aussi très large autorisant ou non la perte d'information. Mais très peu d'entre elles sont adaptées aux images de texte. En effet, même la technique JPEG[11] qui semblait s'approcher le plus de la vision humaine (compression psycho-visuelle retirant l'information non perçue par l'œil humain), a tendance à affecter la lisibilité du texte (à cause d'une mauvaise restitution lors de la décompression des contours des caractères) même dans des taux très réduits 1:10. L'amélioration de cette technique au travers de la nouvelle norme JPEG2000 (transformation en ondelettes qui s'appuie sur une approche pyramidale multi-résolution), n'a pas pu non plus profiter aux images de documents. Les images compressées souffrent d'un effet de flou particulièrement accentué si le taux de compression est élevé. Aussi, des solutions plus adaptées aux images de texte ont été vite avancées en dissociant les images binaires et les images de niveau de gris :

Concernant les images binaires, après des tentatives, le nouveau standard qui émerge est le JBIG2[12], offrant des taux de compression variables. JBIG2 agit comme un système expert : utilise une connaissance a priori sur le contenu avec une séparation des médias texte/graphique et exploite la redondance des caractères dans le texte. Cette alternative est très bonne car les caractères sont considérés en tant que forme élémentaire complète et non comme une image de pixels. La compression par redondance des formes génère une image artificielle où toutes les formes redondantes sont substituées par une forme générique unique. Mais l'efficacité de cette technique reste aléatoire car elle dépend d'une part de la régularité des formes dans les images, d'autre part de l'efficacité des techniques d'appariement des formes, et enfin de l'abondance ou non des médias différents.

---

[5] Agfa-Gevaert, An Introduction to Digital Scanning (Digital Colour pepress ; 4), Mortser (Belgique), Agfa-Gevaert, 1994.
[6] Michael lesk, *Practical Digital Libraries : Books, Bytes and Bucks*, San Francisco, Morgan Kaufmann, 1997.
[7] William Saffady, *Computer Storage Technologies : A Guide for Electronic recordkeeping*, Prairie Village, Kansas, ARMA International, 1996.
[8] <http://www.enssib.fr/autres-sites/dessid/dessid99/gedbouan.pdf>.
[9] Gestion Électronique de Documents.
[10] <http://www.enssib.fr/autres-sites/dessid/dessid99/gedbouan.pdf>.
[11] Joint Picture Expert Group.
[12] P. HOWARD, « Lossless and lossy compression of text images by soft pattern matching », *Proc. of the IEEE Data compression Conference*, p. 210-219, 1996.

Concernant les images de niveaux de gris et en couleur, quelques solutions ont été proposées par des grandes sociétés américaines, comme la compression DjVu[13] d'AT&T et TIFF-FX de Xerox. Ces nouvelles approches permettent d'atteindre des taux de compression supérieurs à 1:100 en utilisant des informations sur la spécificité des images de documents. Ces techniques séparent les plans de l'image ainsi que les différents médias, puis appliquent sur chacun d'entre eux la méthode de compression la plus adéquate. Ainsi, par exemple, l'arrière plan débarrassé de traits et de caractères peut être compressé par une méthode de type JPEG avec un taux relativement élevé. Ce type de méthode a été poursuivi dans le cadre du projet DEBORA[14].

2.2.3 Des logiciels de reconnaissance de caractères performants

Les techniques de lecture automatique de documents ont beaucoup évolué et mûri cette dernière décennie et l'on voit fleurir sur le marché des progiciels de moins en moins chers, de plus en plus complets et fiables offrant des solutions de numérisation très fidèles. Ces progiciels, appartenant à la famille des OCR[15], sont aujourd'hui capables de distinguer les différents médias dans le document (texte, graphique et photographie), d'identifier les structures linéaires et tabulaires, de faire face à une variation importante de la typographie, d'interpréter et de restituer plusieurs styles éditoriaux. Des bancs d'essais sont effectués couramment sur les dernières versions de ces progiciels montrant les nouvelles possibilités offertes et donnant une évaluation correcte et précise de leur capacité de reconnaissance en termes de taux de confiance, de précision et de vitesse d'exécution, par type de document et de typographie utilisée[16][17][18][19].

2.2.4 Des plate-formes d'intégration d'OCR évoluées

Les outils OCR ont également progressé sur le plan de l'intégration puisqu'ils existent aujourd'hui sous la forme d'API facilement intégrables dans un système ou sous la forme de kits de développement offrant des possibilités d'extension et de coopération avec d'autres API permettant l'augmentation de l'efficacité de l'ensemble.
Cependant, leur intégration dans une chaîne de numérisation réelle, en milieu industriel, nécessite de prendre en compte d'autres contraintes liées à la production telles que la volumétrie, la cadence de numérisation et surtout la haute qualité des résultats exigée en bout de chaîne. Peu d'expériences ont été rapportées pour rendre compte de l'efficacité de ces intégrations, mais il est évident d'après les premiers essais qui ont été effectués en entreprise par l'équipe READ à Nancy[20], que leur apport est considérable en termes de gain de productivité, comparé par exemple à la double saisie manuelle.

---

[13] L. BOTTOU *and al.*, «High quality document image compression with DjVu », *Electronics Imaging*, 7(3):410-428, 1998.
[14] F. LEBOURGEOIS *and al.*, « Compression de documents imprimés numérisés », CIFED'2002, Hammamet, Tunisie, 21-23 oct. 2002, p. 195-204.
[15] Optical Character Recognition.
[16] S. V. Rice, J. Kanai, and T. A. Nartker, « An Evaluation of OCR Accuracy », *ISRI 1993 AnnuaI Research Report*, University of Nevada, Las Vegas, April 1993, 9-31.
[17] S. V. Rice, J. Kanai, and T. A. Nartker, « The Third Annual Test of OCR Accuracy », *ISRI 1994 Annual Research Report*, University of Nevada, Las Vegas, April 1994, 11-38.
[18] S. V. Rice, F. R. Jenkins, and T. A. Nartker, « The Fourth AnnuaI Test of OCR Accuracy », *ISRI 1995 Annual Research Report*, University of Nevada, Las Vegas, April 1995, ll- 49.
[19] S. V. Rice, "The OCR Experimental Environment, Version 3", *ISRI 1993 Annual Research Report*, University of Nevada, Las Vegas, April 1993, 83-86.
[20] A. Belaïd, L. Pierron, L. Najman et D. Reyren, « Numérisation de documents : principes et évaluation des performances », Cours INRIA sur les Bibliothèques numériques, Collection ADBS, octobre, La Bresse, 2000.

2.2.4.1 Intérêts de la combinaison

Performance imprévisible de l'OCR : Les logiciels d'OCR sont devenus de plus en plus des systèmes experts à bases de règles intégrant différentes techniques de reconnaissance et entraînés pour maximiser la performance globale. Parce que ces systèmes sont à bases de règles, leur performance n'est pas prévisible de manière théorique. Ce type de système a un comportement qui peut changer de manière imprévisible voyant la performance parfois chuter sur des cas limites. Il suffit parfois d'une petite variation pour que deux caractères, en général bien différenciés, soient confondus. Ces systèmes utilisent des sources d'informations différentes et étagées (taille des caractères, inclinaison des caractères, etc.) pour reconnaître. A cause de la multiplicité de la typographie, ces sources d'information ne peuvent pas être fiables, conduisant à des défaillances dans les cas limites. Ces défaillances sont en plus amplifiées d'étage en étage car chaque étage prend pour acquis les résultats de l'étage précédent.

Amélioration des performances individuelles. Les OCR n'atteindront jamais les 100% de bonne reconnaissance ! On observe par ailleurs, que différents moteurs d'OCR produisent des erreurs différentes. L'objectif de la combinaison est de tirer parti des avantages de chaque OCR et d'écarter leur faiblesse. Dans les différentes études effectuées par[21], il est montré que de l'ordre de 50% d'erreur est éliminée par la combinaison de plusieurs OCR ayant des taux de reconnaissance individuels de l'ordre de 97%. Cela étant, ce gain ne peut être atteint que dans la mesure où les erreurs proviennent des OCR et non de la qualité de l'image, et où les OCR sont de bonne qualité.

2.2.4.2 Types d'évaluation des performances

Il y a deux manières d'évaluer les OCR suivant le type d'information que l'on peut ou que l'on souhaite extraire d'un OCR. On peut considérer l'OCR comme une boîte noire (système fermé) duquel seul le résultat de la reconnaissance des mots est accessible. Dans ce cas, l'évaluation est faite de manière globale. Dans le cas où l'on dispose d'une information plus précise (par exemple, position dans l'image des caractères extraits par l'OCR), on peut affiner la mesure de la performance.

Évaluation globale : C'est le cas où la seule information disponible est la liste des caractères et des rejets. La seule évaluation possible est de calculer le taux d'erreur Γ_{err} ou le taux de reconnaissance Γ_{rec}. Pour se faire, on a besoin de trouver les erreurs. Cela revient à mesurer, par mise en correspondance entre les couples de chaînes de références et de résultats, les ajouts, les suppressions et les substitutions. Si l'OCR se comporte comme une boîte noire, il est impossible, à cause de ce problème de non unicité, de faire un choix entre deux possibilités de coût identique et donc de connaître précisément le nombre de substitutions et de suppressions. Ceci ne permet pas de détailler l'erreur et surtout de déterminer l'erreur de substitution qui dénote la confusion. Or, ce type d'erreur est souvent difficile à détecter par ailleurs car elle conduit à des mots souvent validés par un dictionnaire. Les deux autres erreurs sont moins fondamentales car il y a plus de chance que l'erreur soit flagrante. On souhaiterait utiliser le taux de confusion afin d'évaluer la performance du système. On peut en estimer un à partir de la solution de correspondance de coût minimum. A partir de cette mise en correspondance, on peut déterminer tous les taux précédents, le taux de rejet étant donné par l'OCR.

Évaluation locale : Dans ce cas, on dispose de la position des caractères (glyphes) dans l'image. On maîtrise la réponse sur chaque caractère et on peut déterminer précisément le type d'erreur commise sur chaque caractère. Cela permet par conséquent de calculer le taux de substitution qui posait problème dans le cas de l'évaluation globale.

[21] S. V. Rice, J. Kanai, and T. A. Nartker, « An Evaluation of OCR Accuracy », *ISRI 1993 AnnuaI Research Report*, University of Nevada, Las Vegas, April 1993, 9-31.

2.2.4.3 Combinaison d'OCR

Comme nous l'avons rappelé en introduction, les OCR n'atteindront jamais le taux parfait de 100% de bonne reconnaissance. De plus, ayant remarqué que différents moteurs d'OCR produisent des erreurs différentes, cela a conduit à proposer des techniques de combinaison de moteurs OCR afin d'améliorer le résultat global. L'objectif est de tirer parti des avantages de chaque OCR et d'écarter leur faiblesse[22]. Cela étant, des améliorations de performances ne sont possibles que si les erreurs sont en quantité raisonnable, et proviennent des moteurs d'OCR et non d'une mauvaise qualité de l'image.

La combinaison d'OCR nécessite de connaître les éléments suivants :

Degré d'intégrité : il faut tenir plus ou moins compte du résultat produit par l'OCR en fonction de la confiance que l'on peut avoir dans le moteur. Ceci nécessite donc de connaître ses forces et ses faiblesses.

Détail des réponses : la nature des réponses pour chaque glyphe peut varier depuis un simple nom de caractère jusqu'à la fourniture d'un degré de confiance pour chaque lettre de l'alphabet, en passant par une liste ordonnée de lettres possibles. On imagine que le détail des réponses influe sur la stratégie de combinaison.

Indépendance des moteurs : dans l'idéal, il faut choisir des moteurs les plus indépendants possibles de manière à ne pas trop donner d'importances aux décisions prises sur des critères identiques.

2.2.4.4 Stratégies de combinaison

Il existe plusieurs stratégies en fonction du niveau de détail fourni au niveau de la lettre :

Simple caractère : ce niveau de détail ne permet que la stratégie de combinaison la plus simpliste, celle du vote majoritaire. Le caractère retenu sera celui le plus répandu parmi toutes les réponses des OCR.

Liste ordonnée de caractères : la méthode la plus répandue, appelée « Borda Count » donne une note à chaque réponse en fonction du nombre de fois où la réponse est présente dans les OCR, pondérée par la position dans la liste (rang).

Degré de confiance : dans ce cas, il est d'abord nécessaire de normaliser les degrés de confiance qui ne sont pas nécessairement des probabilités. Suivant la dépendance ou non des classifieurs, on pourra utiliser soit les techniques basée sur la théorie de probabilités conditionnelles (de Bayes), soit celles basées sur la combinaison de la croyance (Dempster-Schafer)[23].

2.2.4.5 Contrôle de l'OCR

Il n'est pas pensable d'éliminer le taux de rejet et de forcer le système à décider de l'identité de chaque échantillon puisque les décisions faites dans le cas incertain peuvent causer une augmentation disproportionnée du taux d'erreur. Ceci étant, il serait naturel pour la mesure de la performance d'un OCR de contenir quelques relations entre les taux de reconnaissance/rejet et le taux d'erreur. Les taux d'erreur et de rejet sont corrélés (en première approximation) en raison inverse selon la formule : $\Gamma err = 1/\Gamma rej$ ce qui veut dire que la seule façon de diminuer le taux de confusion est d'augmenter le taux de rejet. Il y a trois moyens plus ou moins fins de contrôler le taux de rejet :

---

[22] A. Dengel, R. Hoch, F. Hönes, T. Jäger, M. Malburg and A. Weigel, "Techniques for Improving OCR Results", in *Handbook of Character Recognition and Document Image Analysis*, Eds. H. Bunke and P.S.B. Wang, 227-258, Chapter 38.

[23] E. Mandler and J. Schürmann, "Combining the Classification Results of Independent Classifiers based on the Dempster/Shafer theory of evidence", in *Pattern Recognition and Artificial Intelligence*, ed. by E.S. Gelsema and L.N. Kanal, Elsevier Science Publishers B.V., North Holland, 1998, 381-393.

Si on dispose d'un taux de confiance par caractère, on peut contrôler le taux de rejet par rapport à cette confiance. Plus le seuil de confiance est élevé plus le taux de rejet est important de l'ordre de 3% à 5% mais le taux de confusion reste relativement significatif de l'ordre de 1% à 1 pour mille. La conséquence est que le nombre de caractères à corriger devient important et qu'au-delà d'un certain taux de caractères à corriger, il devient économiquement plus rentable de ressaisir la totalité du texte, ce qui rend inutile l'utilisation de l'OCR.

Si au contraire, on dispose d'une appréciation dans une échelle donnée, le contrôle se fait de manière approximative par rapport à l'un des paliers proposés
Si on a le contrôle sur l'apprentissage, et plus précisément sur la distribution des échantillons, on peut agir plus finement sur le rejet en fonction d'un taux de recouvrement plus important des classes.

Dans le cas des OCR du commerce, on ne peut pas parler de contrôle du rejet, car d'une part, on a peu d'information sur la reconnaissance, et d'autre part on a peu de connaissance sur le fonctionnement de l'apprentissage. Donc, on peut soit utiliser les appréciations ou taux de confiance sachant que l'on ne sait même pas à quoi ils correspondent réellement, soit simuler un taux de confiance a posteriori. On utilise une matrice de confusion calculée à partir de la distance d'édition. Mais ces taux de confiance seront une simple estimation car un même caractère aura toujours le même taux de confiance.

2.2.4.6 Une expérience de plate-forme industrielle

S'inspirant de la notion d'Adapter du livre Design Patterns, l'équipe READ[24] à Nancy a proposé une classe OCRAdapter réalisant l'intégration de plusieurs OCR du commerce. C'est une classe abstraite permettant d'associer un ensemble de méthodes. Pour l'incorporation des résultats des OCR, des procédures d'encapsulation, de paramétrisation et d'harmonisation des résultats ont été développés. L'équipe a proposé une DTD appelée XmlLayout permettant de faciliter cette harmonisation. Enfin, contrainte par des taux de performance élevés exigés par la société avec laquelle elle a collaboré, l'équipe a proposé des techniques spécifiques pour améliorer les performances des OCR sur des classes spécifiques de documents. Grâce à des heuristiques, l'équipe atteint les performances exigées par la société, permettant à celle-ci d'améliorer sa chaîne de saisie manuelle. En effet, partant de performances très élevées de la saisie manuelle: une vitesse de l'ordre de 4000 à 5000 caractères/heure en simple saisie, avec une qualité de 2/1000 (erreurs par caractères saisis), et 2000 caractères/heure, en double saisie, avec une qualité de 2/10000, l'équipe a atteint en mode automatique, des performances de l'ordre de 1/10000 avec un rejet de l'ordre de 7%, sur des documents de bonne qualité. Ce niveau de qualité a été atteint par combinaison d'OCR ayant des performances individuelles de l'ordre de 1/100.

3. Pour les documents anciens, des adaptations et des recherches sont nécessaires

Les documents anciens sont des documents d'archives rédigés à une autre époque et obéissant donc à des règles typographiques et de composition différentes de celles appliquées sur les documents modernes.

En effet, l'image numérisée n'est pas structurée. Elle est souvent très tonale, à niveaux de gris ou en couleur. Elle peut comprendre des annotations dans les marges, des illustrations, des lettrines, voire même des écritures manuscrites. Il convient donc de revoir la chaîne de traitement d'images pour en adapter les techniques sur une nouvelle race de documents.

[24] <http://www.loria.fr/equipes/read>.

3.1 Pré-traitement

Le document ancien pose en préambule un problème d'acquisition certain dû d'une part à son positionnement sur le scanner, créant des inclinaisons, des bombages et des pliures du papier, et d'autre part à son contenu hétérogène. Le processus de vieillissement fait apparaître des tâches d'humidité, la transparence de l'encre sur les rectos, la fragmentation des contours fins, etc. L. Likforman dresse dans un article récent[25] (voir Tableau 1) une liste de traitements usuels en fonction des types de problèmes rencontrés.

Tableau 1 : Liste des défauts et pré-traitements appropriés, d'après L. Likforman-Suelem

Défaut	Pré-traitement	Référence
Faible ou forte luminosité	Modification d'histogramme	A. Belaïd[26]
Présence de taches	Filtrages passe haut [FEL 00]	M. Feldbach[27]
Points parasites	Filtrages passe-bas filtrages morphologiques	A. Belaïd[28]
Rotation légère de l'image	Calcul de l'angle par projection redressement par re-échantillonnage	Belaïd[29] DEBORA[30]
Courbure de l'écriture sur un bord de l'image	Calcul de la courbure locale re-échantillonnage	DEBORA[31]
écriture fragmentée	Filtrages (passe-haut, morphologiques, passe-bas)	M. Feldbach[32] DEBORA[33]
Contours de l'écriture flous	Filtrage passe haut filtrage morphologique	I. Lamouche[34]
Écriture du verso apparaissant sur le recto	Combinaison des images recto et verso	Lamouche[35] R. Line[36]

[25] L. Likforman-Sulem, "Apport du traitement des images à la numérisation des documents manuscripts anciens", Numéro special de *Document Numérique*, janvier 2004.
[26] A. Belaid, Y. Belaid, « Reconnaissance de formes : méthodes et applications », Interéditions, 1992.
[27] M. Feldbach « Generierung einer semantischen reprasentation aus abbildungen handschritlicher kirchenbuchaufzeichnungen », Diplomarbeit, Otto von Guericke , Universitat Magdeburg, juillet 2000.
[28] Voir note 26.
[29] Voir note 26.
[30] R. Bouché (coord.), « Présentation du projet européen Debora », projet no LB 5608/A, distribué lors de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon, juillet 2000.
[31] R. Bouché (coord.), « Présentation du projet européen Debora », projet no LB 5608/A, document distribué lors de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon, juillet 2000.
[32] Voir note 27.
[33] R. Bouché (coord.), « Présentation du projet européen Debora », projet no LB 5608/A, distribué lors de CIFED'2000, Colloque International Francophone sur l'Ecrit et le Document, Lyon, juillet 2000.
[34] I. Lamouche, C. Bellissant, « Séparation recto/verso d'images de manuscrits anciens », Actes de CNED'96, Colloque National sur l'Ecrit et le Document, Nantes, juillet 1996, p. 199-206.
[35] I. Lamouche, C. Bellissant, *art. cit.*
[36] R. Ducire Lins, M. Guimaraes Neto, Leopoldo França Neto, L. Galdino Rosa, «An Environment for Processing Images of Historical Documents », *Microprocessing and Microprogramming*, 40 (1994), p. 939-942.

3.2 Binarisation

L'opération de binarisation est nécessaire pour séparer le fond du texte si l'image originale est en niveaux de gris ou en couleur. Elle consiste à produire une image à deux tons : clair pour le fond, et noir pour le texte. Il va de soi après l'exposé des problèmes d'acquisition vus précédemment qu'un soin tout particulier doit accompagner la recherche d'algorithmes de binarisation.

Les méthodes de binarisation se divisent en deux classes : globales et locales.

Les méthodes globales calculent un seul seuil pour toute l'image. Les pixels ayant un niveau de gris plus foncé que le seuil sont mis à noir et les autres à blanc. Les méthodes présentent des façons différentes pour calculer le seuil. Par exemple, l'idée de la méthode de Niblack est de varier le seuil dans l'image en fonction des valeurs de la moyenne locale et de l'écart type local. La taille du voisinage doit être suffisamment petite pour préserver les détails locaux, mais suffisamment large pour supprimer le bruit. Dans la méthode d'Otsu, le problème est vu comme une analyse discriminante, pour laquelle on utilise une fonction critère particulière comme mesure de séparation statistique.

Les méthodes locales calculent un seuil pour chaque pixel en fonction de l'information contenue dans son voisinage. Par exemple, dans la méthode de Bernsen, le seuil est égal à la moyenne entre la plus petite et la plus grande valeur de niveau de gris de pixels situés dans un voisinage carré autour du pixel transformé.

Dans les deux cas, les méthodes de binarisation restent tributaires d'un ou de plusieurs seuils à déterminer. Dans le cas des documents anciens, en général très hétérogènes, ces seuils restent très difficiles à déterminer sans l'aide d'un expert extérieur.

3.3 Segmentation texte/graphique

Une fois l'image binarisée, et le texte séparé du fond, il faut procéder à l'extraction des médias pour des traitements appropriés. Contrairement aux techniques de pré-traitement précédentes, celles-ci se placent aux niveaux des entités et non au niveau des pixels. Il s'agit dans le cas des images de documents anciens de regrouper d'abord les formes en entités similaires, puis de procéder ensuite à leur classification en texte ou en graphique. Les éléments graphiques peuvent être suivant le document, des lettrines, des illustrations, mais aussi des paraphes, des ratures, des signes de renvoi, des grands traits, etc.

Étant régulier et ayant une texture de caractères très homogène, le texte offre une norme pour la classification. On utilise en général la largeur, la régularité et l'abondance des composantes connexes pour la classification. Ainsi, dans un texte, les composantes connexes sont peu larges, très régulières et très abondantes. Dans un graphique, les composantes connexes sont très larges, pas régulières et peuvent être abondantes.

Il existe deux approches générales de segmentation :

La première suppose que les blocs sont homogènes (1 seul média). Dans ce cas, chaque bloc est classé dans le média le plus proche en fonction des caractéristiques textuelles extraites de l'image du bloc.

Dans la seconde approche, on suppose qu'un bloc contient un mélange texte/non texte. C'est le cas des documents médiévaux[37] où des mélanges de graphiques et de texte peuvent être opérés. Dans ce cas, une analyse morphologique fine des composantes connexes, aidée de connaissances a priori sur la position des éléments peut aussi aider à leur élimination.

[37] Gusnard de Ventadert (nom collectif), « Les documents anciens », *Document Numérique*, Hermès, Vol. 3, n° 1-2, juin 1999, p. 57-73.

Image originale

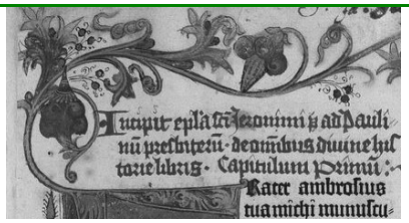
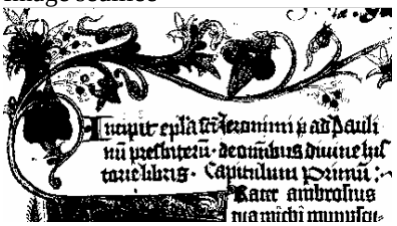


Image de niveaux de gris

Image seuillée

Otsu (globale)

Kitler(globale)



Kapur(globale)

ETM (locale)

Parker (locale)

Bernsen (locale)

Sauvola (globale)

Niblack (globale)

Image seuillée

3.4 Extraction de structures de lignes

L'extraction des lignes est indispensable dans les textes manuscrits qui en forment leurs seules structures. Cette extraction n'est pas facile car le texte est souvent incliné, peut comporter des ratures et des renvois. Elle est cependant indispensable pour rechercher des mots qui constituent les éléments fondateurs du texte.

Dans le projet Philectre[38], une interface a été faite permettant de réaliser une transcription diplomatique du texte, en associant à chaque ligne de transcription, l'image de la ligne dans le document.

Les lignes de texte dans les documents anciens présentent très peu de régularité exploitable. En effet, les lignes sont de différentes longueurs, contenant un enchevêtrement de composantes connexes. La littérature fait état de trois méthodes principales pour l'extraction de lignes dans les images binaires : les méthodes de projection ou groupement de composantes ou de pixels le long d'une direction, les approches multi-résolution[39] ou filtrage

[38] L. Robert, L.Likforman-Sulem, E. Lecolinet, « Image and Text Coupling for Creating Electronic Books from Manuscripts », Actes de ICDAR'97, Ulm, août 1997.

[39] C. Viard-Gaudin, D. Barba, « Extraction robuste et structuration des informations par une approche multi-résolution pour la localisation du bloc adresse sur des objets postaux plats », Actes de CNED'92, Nancy, Bigre, n° 80, p. 48-56.

différentiel[40], et les méthodes de groupement de points caractéristiques[41]. La Figure 1 montre l'application sur une lettre de rémission d'une méthode de type groupement perceptif par L. Likforman. Le groupement perceptif est un groupement de composantes connexes respectant des critères de proximité, de similitude et de direction.

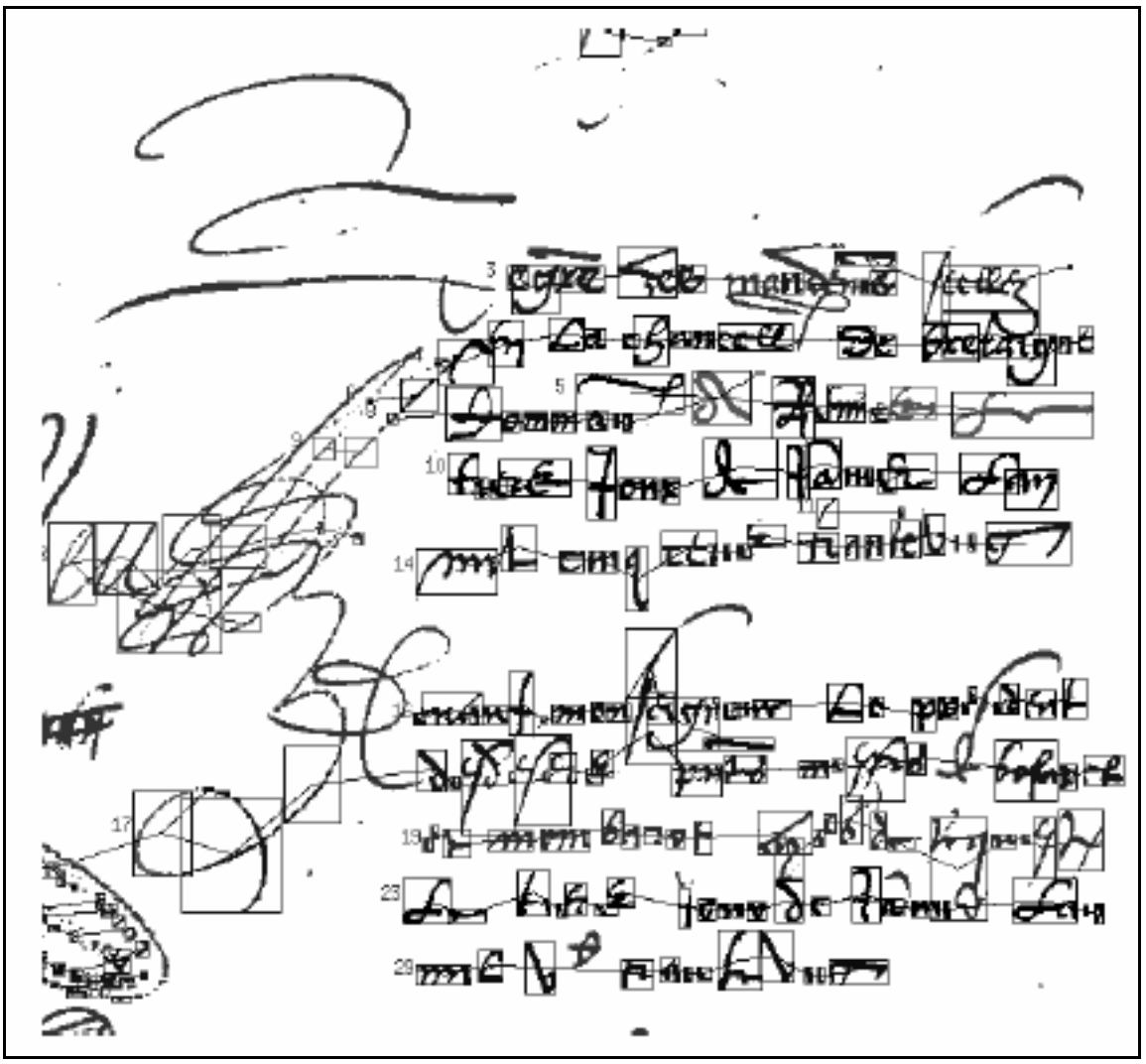


Figure 1 : extraction des lignes par groupement perceptif, d'après L. Likforman[42]

[40] F. Lebourgeois, « Localisation de textes dans une image à niveaux de gris », Actes de CNED'96, Colloque National sur l'Ecrit et le Document, Nantes, juillet 1996, p. 207-214.

[41] M. Feldbach « Generierung einer semantischen reprasentation aus abbildungen handschritlicher kirchenbuchaufzeichnungen », Diplomarbeit, Otto von Guericke Universitat Magdeburg, juillet 2000.

[42] L. Likforman-Suelem, « Apport du traitement des images à la numérisation des documents manuscripts anciens", Numéro special, *Document Numérique*, janvier 2004.

4. Reconnaissance de structures et extraction de méta-données

4.1 Pour les documents modernes, une rétro-conversion de la structure éditoriale

4.1.1 Notion de rétro-conversion

Le document moderne est en général composé suivant une feuille de style et respecte une DTD. Sa rétro-conversion automatique est l'opération inverse de l'édition qui consiste à retrouver ces structures de mise en page et éditoriale que le document aurait perdues lors de son impression sur papier. Quand la structure est évidente du fait de la présence d'indices typographiques flagrants et de la simplicité de l'enchaînement de ses composants, cette opération se ramène à une simple analyse structurelle à partir de grammaires de description, pour retrouver la DTD et la feuille de style. Mais le défi concerne les documents complexes, type formulaire, transaction et archive, où la structure apparente ne correspond pas forcément à une forme syntaxique régulière, et où cette structure s'est complexifiée par la présence de bruit, d'annotations, ou par des manipulations manuelles (papier déchiré, etc.). La difficulté se situe au niveau de la localisation des entités documentaires (tant physiques que logiques) ainsi qu'au niveau de leur séparation (les frontières sont souvent floues). Il faut alors avoir recours à des stratégies d'interprétation plus complexes qui dépassent la simple dérivation de règles syntaxiques. L'existence de bruit et l'incertitude de la connaissance sur ces types de documents conduisent à utiliser des approches hybrides, à combiner plusieurs sources de connaissance et à développer des stratégies relevant plus du raisonnement que de l'analyse syntaxique.

Ceci montre que la rétro-conversion totale ne peut pas se faire sans une modélisation a priori des connaissances. Les points durs de cette recherche concernent la définition et la stabilisation de tels modèles et surtout la définition de stratégies de reconnaissance robustes fondées sur ces modèles. Dans certains cas, on désire simplement retrouver certaines informations pertinentes; dans ce cas, l'utilisation de connaissances a priori n'est pas forcément nécessaire.

4.1.2 Application de la rétro-conversion

Un des problèmes majeurs de la rétro-conversion est qu'une page de document représente une structure *physique*, alors que le but de l'analyse et de la reconnaissance est d'aboutir à une description logique du document, la plus proche possible de celle utilisée par les outils d'édition. Il faut donc résoudre deux sous-problèmes :

- Trouver la structure physique, c'est-à-dire décomposer l'image de document en entités structurelles homogènes : caractères, blocs de texte, primitives graphiques, etc.
- Passer de la structure physique à la structure logique, c'est-à-dire retrouver en quelque sorte la «sémantique » du document. Ce problème peut être assimilé à l'opération inverse de celle effectuée par un logiciel de composition et mise en page.

Bien évidemment, ces problèmes ne se posent pas dans les mêmes termes selon les catégories de documents : composite, postale, formulaire et technique, et selon la granularité des structures : forte ou faible.

4.1.2.1 Cas des documents à granularité forte

C'est le cas des documents composites ou des formulaires où la notion de contenant et de contenu est relativement bien marquée.

La classe des *documents composites* est très diverse. Dans certains cas, le document obéit à des règles de mise en page assez précisément connues et standardisées (lettre commerciale, article scientifique, etc.); il est alors possible d'effectuer une analyse syntaxique pour retrouver la structure logique à partir de la structure physique. Dans d'autres cas, il est difficile de trouver beaucoup de règles fortes sur la structure logique (pages de journaux par exemple); on est alors réduit à effectuer une analyse plus rudimentaire, mais qui permet néanmoins de fournir un enchaînement logique probable des paragraphes, en appliquant les règles usuelles de mise en page et de lecture.

Dans Belaïd, la *reconnaissance des structures éditoriales* a été réalisée par l'emploi de systèmes à bases de connaissances. Les connaissances multiples (typographique, géométrique et logique), pondérées par des degrés de confiance sur la qualité des documents, sont organisées en un modèle *a priori* et interprétées comme des hypothèses de segmentation. Les stratégies d'analyse reviennent d'abord à extraire les composants typographiques (bloc, colonne, style typographique, positionnement topographique, etc.), puis à déduire l'information logique en s'appuyant sur les hypothèses d'association physique-logique du modèle. L'originalité de cette recherche se situe au niveau de l'inférence automatique de ce type de modèle à partir d'exemples[43], et également au niveau de l'étude de la convergence par étude d'entropie[44]. La Figure 2 montre quelques exemples de segmentation effectués sur différents types de documents.

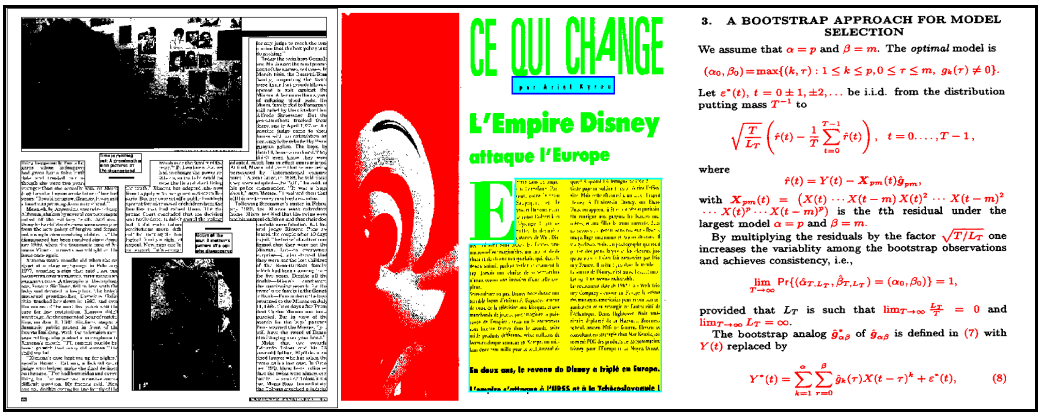


Figure 2: de la gauche vers la droite : Segmentation isothétique, extraction de médias, segmentation de formules mathématiques

Dans la classe des *structures tabulaires*, les méthodes étudiées sont basées sur quelques critères selon la complexité de l'organisation et de la structure de l'information. Ces critères dénotent la complexité progressive dans l'analyse du formulaire qui grandit avec la variabilité de la structure. Il semble évident que le cas le plus favorable est quand l'information est groupée dans les cellules, dans des régions identifiées, dans une structure stable[45],[46],[47]. À

[43] T. Akindele and A. Belaïd, « Un système d'aide à l'acquisition de modèles de documents », Colloque National sur l'Ecrit et le Document, Rouen, juillet 1994.

[44] Y. Chenevoy and A. Belaïd, "Hypothesis Management for Structured Document Recognition. International Conference on Document Analysis and Recognition", St-Malo, septembre 1991.

[45] H. Arai and K. Okada , "Form Processing Based on Background Region Analysis", ICDAR'97, Ulm, Germany, p. 164-169, 1997.

[46] S. Kebairi and B. Taconet, A.Zahour, S. Ramdane, "A Statistical Method For an Automatic Detection of Form types", *Proceedings of the DAS'98*, Nagano, Japan, November 4-6, 1998, p.109-118.

[47] S. W. Lam, L. Javanbakht and S. N. Srihari. "Anatomy of a Form Reader", IEEE ICDAR, p. 579-582, 1995.

l'opposé, plus d'investigations seront nécessaires pour l'information distribuée, non organisée en aucune structure et où la structure change toujours[48],[49],[50].

La littérature mentionne beaucoup de travaux sur les *formulaires administratifs* et les *tableaux*. Dans les travaux de Y. Belaïd sur les formulaires administratifs, la caractéristique filaire de ceux-ci a conduit à utiliser des méthodes d'extraction de traits (Transformée de Hough)[51] et de recherche de cellules par analyse de graphes d'intersections[52]. Enfin, les cellules sont classées par des techniques neuronales[53] prenant en entrée des mesures morphologiques de leur composantes connexes (voir Figure 3, photo de droite).

L'analyse de formulaires à structure non filaire reste un des problèmes les plus délicats de la recherche en analyse de documents. Un des exemples les plus caractéristiques correspond aux formulaires de type VPC (Vente Par Correspondance) traités à Nancy[54], où la mise en page change de manière permanente conduisant à utiliser des méthodes de rétro-conversion adaptatives. La Figure 3 (photo de gauche) en donne un exemple représentatif où les éléments à extraire correspondent aux régions des adresses, du corps de la commande, et aux régions des montants. L'absence de modèle *a priori* général a conduit à utiliser des modèles de points fixes représentatifs des régions mentionnées. Les points fixes correspondent à des mots ou des intitulés qui sont toujours présents quelque soit la mise en page. Aussi, des modèles de contraintes ont été définis pour les représenter. La recherche de ces points fixes se fait par relaxation discrète.

[48] F. Cesarini, M. Gori, S. Mariani and G. Soda, "INFORMys : A Flexible Invoice-like Form Reader System", *IEEE Trans.*, PAMI, 20(7):730-745, July 1998.

[49] J. Lii and S. Srihari, "Location of Name and Address Cover Pages", ICDAR'95, Montreal, Canada, 1995, p. 756-759.

[50] J. J. Yuan , Y. Y. Tang and C. Y. Suen, "Four Directional Adjacency Graphs (FDAG) and their Application in Locating Fields in Forms", *Proceedings of the IEEE*, ICDAR Montréal, Canada, 1995.

[51] Y. Belaïd, A. Belaïd and E. Turolla, "Item Searching in Forms: Application to French Tax Form", ICDAR'95, Montréal, Québec, August 14-16, 1995.

[52] E. Turolla, Y. Belaïd and A. Belaïd, "Form Item Extraction Based on Line Searching", *Lecture Notes in Computer Science*, International Workshop on Graphics Recognition, Pennstate, USA, 1995.

53 Y. Belaïd, J. L. Panchèvre and A. Belaïd, "Form Analysis by Neural Classification of Cells", International Workshop on Document Analysis Systems, Nagano, Japan, November 4-6, 1998.

[54] A. Belaïd, Y. Belaïd, N. Valverde, and S. Kébairi, "Adaptive Technology for Mail-Order Form Segmentation", International Conference on Document Analysis and Recognition, Seattle, USA, 2001, 5 p.

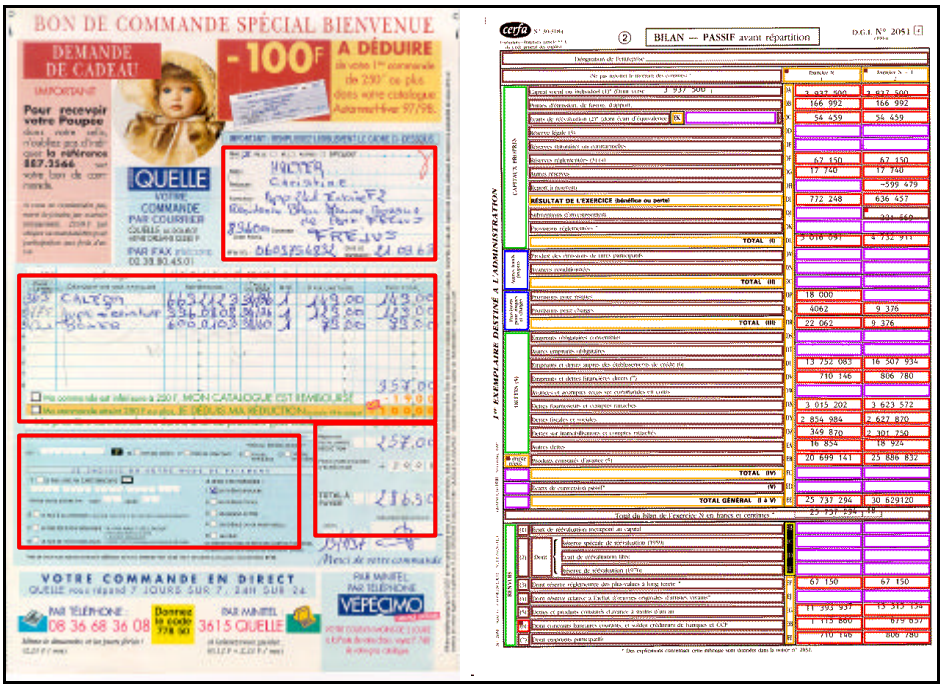


Figure 3: à gauche : extraction des régions d'information dans un formulaire de type VPC, à droite : localisation et classification des cellules par la nature du contenu.

4.1.2.2 Cas des documents à granularité fine

Les documents à structure fine correspondent aux textes à l'intérieur des blocs, qui possèdent une certaine structure interne, représentée par une succession de champs normalisés. C'est le cas des références et notices bibliographiques, des définitions de dictionnaires, des articles des tables de matières, des articles de corps de factures, etc. La rétro-conversion consiste dans ce cas à analyser le contenu et à extraire les champs.

Quand la structure est normalisée et stable, comme cela peut être le cas pour les citations bibliographiques, alors on peut avoir recours à des modèles *a priori* de type syntaxique ou stochastique. Mais souvent, la variation de la présentation oblige à rechercher des techniques ascendantes par analyse directe du contenu. Plusieurs méthodes d'analyse de documents ont été proposées sur la base d'étiquetage de l'information par extraction de mots clés. Un modèle linguistique est ensuite cherché à partir de ces mots clés pour mettre en évidence le contenu informationnel des champs textuels et pour rechercher des unités linguistiques dont la référence à la réalité est stable. Par exemple, l'hypothèse première du modèle SYDO[55] est que les parties du discours construites autour du nom (ou syntagmes nominaux) sont celles qui sont porteuses de référence aux objets de l'univers du discours et donc celles qu'il faut identifier. Le modèle linguistique proposé reflète le mécanisme permettant le passage de mots prédicats au syntagme nominal.

Il y a deux familles de méthodes d'étiquetage automatique de PdD : les méthodes à base de règles[5657] et les méthodes stochastiques[585960], fonctionnant toutes deux en mode supervisé et non supervisé.

[55] Lallich-Boidin Geneviève, Henneron Gérard & Palermiti Rosalba, «Analyse du français : achèvement et implantation de l'analyseur morpho-syntaxique », *Les cahiers du CRISS*, n°16, novembre 1990.
[56] Brill Eric, « A simple Rule-Bases Part of Speech Tagger », *Proceedings of the third Annual Conference on Applied Natural Language Processing*, ACL, 1992.

Les premières utilisent typiquement une information contextuelle pour affecter les tags à des mots inconnus ou ambigus. Ces règles sont souvent connues sous le nom de règles de frame contextuel. En plus de l'information contextuelle, plusieurs taggeurs utilisent l'information morphologique pour résoudre l'ambiguïté provoquée par des mots inconnus. Quelques systèmes vont au-delà de l'information contextuelle et morphologique en incluant des règles prenant en compte des facteurs comme la ponctuation ou l'emploi des majuscules. Les secondes incorporent la fréquence ou la probabilité dans le processus de validation. Les plus simples d'entre eux utilisent la probabilité qu'un mot se présente avec un tag particulier pour son identification, mais ces approches ont un comportement incertain vu leur vision locale qui peut conduire à des séquences inadmissibles de tags. Une alternative à ces approches est de calculer la probabilité d'une séquence donnée de tags occurrents. Cette approche se base la méthode des n-grammes considérant que le meilleur tag pour un mot donné est déterminé par la probabilité qu'il se présente avec les n tags précédents[61]. La méthode généralement retenue dans un taggeur stochastique combine les deux approches précédentes, utilisant les probabilités de séquences de tags et les mesures de fréquences de mots. Ceci est connu sous le nom de HMM[62][63].

La Figure 4 donne deux exemples de rétro-conversion de documents textuels, issus des travaux de Belaïd, en utilisant une méthode d'étiquetage par partie de discours.

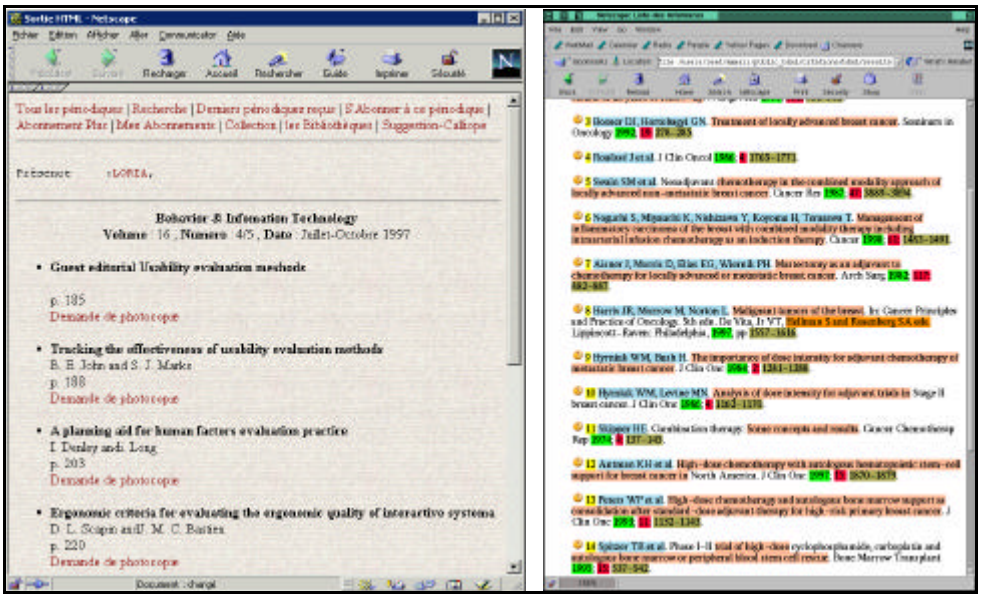


Figure 4: à gauche : page XML du résultat de la reconnaissance d'une table de matières, à droite : résultat d'étiquetage de champs de références bibliographiques

[57] Tapanainen, Pasi and Voutilainen, Atro, « Tagging accurately: don't guess if you don't know », Technical Report, Xerox Corporation, 1994.

[58] DeRose, Stephen J., « Grammatical category disambiguation by statistical optimization », *Computational Linguistics*, 14.1, 1988, 31-39.

[59] Marshall, Ian, « Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB corpus », *Computers and the Humanities*, 17, 1983, 139-150.

[60] Merialdo, Bernard, «Tagging English text with a probabilistic model », *Computational Linguistics*, 20.2, 1994, 155-172.

[61] Kupiec, J., « Robust Part-of-speech tagging using a hidden Markov model », *Computer Speech and Language*, 6, 1992.

[62] Merialdo, Bernard, «Tagging English text with a probabilistic model », *Computational Linguistics*, 20.2, 1994, 155-172.

[63] Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L. and Palmucci, J., « Coping with ambiguity and unknown words through probabilistic methods », *Computational Linguistics*, 19, 1993.

4.2 Pour les documents anciens, une extraction de méta-données

Comme cela a été précédemment indiqué, les documents anciens n'obéissent pas tous aux mêmes règles de structuration que les documents modernes car ils sont écrits différemment. Certains sont manuscrits, comme c'est souvent le cas pour les documents d'archives, tandis que d'autres sont calligraphiés. Une troisième catégorie imprimée offre une typographie reconnaissable par OCR avec un entraînement spécial sur certains caractères.
Nous allons dans la suite faire état des travaux de recherche effectués sur les documents anciens.

4.2.1 L'indexation et l'annotation d'images

Quand le contenu est difficile à rétro-convertir, ce qui arrive souvent quand le document est ancien, on indexe alors son contenu pour permettre de le consulter. L'opération d'indexation est une opération très délicate car il s'agit de passer de l'interprétation d'une représentation visuelle à sa description textuelle. Cette opération a été beaucoup utilisée sur les archives anciennes pour les ouvrir au public le plus large par le biais d'Internet, le plus souvent. Il s'agit en général d'une description très condensée par une liste de mots clés choisis dans un dictionnaire approprié. Les mots clés font référence à des méta-données structurelles.
L'expérience menée dans le cadre du projet LIB.R.I.S. de réalisation d'une base de données iconographique par le CRHENO[64] est particulièrement intéressante. Elle repose sur un mode d'observation interactif qui consiste à faire participer les lecteurs dans la constitution de la base de données et surtout des index. On demande à chaque lecteur de regrouper les images semblables ensemble, puis de donner une légende par image signalant les éléments les plus marquants et les plus significatifs, et enfin de décrire le contenu de chaque image par quelques phrases. Ce style de participation a fait dégager un style narratif très riche par rapport aux modes d'indexation classiques.

Considérant par ailleurs la base comme un ensemble cohérent et structuré, l'auteur[65] a exploité la programmation objet pour proposer un mode de description basé sur le principe de l'héritage de propriétés et de méthodes au sein d'une classe d'objets. Les relations établies entre classes d'images permettent ensuite de retrouver les images en fonction de leurs caractéristiques propres mais également grâce à leur appartenance à différents « scénotypes ».
Concernant les images de texte manuscrit, il est souvent tentant de vouloir produire automatiquement des indexations ou plutôt des annotations. Cependant, certains documents manuscrits d'archives sont particulièrement difficiles à reconnaître et il faut avoir recours à des méthodes de localisation ou d'approche. Ceci n'est possible que si le document est très structuré, contenant des indices permettant d'approcher l'information manuscrite. B. Couasnon[66] propose une méthode d'annotation permettant d'accéder aux documents manuscrits à partir de leur contenu. Il présente deux manières complémentaires de production des annotations : une automatique en utilisant la reconnaissance de documents, et une manuelle en utilisant Internet et une saisie au clavier effectuée par les lecteurs eux-mêmes. Cette deuxième méthode vient au secours de la première quand le document n'est pas très structuré, ou tout simplement pour compléter la première. Une plate-forme de gestion des annotations a été réalisée permettant l'accès au contenu manuscrit d'archives. Cette plate-forme a été expérimentée sur des registres paroissiaux et d'état civil, des formulaires militaires (registres matricules) et des décrets de naturalisation.

---

[64] Centre de Recherche sur l'Histoire de l'Europe du Nord-Ouest de l'Université de Lille 3.
[65] M. Després-Lonnet, « Pour une Indexation Orientée Objets », Numéro spécial, *Document Numérique*, Janvier 2004.
[66] B. Couasnon et J. camillerap, « Accès par le contenu aux documents manuscrits d'archives », Numéro spécial, *Document Numérique*, Janvier 2004.

La méthode automatique s'intitule DMOS[67]. C'est une méthode de localisation générique des emplacements des annotations. Elle est constituée d'un langage de description des emplacements géométriques, d'un analyseur syntaxique dédié, et d'un analyseur lexical basé sur les filtres de Kalman pour l'extraction des termes.

Les autres travaux sur les annotations que la littérature mentionne sont principalement autour d'XML et de RDF[68].

Photo-RDF[69] est un projet pour décrire et retrouver des images numériques grâce à des méta-données en RDF. Des schémas RDF ont été définis ou utilisés pour associer différentes informations aux photos : titre, date, appareil photo, focale, etc. Un des problèmes signalés dans Photo-RDF est qu'il n'est pas possible d'y associer une position précise dans l'image, mais seulement l'image toute entière.

Hunter et Zhan[70] proposent d'inclure des méta-données, définies également en RDF, dans des fichier PNG. Dans ce schéma, il est possible de définir une région dans l'image à l'aide d'un identificateur, un titre, un peu de texte et ses coordonnées. Même si ceci offre la possibilité d'associer une annotation à une position précise dans l'image, cette position n'est qu'un attribut d'une annotation textuelle. Or dans le cadre de l'accès aux documents, il est nécessaire de pouvoir considérer une position dans l'image comme une annotation au même titre qu'une annotation textuelle.

Phelps *et al*[71] offrent un cadre d'annotations multivalent de documents sous des formats très variés : images numérisées, HTML, DVI, etc. Cependant, une position dans l'image n'est toujours pas considérée comme une annotation à part entière, ce qui ne permet pas d'associer par exemple plusieurs annotations textuelles à une même zone de l'image.

4.2.2 Recherche d'information

Dans cette partie, il est question de navigation et d'accès à des collections inaccessibles sous forme textuelle (cas des documents manuscrits anciens, par exemple). La démarche est différente de celle vue précédemment, car il s'agit d'offrir à un lecteur spécialiste un moyen de navigation particulier. Les besoins d'utilisation peuvent être différents. Certains utilisateurs cherchent par exemple à retrouver les documents de la base présentant certaines calligraphies correspondant à certains scripteurs. D'autres cas d'utilisation peuvent concerner la détection des différentes mains présentes, ou bien la datation des documents par rapport à la chronologie de l'œuvre de l'auteur.

On peut considérer que ces deux cas d'utilisation relèvent d'un problème de recherche d'information soit textuelle soit graphique. Ces deux tâches ont été largement étudiées soit dans le domaine documentaire soit en traitement d'images. En ce qui concerne spécifiquement l'analyse des écritures manuscrites, cette tâche relève de l'identification du scripteur d'un document.

---

[67] Description et Modification de la Segmentation.

[68] Resource Description Framework (RDF), « Model and syntax specification », W3C Recommandation, février 1999. <http://www.w3.org/TR/REC-rdfsyntax/>.

[69] Photo-RDF, « Describing, retrieving photos using RDF, and http », W3C Note, April 2002. <http://www.w3.org/TR/photo-rdf/>.

[70] Hunter J., Zhan Z., « An indexing and querying system for online images based on the png format and embedded metadata », *Proc. of the ARLIS/ANZ Conference*, Brisbane, Autralia, septembre 1999.

[71] Phelps T.A., Wilensky R., « Multivalent annotations », *Proc. of the First European Conference on Research and Advanced Technology for Digital Libraries*, Pisa, Italy, 1997.

L'équipe de Rouen[72] a particulièrement investi dans la recherche sur la description des écritures et l'identification des scripteurs, d'abord sur des documents synthétiques, puis sur des documents réels correspondants aux manuscrits de correspondances de Zola. Elle a défini une chaîne complète d'identification des scripteurs permettant d'identifier les caractéristiques importantes à prendre en compte dans les mots. Ces caractéristiques ont permis d'identifier des espaces de recherche de scripteurs.

La littérature mentionne plusieurs autres types de modèles de Recherche d'Information qui peuvent être appliqués sur les images de texte, d'après Song et Bruce Croft[73] : le modèle booléen, le modèle probabiliste et le modèle vectoriel (VSM) sont les plus connus. Ce dernier, proposé par Salton[74], est un des modèles de recherche d'information les plus utilisés. Les documents de la base ainsi que la requête sont représentés par un vecteur dans un espace de grande dimension. Bien que très simple et de conception assez ancienne, ce modèle reste très efficace[7576]. Dans ce modèle, la stratégie de recherche s'effectue en deux phases : une phase d'indexation permettant de décrire chaque document par un vecteur de grande dimension; une phase de recherche où sera évaluée la pertinence de chaque document de la base D_j par rapport à une requête spécifique Q. Cette évaluation n'est rien d'autre qu'un produit scalaire entre le vecteur décrivant la requête Q et celui décrivant un document de la base D_j.

4.2.3 Reformulation de documents et lecture savante

La lecture savante consiste, à partir de documents diversifiées, d'essayer de répondre à la diversité des attentes et des besoins exprimés sur plusieurs plans par des utilisateurs multiples. Il s'agit alors de lever l'obstacle à l'accès aisé à la forme et au contenu du document et de permettre une interaction spontanée avec lui. C'est un problème de ré-édition qu'il faut étudier par rapport à la spécificité originale du document et en en conservant toutes les qualités morphologiques, syntaxiques, sémantiques et visuelles d'origine.

En théorisation comme en pratique effective, on voudrait aller plus loin que la simple transformation du document d'origine et permettre une lecture savante ou intelligente de celui-ci en conformité avec le support sélectionné, même si celui-ci n'est pas spécifiquement dédié à la lecture.

La lecture savante consiste à équilibrer les contraintes posées par les besoins de l'utilisateur, par le support et par le contrat de lecture original pour produire un nouveau document, conforme à l'attente de l'usager. C'est un problème de reformulation électronique.

---

[72] A. Bensefia, Th. Paquet et L. Heutte, « Documents Manuscrits et Recherche d'Information », Numéro spécial, *Document Numérique*, janvier 2004.

[73] Song F., Bruce Croft W., « A General Language Model for Information Retrieval », *Eighth International Conference on Information and Knowledge Management* (ICIKM'99), 1999.

[74] Salton, Wrong, «A vector Space Model for Automatic Indexing », *Information Retrieval and Language Processing*, p. 613-620, 1975.

[75] Memmi D., « Le modèle vectoriel pour le traitement de documents », *Les Cahiers du Laboratoire Leibniz*, IMAG-Grenoble, France, n°14, 2000.

[76] Pouliquen B., Delamane D., Lebeux P., « Indexation des textes médicaux par extraction de concepts et ses utilisations », 6e Journée internationale d'Analyse statistique des Donnés Textuelles (JADT'02), 2002.

4.2.3.1 Contexte

Ces dernières années, les supports électroniques permettant de restituer les documents se sont multipliés. En effet, en plus de l'ordinateur classique, on trouve maintenant les butineurs, les assistants personnels, les téléphones portables, les livres et cartables électroniques. Pour les livres électroniques (format OeB - Open eBook) la situation est radicale : ils sont conçus comme les maillons terminaux d'une chaîne d'édition et à ce titre ils sont fermés d'un point de vue format et usage. Il n'est pour l'instant pas question d'une fonction de transformation permettant à tout un chacun de transférer un document sur un livre électronique.

Les utilisateurs sont de plus en plus confrontés à des masses d'information qu'ils reçoivent sous des formes très diverses, sur des supports ayant des limites intrinsèques d'affichage et de manipulation. La recherche s'est donc portée naturellement vers la restructuration de l'information de manière à faciliter la réponse aux besoins de l'utilisateur.

Au niveau de la recherche et de la standardisation internationale, plusieurs types de formats existent pour représenter des documents structurés en fonction de la finalité recherchée. Initialement, la recherche s'est intéressée à la modélisation des structures physiques et logiques en explicitant leur dualité. La famille des standards SGML[77] a servi de cadre pour concrétiser cette modélisation. Plus récemment XML a permis d'étendre la structuration aux protocoles d'échange de documents de différents types. Concernant la dimension sémantique plusieurs projets de recherche ont été développés dans le domaine de l'hypertexte. Plusieurs modèles ont été proposés pour réaliser cet objectif qui s'appuient principalement sur la définition et l'utilisation d'hyperliens puissants. Ils permettent également la modélisation de méta-données associées au document et la spécification d'interactions avec l'utilisateur.

La recherche sur l'hypertexte a montré les limites des modèles liens nodaux simples dans la représentation de la connaissance. Des chercheurs ont proposé diverses structures pour améliorer la capacités des hypertextes à exprimer les relations conceptuelles (par exemple le système MacWeb de Nanard[78]). La proposition RDF (Ressource Description Framework) du W3C constitue une tentative pour adjoindre une telle structure conceptuelle à la couche supérieure du Web.

La recherche concernant les méthodes de transformation est féconde quant à elle : il existe de nombreuses méthodes conventionnelles de transformation de formats (cf. supra). En revanche, la structuration conceptuelle vers laquelle on tend pour un usage personnalisé n'a pas fait l'objet d'une recherche scientifique approfondie.

4.2.3.2 Objectif scientifique

De manière générale, ces recherches ont ouvert la voie sur la prise en compte de la sémantique et les conversions de formats, mais elles font complètement abstraction des contraintes du support et celles liées aux conditions spécifiques d'utilisation du document. C'est sur ce point que l'équipe READ à Nancy oriente sa recherche en collaboration avec des partenaires industriels. Cette recherche agira plus spécifiquement sur les points suivantes :

Les contraintes : on peut isoler trois types de contraintes fortes possibles sur un document:

[77] Standard Generalized Markup Language.

[78] J. Nanard and M. Nanard, "Should anchors be typed to ? an experiment with MacWeb", *Proc. Of the Hypertext Conference*, Seattle, 1993.

- La première contrainte concerne la dimension du réceptacle. On ne peut pas faire abstraction de cette contrainte car elle a une forte influence sur la bonne lecture des données et le bon affichage des informations pour l'utilisateur.

- La deuxième contrainte concerne le contrat de lecture passé entre l'éditeur du document et son utilisateur. Cette contrainte est un peu plus souple et la rétro-conversion classique des données peut permettre de répondre dans une certaine mesure à l'ensemble de ces deux contraintes.

- La troisième contrainte concerne les désirs de l'utilisateur. Étant purement informelle, cette contrainte est souvent évacuée au bénéfice des autres. Lorsqu'elle est prise en compte, cela indique souvent que les deux autres contraintes ne se posent pas dans l'environnement présent (ce qui est le cas du Web avec ses hyperliens qui suppose que l'utilisateur se plie à ses règles de présentation, et les contraintes de dimension sont souvent imposées; on peut alors prendre en compte plus simplement les besoins de l'utilisateur).

L'objectif est donc de prendre en compte ces trois contraintes et d'en tenir compte de manière pondérée pour reformuler des documents de la manière la plus proche possible des désirs de l'usager.

Le lectorat et les usagers : Il faut distinguer en matière de lecture de document électronique entre l'***usager*** à la recherche d'une information ponctuelle et le ***lecteur*** désireux de saisir l'ensemble de l'information du document. La diversité des attentes et des compétences de ces lecteurs amène à formuler la notion de ***lectorat*** comme ensemble des particularités et des besoins exprimés par chaque lecteur individuel. Il s'agit alors de définir l'extension et les modalités constitutives de cet ensemble (marques socio-culturelles, historiques, intentionalité pratique, etc.).

Cette distinction emporte avec elle une différence dans le traitement informatique des documents. Dans le cas du simple ***usager***, l'indexation ponctuelle paraît suffisante. Dans le cas du ***lectorat***, se pose le problème de l'accumulation d'informations et de l'accès intelligent au sens global du document.

Ce problème peut se traduire en termes d'interprétation de requêtes lancées par l'utilisateur au système de ré-édition. Dans ce cas, l'***usager*** formule des requêtes de type simple (recherche d'un mot, d'une illustration, d'un titre, etc.), tandis que le ***lecteur*** adresse plus des requêtes complexes destinées à lui faciliter la compréhension globale du texte par delà les enchaînements d'unités informatives. Ces requêtes peuvent prendre plusieurs formes, soit une combinaison logique de requêtes simples, soit une formulation adaptée au résultat recherché, étant donné que le sens du document constitue un tout qui excède la somme de ses parties constituantes, ce qui pose le problème du passage du discontinu des requêtes au continu sémantique. On parlera alors de ***lecture augmentée***.

Lecture augmentée, reliure personnalisée et re-lecture : La *lecture augmentée* cherche tout d'abord à apporter au lecteur un suivi du document comme il le désire. Selon la nature du désir, cela peut aller d'une restructuration physique propre à restituer la lisibilité originelle du document, à une restructuration ontologique pouvant modifier complètement la structure du document, en passant par une reformulation structurelle du document permettant tout simplement une meilleure prise en main. Les différents supports permettent de plus une interaction plus ou moins complexe avec le document (par ex. bouton, stylo, etc.), que l'on appelle *reliure personnalisée*.

Une fois le texte compris, la *re-lecture* permet au lecteur d'agir sur le document en y insérant commentaires, notes, balises définissant un parcours de lecture, etc. Pour atteindre cet objectif, il est nécessaire d'intégrer à la fois les interfaces graphiques adaptées à cet usage et les techniques de reconnaissance du manuscrit concomitantes. La recherche portera à la fois

sur les aspects d'intégration, mais également sur ceux de l'adaptation au support. Il existe un acquis relativement important en recherche qu'il s'agira de reprendre ici.

L'étiquetage sémantique : Il s'agit de repérer les éléments linguistiques qui structurent sémantiquement le document, d'en dégager la hiérarchie (noyaux et catalyses) et de les lier ontologiquement. L'idée est de se baser sur ces indexes pour sélectionner les composants du document les plus proches des parties du discours que ces composants désignent (langage et métalangage) et qui correspondent aux attentes du lecteur.

Les ontologies formalisent les concepts et les organisent dans des structures hiérarchisées. C'est tout naturellement que les logiques de descriptions ont été proposées pour décrire les ontologies, car plus qu'un simple formalisme, elles possèdent des mécanismes qui garantissent la cohérence des définitions conceptuelles et de leur organisation hiérarchique comme il a été dit plus haut.

4.2.3.3 Objectif applicatif

Le livre objet étant jusqu'à aujourd'hui l'espèce matérielle permettant l'interactivité de lecture la plus aboutie, le défi que l'on vise à relever consiste à permettre un égal accès à la forme et au contenu des documents à travers toute la déclinaison des modes électroniques sous lesquels l'objet peut se présenter: livre électronique bien sûr mais au-delà l'écran d'ordinateur, les PDA, etc.

Il va de soi que les problèmes numériques de compression, de visualisation et d'affichage ressortent immédiatement. C'est un problème de ré-édition électronique que l'on pense pouvoir résoudre en se ralliant à un schéma de transformation de formats pour lequel il existe des standards (W3C).

4.2.3.4 Un exemple pratique, la reformulation de dictionnaires anciens

C'est un projet entre l'ATILF et le LORIA[79] permettant de rendre accessible à un public diversifié de chercheurs, grâce aux nouvelles technologies de numérisation et d'indexation, le contenu de documents anciens imprimés, du XVIe au XIXe siècle. Forts des expérimentations précédentes de numérisation en mode image, les chercheurs de l'ATILF visent à approfondir l'usage de l'indexation par la structure formelle représentative de certains contenus pour en optimiser l'exploration.

Les dictionnaires constituent par définition la matière la plus propre à une indexation fine. Contrairement aux dictionnaires contemporains qui offrent d'emblée une indexation systématique (théorisée, i.e. fondée sur une spécification rigoureuse des entrées), les dictionnaires anciens oblitèrent souvent le système de traitement des items lexicaux (définition et repérage des entrées, des sous-entrées, des entrées cachées, des renvois, etc.). En effet, la fonction annoncée des choix typographiques (polysémie des grandes capitales, petites capitales, italiques) n'est pas toujours rigoureusement respectée, ce qui crée des ambiguïtés sur le statut des entrées et rend difficile la consultation manuelle et automatique, d'où l'urgence d'une réflexion sur une indexation efficace.

L'index minimal consiste à produire la liste des entrées marquées d'un dictionnaire ; la réalité des textes étudiés conduit à définir un principe d'indexation qui tienne compte de l'ensemble des mots traités sans être marqués comme des entrées évidentes.

La reformulation porte essentiellement sur la structure des articles et le contenu des champs informationnels. Pour cela chaque champ doit avoir été défini par son image. La reformulation portera soit sur le changement d'ordre des champs, soit sur le contenu du

[79] A. Belaïd, I. Turcan, J.M. Pierrel, Y. Belaïd, Y. Rangoni and H. Hadjamar, «Automatic indexing and reformulation of ancient dictionaries », Workshop DIAL, January 24-26, AT&T Labs, San José, USA, 2004.

champ lui-même susceptible d'être affiché (repris tel que), ou associé à d'autres éléments, voir complété par d'autres champs (collés ou liés, etc.)

Ainsi, ce travail consiste à proposer à tout lecteur, averti ou non, une nouvelle forme de structuration des champs des articles. Par exemple :

Reformulation de la nomenclature quel que soit son mode de repérage (liste exhaustive des entrées). L'étiquetage inter-structurel permet de neutraliser la difficulté des variantes graphiques sur un corpus d'ouvrages anciens ;

Reformulation de la terminologie grammaticale ;

Reformulation des domaines de spécialités représentées (vocabulaire des artisans , marine…).

Georges VIGNAUX
(CNRS – Laboratoire Communication et Politique, Paris)

La notion de « collection » : genèses, développements, valorisations

0. Avant-propos

Ce texte doit être considéré comme une contribution aux travaux de l'Action spécifique « Numérisation et valorisation des collections », lesquels travaux ont permis de capitaliser durant la durée de l'Action, nombre de réflexions fécondes sur cette thématique somme toute considérable. La diversité des apports et des expériences des participants nous a conduit a opter pour la formule de contributions spécifiques. Tel est le cas de ce document consacré à une certaine approche de la notion de collection.

1. La notion de collection

1.1 Parcours historique conceptuel

A suivre la définition de Krzysztof Pomian[1], le terme « collection » désignerait « tout ensemble d'objets naturels ou artificiels maintenus temporairement hors du circuit d'activité économiques, soumis à une protection spéciale dans un lieu clos aménagé à cet effet, et exposés au regard ». On pourrait corriger cette définition d'historien en précisant qu'une collection n'est pas nécessairement un regroupement d'objets, mais plutôt un ensemble d'entités, d'éléments tels que l'acte de les regrouper implique un certain regard sur eux ou une certaine représentation construite sur eux. Ces entités sont alors considérées comme dotées de d'attributs qui les relient entre elles. Autrement dit, «ce qui fait collection », c'est aussi l'œil de l'observateur.

Jusqu'au XVIII^{e} siècle, la collection se révèle comme la manifestation d'une activité à laquelle contribuent selon les époques, la foi, la sociabilité ou la soif de connaissances. Dans son *Histoire naturelle,* Pline fait remarquer qu'une seule gemme précieuse ne peut suffire à

[1] Pomian, K., *Collectionneurs, amateurs et curieux, Paris, Venise : XVIe -XVIIIe siècles,* Paris, Gallimard, 1987.

exprimer la totalité des choses dans la nature. Il organise alors les gemmes selon l'ordre descendant de leur renommée, c'est-à-dire selon leur importance dans la tradition à travers les siècles. C'est l'idée de compilation. Cette idée se retrouve chez le Père Richelet, qui dans son *Dictionnaire françois* (1680), renvoie à l'article « Recueil » ainsi défini : « Extrait de ce qu'il y a de beau dans plusieurs auteurs. Ramas de différentes pièces. Assemblage de diverses choses qui concourent toutes à une fin. »

A cette époque, chaque collection est censée fournir à l'observateur des formes abrégées pour l'aider à se représenter une nature en soi, au-delà de l'apparence des choses. Hérodote mentionne ainsi une collection de statues d'anciens prêtres égyptiens qui a pour vocation d'aider à visualiser les générations qui ont dirigé le pays et donc son histoire.

La même fonction didactique de la collection se retrouve chez les académiciens *del Cimento* à Florence, au XVII^{e} siècle. «Les objets les conduisent "géométriquement" (par étapes graduelles) à la connaissance des choses naturelles » au moyen d'une confrontation répétée entre les assertions des philosophes, naturalistes et honnêtes gens et ce qui se passe devant leurs yeux lors de la reconstitution appelée "expérience naturelle" des circonstances décrites auparavant à propos de quelque changement que subit un objet dans une circonstance donnée » (*Dictionnaire des notions philosophiques*)[2].

C'est au début du XVIII^{e} siècle que se clarifie le débat sur la notion de collection, jusque-là traitée avec ironie. Rappelons le portrait du curieux brossé par La Bruyère dans *Les Caractères* (1688) où le collectionneur de tulipes devient objet de ridicule puisque sa passion l'éloigne du monde.

C'est Joseph Pitton de Tournefort, qui le premier, dégage dans ses *Elémens de botanique* (1694) une méthode pour l'analyse systématique des plantes, fondée sur les «caractères » de la plante, le fruit et la fleur et les parties anatomiques qui la composent. Tournefort fonde les critères sur la structure même de la plante et par là, éloigne toute autre considération douteuse. C'est le premier regard atemporel et dépersonnalisé. De là viendront les grandes projets des sciences de la nature au XIX^{e}, les grands schémas analytiques et classificateurs des Linné, Lamarck et Darwin.

1.2 Les définitions du dictionnaire (Trésor de la Langue française informatisé)

<http://atilf.inalf.fr/tlfv3.htm>

- Vx. [En parlant d'inanimés] Action de réunir, recueillir, rassembler.
- [Ce qui est réuni est considéré comme un tout] Les collections d'eau stagnante (E. BRUMPT, *Précis de parasitologie*, 1910, p. 73).

[En parlant d'animés ou d'inanimés] Etat résultant dune réunion d'éléments

- [Ce qui est réuni est constitué d'éléments juxtaposés, conservant leur individualité] Ensemble d'éléments groupés en raison de certains points communs.

[La collection résulte de choix successifs faits systématiquement par un individu ou un groupe d'individus dans une intention particulière]

Ensemble non fini (le plus souvent classé) d'objets réunis par un amateur, en raison de leur valeur scientifique, artistique, esthétique, documentaire, affective ou vénale. Collection d'insectes, de tableaux, de timbres; collection particulière, publique; pièce de collection; vente de la collection (de) X; faire (la) collection de; dépareiller une collection:

Série de volumes contenant les œuvres d'un auteur ou de volumes publiés sous un titre commun, et édités le plus souvent de façon uniforme.

[2] *Dictionnaire des notions philosophiques*, sous la direction de Sylvain Auroux, Paris, PUF, 1990.

Ensemble complet de fascicules d'un périodique parus pendant une certaine période.

Ensemble des modèles (vêtements ou accessoires) créés par un couturier pour une saison. Collection de prêt-à-porter; présentation de (la) collection; la, les collection(s) de printemps-été, d'hiver:

Fam. Toute une collection de. Un grand nombre, une grande quantité de choses ou de personnes mises ensemble sans beaucoup de discernement :

1.3 Art et collection

« D'une façon très générale, les œuvres-collections […] sont des œuvres dans lesquelles sont mis en scène divers aspects propres à ce mode particulier d'intelligibilité du monde qu'est l'acte de collection en lui-même »[3].

On a pu retrouver « des formes subtiles de collection utilisées comme moyen d'expression artistique. Un grand nombre d'artistes utilisèrent la collection comme activité directe ou ils s'en inspirèrent en réunissant des archives ou des inventaires, ou alors en créant des reconstructions et des musées. Cette façon d'élever la collection au niveau de forme d'art découle d'une longue tradition liant les collections et la production artistique »[4].

C'est «à partir de l'acceptation du fragmentaire et d'une dissémination du réel, que le principe général de collection peut ainsi être mis en œuvre, dans un contexte artistique qui semble rompre pourtant avec les principes fondamentaux ont relève l'acte de collection »[5].

Mentionnons pour conclure provisoirement, cette citation de Claudine Robert, fille du défunt mathématicien Laurent Schwartz (médaille Fields), qui fut aussi – ce qu'on sait moins – un très grand collectionneur de papillons : «On commence avec très peu de choses, puis cela s'agence, cela s'organise. On se récite les monographies. Une collection à bâtir, c'est une langue qui se construit. »[6]

2. La notion de collection : genèses et valorisations

La notion de collection repose sur trois grandes « idées » :

• celle de réunir, de mettre ensemble des objets,

• celle d'éléments juxtaposés ainsi réunis, mais qui conservent leur individualité,

• celle d'un principe de réunion au nom d'éléments communs ou encore fondé sur des caractéristiques communes.

Plusieurs types d'opérations et de processus sont ainsi mobilisés à l'occasion de l'émergence d'une collection :

• le choix de quelques objets initiaux (livres, tableaux, bibelots, archives),

• la mise en évidence de propriétés communes ou proches (aux plans des formes ou des fonctions) entre ces objets,

• la confirmation d'une certaine permanence de ces propriétés : la collection doit se prévaloir d'une stabilité mesurable dans l'espace et dans le temps (physique, fonctionnelle, objectale : collectionner les compteurs à gaz peut se comprendre…),

• la légitimation de cette mise en commun au nom de valeurs qui peuvent être, selon les circonstances, symboliques (les fèves des galettes des rois), matérielles (les flacons de

[3] Dubois, Christine, « L'œuvre-collection : de la taxinomie du visible à l'utopie », *Parachute*, 1989, n° 54.

[4] Grasskamp, Walter, « Les artistes et les autres collectionneurs », Museum by Artists, Toronto, Art Metropole, 1983.

[5] Dubois, Christine, « L'œuvre-collection : de la taxinomie du visible à l'utopie », *Parachute*, 1989, n° 54.

[6] « Les papillons de Laurent Schwartz», *La Recherche*, mars 2003, 72-78.

parfum), monétaires (les autos de luxe), patrimoniales (les objets de Picasso), historiques (les ivoires de Dieppe) voire sentimentales (mes « peluches »).

En conséquence, une collection traduira la mise en œuvre d'une ou plusieurs valeurs :

• *cognitives :* on l'a conçue et construite ; elle est donc porteuse d'idées, de représentations, de connaissances ;

• *pratiques :* elle est mesurable, on peut la référencer à un domaine, elle a des limites, des clôtures ;

• *fonctionnelles :* on peut la traiter, l'analyser, la décomposer, la recomposer, la vendre comme tout ou partie ;

• *symboliques :* historiques, originales (objets insolites), emblématiques (les objets retrouvés du Titanic).

De ces processus dépend la valorisation des collections. Cette valorisation peut suivre plusieurs types de stratégies concrètes et symboliques, centrées sur les objets-exemplaires de la collection ou centrées sur les extensions de ces objets :

• *référentielles :* définition et illustration d'un domaine sous forme d'ancrage dans ce domaine soit en extension (les timbres poste français de 1960 à nos jours) soit en rareté (les timbres des îles Comores de 1970 à 1980) ;

• *testimoniales :* la collection témoigne d'une époque ou d'un art (les épées du XV^{e} siècle espagnol ; les affiches art déco) ;

• *promotionnelles :* il s'agit alors de la mise en valeur de créations (la mode) ou d'époques ou de lieux (les châteaux de la Loire) ;

• *fiduciaires :* fondées sur la rareté (telle ou telle collection de bijoux ou de voitures anciennes).

Nous sommes ici proches de la notion de document, telle que je la déclinais lors de ma contribution aux réflexions du groupe « Roger T. Pédauque » :

La notion de document peut se décliner selon différentes facettes ou figures :

Selon le statut physique ou matériel (papier, image, photo, vidéo, film),

Selon le statut social ou communicatif : il s'agit de la place de ce document dans un domaine, un contexte, une société,

Selon le statut cognitif et informatif : qu'est-ce que le document apporte comme connaissance, comme information et sur quoi ?

On peut dégager trois grands cas de figures :

Le document est une pièce unique, isolée : il est une preuve d'existence : c'est la carte d'identité, qui témoigne de l'inscription du porteur dans une société, dans un pays. Sa définition, c'est sa *dénomination :* un extrait d'acte de naissance, un passeport, etc. Il peut aussi avoir une valeur indiciaire : c'est un *témoignage*, une *preuve* dans un dossier (criminel, judiciaire, fiscal). Dans ce dernier cas, c'est un élément d'un regroupement opportun en vue d'un but, d'une fin. Le document, dans cette situation générale est *ancré* comme pièce, il est *autosuffisant*, encore que certains supermarchés aujourd'hui réclament deux pièces d'identité pour accepter un chèque.

Le document est *une partie de, un élément d'un dossier ou d'une collection* : ce qui compte alors, c'est sa qualification, son *indexation.* Tel timbre est un document d'un tirage postal historique, telle photo est un élément d'un reportage sur un voyage. Etc.

Le document est *un matériau d'une collection* ; il est composable avec d'autres éléments. Il est porteur d'une *interprétation* qui lui confère une certaine *détermination,* c'est-à-dire un mode d'être dans un domaine ou un univers de connaissance et de culture. L'important alors,

ce sont les critères qui vont lui être appliqués pour légitimer sa présence dans une *classification.* C'est sa *valorisation.*

D'où l'exemplarité de certaines entreprises de classification qui prennent forme de véritables collections. Tel est le cas historique et exceptionnel de l'*Encyclopédie* de Diderot et d'Alembert.

3. De l'encyclopédisme à l'encyclopédie, à travers Diderot et les Lumières : la naissance de la classification ouverte…

3.1. Une aventure tumultueuse

En 1745 à la suggestion d'un Anglais, John Mills, et d'un Allemand, Sellius, le libraire Le Breton annonce le projet de publier une traduction d'un *Dictionnaire des arts et des sciences*, paru en Angleterre, la *Cyclopedia* d'Ephraïm Chambers (1728), mais la brouille survient rapidement entre eux. C'est pourquoi, le 27 juin 1746, par devant d'Alembert et Diderot témoins, l'entreprise est confiée à l'abbé Gua de Malves, qui abandonne au bout de treize mois ; la main passe aux témoins, nommés codirecteurs, le 16 octobre 1747. Le projet s'élargit : un *Prospectus* le fait connaître en 1750. Rédigé par Diderot, il promet « un tableau général des efforts de l'esprit humain dans tous les genres et dans tous les siècles ». On prévoit huit volumes de textes et deux de planches et cinquante-huit collaborateurs (au total, ils seront plus de cent soixante).

Diderot est incarcéré à Vincennes pour sa *Lettre sur les aveugles* (1749) dans laquelle il dévoile son athéisme. La publication de celle-ci et bientôt l'affaire de l'abbé de Prades accusé de défendre la religion naturelle (1751) alertent les ennemis des Lumières contre l'entreprise encyclopédiste, qui engage de plus en plus de capitaux et de souscripteurs : mille à la parution du 1er tome, deux mille en 1752, trois mille en 1754 et 4200 en 1757. Dès février 1752, après la parution du tome 2, un arrêt du Conseil du roi interdit l'ouvrage.

Mais la publication se poursuivra, grâce à l'appui de Guillaume de Lamoignon de Malesherbes, qui nommé directeur de la librairie en 1750 et chargé de la censure, est un ami de Diderot. Dès mai 1752, d'Alembert pourra écrire : « A l'égard de l'Encyclopédie, toute la France désire qu'on la continue ; tout paraît apaisé et même d'accord ». Une nouvelle crise éclate quand est publié le tome 7 en novembre 1757. Ce tome 7 contenait l'article « Genève » dû à d'Alembert et où celui-ci faisait allusion au « socinianisme » des pasteurs genevois, en référence à Socin (1525-1562), réformateur italien niant la divinité de Jésus-Christ et le dogme de la Trinité. Cette affaire va consacrer la rupture avec Rousseau. Guerre de pamphlets. Voltaire conseille à d'Alembert de renoncer à ce « maudit travail ».

Diderot reste seul, refusant de décevoir les lecteurs et de ruiner les libraires. Avec l'attentat de Damiens (1757) et le scandale du *De l'Esprit* d'Helvétius (1758), défendant le matérialisme avec énergie, toutes les forces conservatrices sont rameutées. En juillet 1759, le Conseil du roi révoque les lettres de privilège de l'Encyclopédie et décrète même le remboursement des souscripteurs, mais aucun ne remboursera. Début septembre de la même année, le pape Clément XIII interdira l'ouvrage. L'œuvre doit donc s'élaborer dans l'ombre. Le Breton, l'éditeur apeuré, censure les textes. Mais une fois de plus, Malesherbes sauvera l'entreprise.

Le 8 septembre, cinq jours après la condamnation papale, un privilège est accordé à Le Breton pour un recueil de « mille planches… sur les Sciences, les Arts libéraux et les Arts mécaniques. » Enfin, les tomes 8 à 17 sont prêts et d'abord livrés sous le manteau au début de 1766 sous le couvert de « Samuel Faulche et Compagnie, libraires imprimeurs à Neufchastel ». Censés avoir été imprimés à l'étranger, ils sont l'objet, comme l'écrit Diderot, d' « une tolérance tacite ». Entre-temps, les volumes de planches seront distribués jusqu'au 11e et dernier en 1772.

3.2. Un modèle importé d'Angleterre

L'encyclopédie multi-volumes moderne est issue des petits dictionnaires des arts et des sciences apparus dès le début des années 1700. Le premier est celui d'Antoine Furetière : *Dictionnaire des arts et des sciences* paru en 1694. Dix ans plus tard, est publié le *Lexicon Technicum or an Universal English Dictionnary of Arts and Sciences*, de John Harris, qui connaîtra un grand succès. Puis en 1728, paraît la *Cyclopaedia* de Chambers dont on a vu qu'elle est à l'origine de l'*Encyclopédie* de Diderot et d'Alembert.

Chambers se donne pour objectif de fournir des définitions précises des différents termes utilisés dans les sciences et les arts. Ainsi que l'annonce, en 1726, la publicité faite pour inciter à souscrire : « Le caractère de cet ouvrage est d'être à la fois un DICTIONNAIRE et un SYSTEME. Il se compose d'un Nombre infini d'Articles qui peuvent être pris séparément comme autant de Parties distinctes du Savoir ; ou collectivement, constituant en cela un Corps. » Chambers vise d'emblée à rendre compte de l'unité et de la cohérence du savoir. De même, pour Diderot, il est clair que les premiers dictionnaires encyclopédiques, cantonnés au seul ordre alphabétique, sont insuffisants. Il reconnaît à Chambers le mérite d'avoir tenté de construire une rationalité au-delà du simple ordre alphabétique : « Il a bien senti le mérite de l'ordre encyclopédique, ou de la chaîne par laquelle on peut descendre sans interruption des premiers principes d'une science ou d'un art jusqu'à ses conséquences les plus éloignées et remonter de ses conséquences les plus éloignées jusqu'à ses premiers principes. »

Lorsque la *Cyclopaedia* paraît en 1728, ce qu'elle présente, c'est en effet une carte des savoirs. L'objectif, exposé dans une longue préface de vingt-quatre pages, tient en fait en une phrase[7]. « Les anciens lexicographes, écrit Chambers, ont rarement fait preuve d'une quelconque structure dans leurs ouvrages ; pas plus qu'ils ne semblent avoir eu conscience qu'un dictionnaire pouvait, dans une certaine mesure, offrir les avantages d'un discours continu. »

La classification des sciences de la *Cyclopaedia* incarne cette carte qui doit permettre au lecteur de naviguer dans le dictionnaire alphabétique selon une « Vue du Savoir ». Ce savoir est divisé en deux groupes : « Naturel et Scientifique » et « Artificiel et Technique », eux-mêmes fractionnés en subdivisions, à l'image de la « méthode des dichotomies » prônée par Petrus Remus dans les années 1600. Exemple : à l'issue de la première division, le savoir scientifique de la nature est scindé en deux sous-catégories relevant l'une du sensible, l'autre du rationnel, et qui permettent de distinguer – autre exemple -, d'un côté la météorologie, de l'autre, la géométrie. Mais la Cyclopaedia se distingue surtout par son *tableau*, clé selon Chambers, d'un dictionnaire scientifique.

Ce tableau se compose de 47 rubriques, numérotées selon leur position sur le diagramme, de Météorologie à Poésie. Les notes annexées à chaque art ou science indiquent les termes qui leur sont apparentés. L'objectif de ce tableau et des 47 notes de bas de page énumérant les termes affiliés à chacune des rubriques, est de rétablir le système des « lieux communs ». A la Renaissance, tout lecteur érudit était invité à mémoriser des citations d'auteurs classiques sur des thèmes littéraires et moraux. Ces citations étaient classées sous des « rubriques » ou notées comme « lieux communs » et chacun pouvait en disposer pour rédiger des discours ou des essais. Avec cette stratégie, Chambers entend différencier son encyclopédie vis-à-vis d'un simple dictionnaire de mots. Sa définition du « Livre de Lieux communs » décrit cette pratique comme une « collection ordonnée » de matières sous une « multiplicité de Rubriques »[8]. Il invite donc les lecteurs à utiliser le tableau et les notes comme guide de lecture en vue de

[7] Yeo, Richard, « Modèles d'outre-Manche », La Grande Encyclopédie, *Les Cahiers de Science et Vie*, 1998, n° 47, 24-27.

[8] Yeo, Richard, « Modèles d'outre-Manche », La Grande Encyclopédie, *Les Cahiers de Science et Vie*, 1998, n° 47, 24-27.

regrouper les articles épars afin de restituer à chacune des sciences sa cohérence. A l'intérieur du texte, il introduit également des renvois reliant les termes apparentés.

Pour Chambers, il y a déjà trop de livres. Il en appelle donc à la «réduction de la vaste masse du savoir universel à un plus petit volume » : un résumé pratique des connaissances essentielles d'une série de sciences, accessible à tous. C'est la raison pour laquelle, les unités de base de la *Cyclopaedia* ne se constituent que de mots et de termes.

De façon similaire, Diderot prédit un jour où le volume total des livres submergera les spécialistes eux-mêmes. Comme Chambers, il assimile l'*Encyclopédie* à une collection résumant le savoir d'une époque. L'un comme l'autre savent que leur œuvre nécessitera d'être revue en fonction des progrès du savoir.

Mais en prônant la nécessité d'une classification rationnelle des arts et des sciences, et en considérant l'encyclopédie comme le moyen de gérer le développement exponentiel du savoir, Chambers a produit un dictionnaire qui n'est pas un simple précurseur mineur de *l'Encyclopédie.* Néanmoins, les différences entre les deux entreprises demeurent : Chambers pouvait concevoir sa *Cyclopaedia* comme un ouvrage résumant le savoir en un nombre plutôt restreint de rubriques ; Diderot en revanche, a pressenti un avenir au savoir incontrôlable. « Et bien que les renvois contenus dans l'*Encyclopédie* soulignent l'intégrité des disciplines, ils invitent les lecteurs à un ensemble de discussions dans un domaine illimité. »[9] Déjà l'hypertexte ?

3.3 Que contient l'Encyclopédie ?

Dès le *Prospectus* publié par Diderot en 1750, le paradoxe de l'ordre encyclopédique est posé. Deux ordres se combineront en effet : celui alphabétique du dictionnaire, qui allouera à chaque matière une place identifiable dans les volumes, et celui systématique de l'arbre des connaissances humaines. Cet arbre traduit le souci philosophique de mettre en évidence les proximités, les comparaisons, les dépendances. Il reprend, en le remaniant, celui élaboré par Bacon cent trente ans plus tôt. Le philosophe anglais disposait l'ensemble des « doctrines humaines » en trois catégories : *Memoria, Phantasia, Ratio.* L'*Encyclopédie* reprend ces trois grandes subdivisions, mais dans un ordre différent. L'entendement humain procède d'abord de la *Mémoire,* cette faculté d'enregistrer les faits, puis de la *Raison* qui les compare, c'est ici la place de la science et de la philosophie, enfin de l'*Imagination*.

Les enseignements donnés aux jeunes gens à l'époque enregistraient, à l'exception de la rhétorique ou de la morale, nombre de résultats établis par les savants depuis la seconde moitié du XVII^{e} siècle. L'histoire naturelle, l'astronomie, la mécanique, les mathématiques, tous ces domaines ont connu des bouleversements. Les précepteurs et les professeurs se sont querellés, les ordres religieux ont dû statuer sur le bien fondé de l'introduction de telle ou telle matière nouvelle. Les jeunes collégiens ont assisté à ces tensions, se sont parfois trouvés confrontés à des débats de science et de métaphysique[10].

Dans ce monde, Bacon est l'une des figures emblématiques de la science nouvelle. Il symbolise le recours à l'expérience et au constat des faits. La référence à Bacon permet de ne pas raviver les querelles théologiques attachées à des savants continentaux tels Galilée, Descartes, Pascal ou Leibniz. Former le tableau des connaissances en s'inspirant de celui de Bacon, c'est aussi rendre possible, la comparaison entre l'état des connaissances un siècle plus tôt et celui que l'on va donner au lecteur. Le *Prospectus* promet un compte rendu raisonnable des connaissances. Or, dès qu'on parcourt l'arbre figuré, le renouvellement est évident. De Bacon, on reprend certes les trois grandes divisions, mais au sein de chacune, les catégories et leur articulation ne sont plus du tout les mêmes.

---

[9] Yeo, Richard, *op. cit.*

[10] Brian, Eric, « L'ancêtre de l'hypertexte », La Grande Encyclopédie, *Les Cahiers de Science et Vie,* 1998, n° 47, 30-38.

Dans l'article « Encyclopédie », Diderot expose les motifs qui animent les deux éditeurs. Toute ambitieuse qu'elle est, l'entreprise, assure-t-il, sera menée à son terme. Une nouvelle image de l'arbre encyclopédique est présentée, à travers le *Discours préliminaire* lui-même, le « Système figuré des connoissances humaines », et son explication en appendice. D'Alembert de son côté, répond aux objections qu'a suscitées le *Prospectus* en même temps qu'il obéit à la loi du genre : un ouvrage adressé au public lettré impose en effet une ample introduction, en forme d'exposé des motifs. L'ensemble formé par ces textes, par le *Prospectus* et par les articles «Dictionnaire » et «Encyclopédie » montre que les deux éditeurs ont voulu indiquer très explicitement une « vaste opération » (le mot est de Diderot dans le *Prospectus* qui désigne ainsi le travail préparatoire des animateurs du projet).

Mais au fil de ces textes, tantôt Diderot tantôt d'Alembert convoquent une riche panoplie de métaphores. Par leur diversité, elles suggèrent que l'essentiel n'est pas dans une forme particulière de représentation – l'arbre, la chaîne, le tissu, la carte, la machine, le labyrinthe ou l'édifice – mais dans le recours que tout artifice procure aux auteurs et aux lecteurs pour se déplacer par l'esprit dans l'ensemble du corpus.

Les deux éditeurs le soulignent : toute classification comporte une part d'arbitraire. Il faut entrer dans la matière et, comme l'écrit par exemple d'Alembert, « *ne […] pas attribuer à notre arbre encyclopédique plus d'avantage que nous ne prétendons lui en donner* ». Le « Système figuré » est donc un anti-système, un garant de la liberté de jugement du lecteur.

Le présupposé de l'«opération» encyclopédique, affiché par Diderot, est l'unité des êtres de la Nature et la diversité des opérations abstraites par lesquelles la raison s'en empare. Ainsi la section philosophique ou scientifique, qui occupe la colonne « raison», s'ordonne-t-elle selon un critère de degrés successifs d'abstraction.

Aux divers niveaux de l'arborescence, les matières sont toujours disposées, de bas en haut, depuis les sciences les plus proches des données physiques jusqu'aux formes les plus abstraites. Cette disposition, qui est l'objet même de la méthode, renvoie à deux siècles d'histoire des mathématiques. Des débuts de l'algèbre symbolique à l'essor du calcul infinitésimal, construire un calcul, le former pour le rendre adéquat à la solution d'un problème ou à la physique d'un phénomène, c'est avant tout savoir le bien disposer. C'est jouer sur sa forme même, au moyen des analogies et des combinaisons[11].

Pour d'Alembert, on peut multiplier les observations de faits, on peut écrire une histoire naturelle qui les enregistrerait, on peut les mesurer, les comparer, calculer des proportions, construire un calcul, le mener à son terme et ainsi dégager des lois : ce sont les étapes successives de l'abstraction mathématique. Pour un géomètre du milieu du XVIII^{e} siècle, l'analyse des modernes, entendue comme la tradition mathématique inaugurée par Descartes, est l'art de résoudre les problèmes en combinant la maîtrise de la décomposition et celle d'une panoplie de techniques de calcul.

A la même époque, Condillac théorise les rapports entre nature, langage et calcul, même si son influence au milieu du siècle est moins importante qu'elle ne le sera cinquante ans plus tard. Sa conception de l'analyse est autre : pour Condillac, le calcul est le modèle même d'un langage bien formé. Demeurer la question de la genèse des connaissances. Pour Condillac, elle est homologue du processus d'acquisition du savoir chez l'enfant. On peut donc comprendre l'attachement de d'Alembert à bien distinguer, dans le *Discours préliminaire*, l'explication analytique du « Système figuré» de l'exposition de sa genèse historique, entendue comme celle des progrès de l'esprit humain[12].

[11] Brian, Eric, « L'ancêtre de l'hypertexte », La Grande Encyclopédie, *Les Cahiers de Science et Vie*, 1998, n° 47, 30-38.
[12] Brian, Eric, *op. cit.*

3.4 Le mode d'emploi de l'Encyclopédie

Pour cerner comment opère l'*Encyclopédie*, il faut aller au-delà du tableau systématique, entrer dans l'ouvrage, et feuilleter les volumes eux-mêmes. Chaque article est introduit par le mot qui désigne la matière traitée. Il est le plus souvent suivi d'une abréviation en italique qui indique le domaine de compétence dont relèvera le commentaire : jurisprudence, astronomie, mathématique, médecine, etc. Ces termes évoquent le plus souvent les différentes branches de l'arbre du « Système figuré ». Un examen attentif montre que la logique de ces mentions repose plus sur le partage des compétences entre les différents collaborateurs du dictionnaire que sur la place exacte qu'on pourrait restituer à chaque article dans le tableau initial. Il en va de même pour les renvois vers d'autres articles dont on pourrait penser qu'ils traceraient un réseau cohérent couvrant l'un des embranchements de l'arborescence exposée au *Discours préliminaire*. C'est rarement le cas. Les tentatives de reconstitution menées à ce jour, par exemple en vue de l'édition critique des articles rédigés par d'Alembert, montrent que la logique des renvois est induite par la rédaction des articles, et non issue d'un schéma *a priori*. Même en ne considérant que les premiers volumes pour lesquels les conditions de préparation sont assez constantes, il est possible de discerner des dynamiques d'écriture étrangères au schéma de départ. Elles paraissent provenir de ce que les auteurs travaillaient en rassemblant des dossiers personnels, la lecture d'ouvrages de références, des corpus issus de l'activité de sociétés savantes telle l'Académie royale des sciences. Ainsi les renvois tissent-ils le plus souvent les liens entre les pièces éparpillées d'un puzzle rassemblé par les auteurs à titre préparatoire. Il s'y ajoute des renvois à des matières dont on attend qu'elles soient traitées par d'autres. Les aléas de la préparation, des collaborations et de la publication procurent parfois la surprise de renvois sans objet. Ces incohérences et ces imperfections s'expliquent le plus souvent par la durée de l'entreprise et par ses vicissitudes[13].

Il arrive que l'écriture même d'une série d'articles engendre la formation d'un petit traité, livré comme en feuilleton, qui prend ensuite son autonomie. C'est le cas après la publication de l'article « Absent » (sur les personnes disparues) au premier tome. Trois auteurs s'y succèdent : le juriste Toussaint, d'Alembert et Diderot. Ce dernier a le malheur de conclure l'article en vantant les mérites d'une théorie des probabilités attribuée à Buffon. D'Alembert, sceptique sur la question, rédige alors une série d'articles sur l'« analyse des hasards » destinés à exposer ses objections. Il livre la clé de la série en indiquant à l'article « Combinaison »: « *Voyez Jeu, Pari, Avantage, Probabilité, Certitude, etc.* ». Diderot, de son côté, saisit l'occasion d'une autre série d'articles pour s'essayer à cette matière, calculant, par exemple, les combinaisons géométriques qu'on peut former avec un carrelage particulier à l'article « Carreau ». Il y renvoie à une planche publiée ultérieurement sous le titre « Table des 64 combinaisons de deux carreaux mipartis de deux couleurs». A son tour, d'Alembert saisit l'occasion de ce même mot « Carreau » pour introduire un article titré « Franc-Carreau » (« Franc » en italique pour justifier la publication à cet endroit). Cela lui permettra de commenter les calculs de Buffon sur ce jeu et de renvoyer vers les articles de probabilité à venir. Ainsi, la puissance de la combinaison de l'ordre alphabétique et de l'ordre encyclopédique, les libertés ainsi permises et les contraintes alors induites, furent puissamment explorées par les encyclopédistes. L'*Encyclopédie*, dans sa première édition, fut un moment intellectuel exceptionnel[14].

Auguste Comte, Emile Durkheim, et Marcel Mauss ont abordé le problème des classifications des connaissances de façon peu orthodoxe dans les années 1900. Ils ont comparé les systèmes de représentations collectives étudiées par les ethnologues et les classifications opérées par les savants. C'est-à-dire, au fond, les deux extrêmes de ce qu'un encyclopédiste aurait pu attribuer aux progrès de l'esprit humain. Quelles sont leurs conclusions ? Dans tous les systèmes de classification, ils retrouvent une série de caractères essentiels communs : le recours à un système de notions hiérarchisées, le caractère spéculatif

[13] Brian, Eric, *op. cit.*
[14] Brian, Eric, *op. cit.*

de l'élaboration symbolique, la fonction d'unification des connaissances. Ils invitent leurs lecteurs à s'interroger sur la place des systèmes qu'ils qualifient de primitifs dans la genèse des classifications logiques. Le programme qu'ils esquissent n'est toujours pas accompli aujourd'hui.

3.5 Le progrès humain

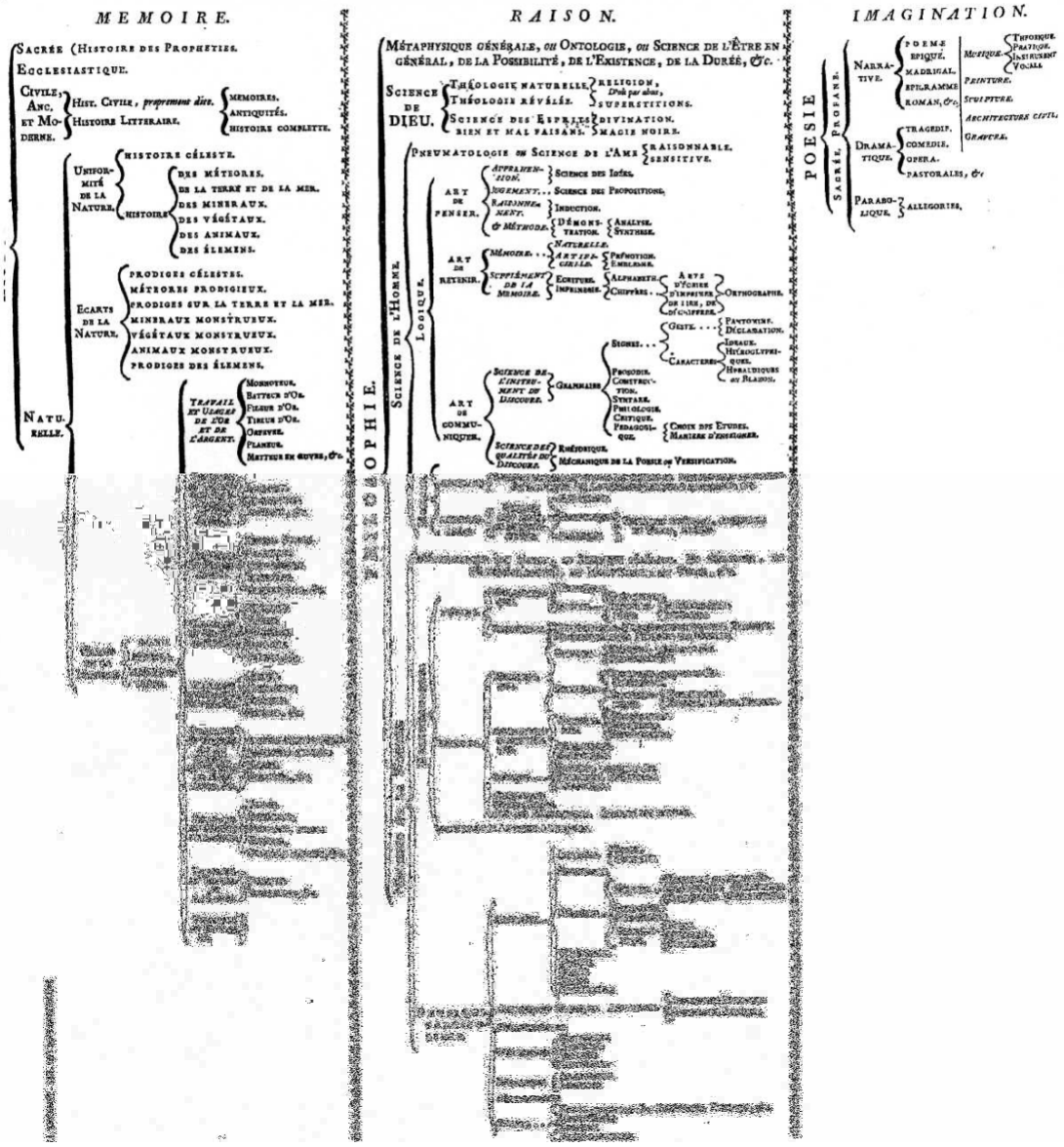
On peut s'interroger sur deux propriétés des systèmes de classification primitifs. La première est le fait que chacun forge son identité et donc ses relations avec les autres dans la référence à un élément situé dans l'unité du monde connu : un animal, une plante, qui procure un pouvoir magique. La seconde provient de ce que nous savons mieux aujourd'hui que les récits structurés par de tels systèmes de représentation sont flexibles, réguliers, mais non pas figés. Songeons au lectorat enthousiaste contemporain de l'*Encyclopédie* : le dictionnaire et son « Système figuré » pourrait bien présenter les mêmes propriétés. Un lecteur averti n'adhérait pas à tout l'édifice. Il avait ses prédilections et pouvait reconnaître comme sienne l'une des branches de l'arbre encyclopédique, l'un des territoires de la carte des connaissances. En identifiant la place de ses prédilections dans le panorama des connaissances humaines, on pouvait s'inscrire dans le tableau des progrès de l'esprit humain. Quant au récit global sur ces connaissances, chacun pouvait le reprendre à son compte à sa manière.

Est-ce à dire que l'emblème de la raison au XVIII^{e} siècle équivaudrait à un système mythologique ? Ce serait escamoter précisément la lente genèse des systèmes de classification, et par exemple l'histoire de la formation des critères logiques de classement. Aujourd'hui, une fois affranchis du dogmatisme des classifications du XIX^{e} siècle, les historiens perçoivent peut-être mieux les raisons conceptuelles et culturelles du succès de l'*Encyclopédie :* l'ouvrage était fait pour être consulté en tout sens, pour que, par la comparaison des différentes lectures se forme le jugement du lecteur. C'est une matière dans laquelle on naviguait, et les actuels concepteurs d'hypertextes y trouveraient volontiers un précédent. La grande *Encyclopédie* est déjà sur Internet. On peut y explorer le corpus avec une mobilité sans égale. Pourtant, la puissance de la première édition nous échappera encore tant elle provenait de l'adéquation d'un projet éditorial à un moment de l'histoire intellectuelle. Une aventure qui dura une vingtaine d'années au milieu du XVIII^{e} siècle. « Ce furent les conditions de la magie de ce moment de raison. »[15] Demeure l'exemplarité des problèmes de *catégorisation* formulés dans ce grand moment historique et dans ce grand œuvre.

[15] Brian, Eric, *op. cit.*

*SYSTÊME FIGURÉ
DES CONNOISSANCES HUMAINES.

ENTENDEMENT.



4. Catégorisation et schématisation : des « objets » au langage et à la collection

4.1. Une double interrogation récurrente

Il est une double interrogation récurrente depuis au moins deux mille ans, laquelle est : « Comment savons-nous et pouvons-nous savoir ce que nous croyons et disons savoir? » D'un côté, l'acception commune tend à nous faire considérer, même intimement, que ce que nous savons demeure tributaire de l'existence réelle des choses. D'un autre côté, la question de savoir si la réalité existe vraiment ou non demeure vaine, et tous les efforts des hommes n'y pourront rien.

Reste donc la question « Comment savons-nous ce que nous savons? » C'est la préoccupation actuelle, induite par notre propension à admettre que la connaissance, le corps, le cerveau, sont autant de « systèmes » et qu'ils peuvent se modéliser. A vrai dire, cette question, croyant éviter la précédente, multiplie la difficulté. Car pour comprendre comment nous faisons connaissance avec le monde et les choses, il nous faut sortir en quelque façon de nous-même, et dès lors, se risquer à des hypothèses sur la manière dont fonctionneraient nos processus mentaux.

La nature même de ces processus mentaux – et l'histoire de la psychologie le prouve –, n'est jamais évidente à cerner, encore moins à démontrer. Voici alors, réitérée l'interrogation première : si ce que nous savons est conséquence de la façon dont nous sommes parvenus à le savoir, alors notre conception de la réalité est entièrement déterminée par les processus qui nous ont menés à cette conception, sans que pour autant, nous ne sachions jamais ce qu'elle recouvre « réellement ».

Les développements de la connaissance sont donc autant tributaires des configurations d'objets sur lesquels ils portent que des moyens par lesquels ces configurations seront « présentées », c'est-à-dire « exprimées ». C'est dire que nous sommes toujours, ce faisant, en situation de « représentation des choses », et que les interrogations se trouvent en définitive, déportées sur ce moyen essentiel que nous avons de représenter et de nous représenter le monde et qui est le langage.

De ce côté-là, règnent encore quelques confusions, liées aux avatars de la linguistique. Avec le XX^{e} siècle, s'opère en effet, un renversement en regard des études antérieures largement consacrées aux langues indo-européennes : la linguistique sera présentée comme une discipline dont l'objectif est d'analyser le fonctionnement des signes du langage dans la vie sociale. Prennent essor alors, divers mouvements: descriptivistes (Sapir, Bloomfield), puis distributionnalistes (Hockett, Harris), enfin fonctionnalistes (Martinet). Mais avec les années soixante-dix, une rupture est délibérément introduite dans ces approches linguistiques: rupture concrétisée par la proposition faite par Chomsky, d'une *grammaire générative.* L'entreprise signifiait un retour au primat du syntaxique comme élément-clé d'une explication synchronique des phénomènes linguistiques ; elle réactivait le vieux débat entre empirisme et rationalisme au travers des controverses d'alors sur les parts respectives de *l'inné* et de *l'acquis.* De ce débat, nous ne sommes pas tout à fait sortis, bien que les divers courants génératives aient depuis, décliné. On comprend mieux aujourd'hui l'importance de l'acquis venant se construire sur les dispositifs innés fondant notre cognition ; on comprend mieux la nature empirique de la science du langage.

4.2. Réhabiliter l'empirisme

Les conceptions de l'empirisme ont souffert elles aussi de nombre d'ambiguïtés sinon de déformations. En réalité, l'empirisme, en tant que tradition philosophique, est cette conception qui consiste à valoriser le pouvoir qu'a notre esprit de traiter en permanence les expériences qu'il acquiert pour les reconstruire ensuite comme rationalités s'appliquant ultérieurement aux phénomènes qu'il rencontrera.

John Locke, mieux encore que Hume ou Mill, s'est attaché à décrire et analyser ces processus généraux qui nous font accéder à la connaissance. Plus que tout autre, il s'est insurgé contre le postulat métaphysique d'une constitution du savoir à partir d'idées ou de principes innés. Une telle innéité supposerait que dès la naissance, nous aurions une sorte de connaissance virtuelle. Pour Locke, faire un tel postulat, c'est nier toute conscience que nous aurions des corrélations entre nos idées. Or, à l'évidence, souligne-t-il, c'est l'esprit de l'homme qui élabore ses propres notions, qu'il s'agisse de raison, d'identité et de non-contradiction ou de morale. *Tout cela provient de principes abstraits qui sont produits par la réflexion de l'homme sur les idées que lui fournit l'expérience.*

Locke distingue deux types d'idées : les *idées simples* que nous impose l'expérience sensible, sans intervention de l'esprit ; les *idées complexes* qui viennent d'un travail de corrélation sur ces données de l'expérience et qui sont de trois sortes : les idées de *modes*, de *substances* ou de *relations*. Les idées de modes ne sont que des représentations d'idées simples; ainsi l'espace est-il la donnée sensible grâce à laquelle nous parvenons à concevoir les rapports d'étendue et de lieu. Les *idées de substances* se génèrent dans notre perception même : si nous partons en effet, des idées de sensations que nous avons – couleur, odeur, saveur, mobilité, solidité, etc. –, nous voyons que ces idées nous sont *secondes* par rapport à la perception immédiate, mais en même temps, nous découvrons que certaines de ces propriétés appartiennent aux réalités matérielles et qu'elles sont en vérité, des *qualités premières* des objets : stabilité ou mouvement, étendue ou solidité, etc. Ce qu'opère le travail de l'esprit, c'est de relier ces qualités secondes qu'il se construit des objets à ces qualités premières qu'il découvre en la matière ; en la réside son *pouvoir*, dit Locke[16].

Cette idée de *pouvoir* renvoie à la capacité qui est en chacun de nous, de traiter objectivement et de façon consciente, les expériences psychologiques et perceptives qu'il est à même de faire quotidiennement.

Ainsi, ce que nous propose Locke, c'est de faire au sens sémiotique - médical du terme -, l'analyse des signes à chaque fois, constitutifs des organisations de nos savoirs, et parmi ces signes, en premier lieu : le langage. Sans doute, celui-ci est-il à l'origine, pure convention, mais Locke insiste sur ce que le sens des mots est toujours relatif aux corrélations entre l'esprit et les signes linguistiques; de là, nos confusions. D'où la nécessité de travailler les déterminations du langage si l'on veut comprendre et analyser les contributions de celui-ci à la construction permanente de nos savoirs.

C'est le langage qui s'avère facteur constitutif de nos représentations abstraites des choses, à la fois instrument et signe obligé pour nous faire progresser d'idées simples aux connaissances et aux savoirs génériques. «Connaître, précise Locke, ne me semble rien d'autre que percevoir la connexion et la convenance ou le désaccord et la disconvenance entre n'importe laquelle de nos idées. »[17] Connaître, c'est tenter d'appréhender les relations de nos idées entre elles. Tout notre univers cognitif n'est ainsi qu'un univers de signes.

Une conception empirique, telle celle de Locke, peut inspirer nos précautions vis-à-vis des contradictions « métaphysiques », inhérentes au rationalisme de type chomskien ou post-cartésien, en même temps que contribuer à restituer au langage, ses deux propriétés essentielles :

Il est un système ouvert et au-delà du rapport désignatif à des «réalités », il est l'objet même de ses propres constructions;

Il est à la fois, support, instrument et fondation de nos activités de représentation et par suite, de nos actions symboliques sur le monde ; on ne peut ainsi analyser le linguistique sans prendre en compte cette capacité structurelle de « manipulation des signes », en conséquence

[16] *Essai*, III, XXI, I.
[17] *Essai*, IV, I, 2.

d'abstractisation générique vers la catégorisation permanente des choses et du rapport à ces choses.

4.3 Les « catégories » de la grammaire sont-elles les « catégories de la pensée »?

Il n'y a pas de rapport direct entre catégories de la pensée et catégories grammaticales, bien que demeure l'idée de la langue naturelle comme système reliant des signes à des significations[18]. Le statut des catégories grammaticales est ainsi ambigu : elles ne sont pas des classements univoques de la langue; elles ne suffisent pas à expliquer comment se génère une phrase ni *a fortiori* comment cette phrase sera interprétée. Elles ne sont en vérité, que des *états*, des « traces ». S'il nous faut comprendre ce qui opère dans l'activité langagière, cela se passe nécessairement à un autre niveau du type « générique », intriquant les rapports entre le sujet et le monde au travers de ces contraintes expressives que la grammaticalisation impose au langage pour qu'il soit «système », c'est-à-dire communément repérable d'un locuteur à l'autre. Il faut alors se donner les moyens théoriques et pratiques du passage d'une analyse des faits de langue, c'est-à-dire d'une « linguistique des états » à une « linguistique des opérations », essentielle si l'on veut rendre compte des rapports langage-cognition[19].

Les catégories de la grammaire ne sont donc pas la traduction de quelconques « lois de la pensée ». Ce dont elles témoignent, c'est de cette nécessité qui incombe à l'activité de langage de nous marquer des *repères* empruntés au système de la langue, repères des manipulations qui sont faites de ce système, sous forme de combinaisons syntaxiques, et conjointement, d'organiser ces agencements de repères de sorte qu'ils orientent vers des jeux expressifs sur des champs déterminés de référence. Celui qui écoute un discours de même que celui qui lit un texte ne va pas devoir se remémorer les distinctions subtiles qu'on lui aura apprises entre l'imparfait, le passé simple ou composé et le présent, mais il saura d'emblée que du point de vue aspectuel, la situation qu'on lui évoque, est actuelle ou révolue, ou renvoie à un passé lointain ou à ce qui vient juste de s'achever. Ce faisant, il utilise les marques que lui procure la langue comme «instructions de repérage » en vue d'une certaine compréhension de ce que son locuteur ou l'auteur du texte vise à représenter d'un certain état des choses. Il interprète le discours ou le texte en termes d'opérations que ce discours ou ce texte font sur la langue pour pouvoir communiquer une certaine visée de sens.

Les opérations langagières sont donc indissociables des activités cognitives, celles-ci prenant appui des premières ; celles-là n'étant motivées que par la visée d'organiser et de moduler les secondes. Toute catégorisation langagière – et ne disons plus grammaticale – va s'organiser comme une sorte de compromis tactique entre un système de marques expressives – la langue, mais aussi les jeux sur ses usages –, et le système des « régulations cognitives », c'est-à-dire des visées de connaissance ou plutôt de représentation des connaissances, qu'à chaque fois, de tels emplois permettent ou favorisent. Cette négociation permanente entre deux « systèmes » – celui du penser pour faire dire, et celui de l'exprimer pour amener à ainsi penser –, ne peut que résulter sous forme d'*opérations* qui ne relèvent ni du mental isolé en tant que tel, ni du linguistique, tel que la tradition nous accoutume de le classifier. Au contraire : il s'agit plutôt de jeux sur l'extension des formes et sur l'intension que ces formes peuvent véhiculer : des objets aux situations et aux collections, des schémas aux catégories.

4.4 De la schématisation à la catégorisation

Cette activité fondamentale qu'est la cognition, c'est-à-dire l'incessante dialectique de construction et reconstruction de nos connaissances en vue de les échanger, prend donc

[18] Jakobson, R. *Essais de linguistique générale*,. Paris, Minuit, 1963.

[19] Culioli, A., « Notes du séminaire de D.E.A, 1983-1984 », Paris, Université Paris 7, Unité de Formation et Recherche en Linguistique, 1985 et *idem, Pour une linguistique de l'énonciation,* Paris-Gap, Ophrys, 1990.

fondement dans le discours, medium essentiel de nos communications. Une double responsabilité pèse en conséquence sur celui-ci, et à travers lui, incombe au sujet énonciateur : celle d'abord, d'assurer un contrôle suffisant des formes de l'agencement linguistique pour que ces formes constituent des repères de marques et donc d'intentions ; celle ensuite, d'orienter, c'est-à-dire d'architecturer dans le discours, ces agencements de formes par la construction de parcours expressifs visant à l'indication *(deixis)* et à la recomposition intentionnelle de domaines ou de champs de connaissances, allant du concret à l'abstrait, du singulier au général.

Il y a jeu entre l'intensionnel – les « sens » retravaillés ainsi par le discours – et l'extensionnel – les types d'indicateurs et d'attributs choisis pour caractériser tel ou tel type de domaine ou de collection. Ces processus sélectifs que chacun de nous applique à ce qu'il exprime d'états ou de moments du monde font que chaque discours n'est guère autre chose qu'une *schématisation* à la fois de visions personnelles et de toutes les visions que chacun pourrait appliquer à ce que le discours là, désigne.

La notion de « schéma » appliquée ainsi à l'activité de langage répond à l'idée que tout discours n'est qu'une construction simplifiée, élaborée par un sujet énonciateur toutes les fois qu'il dit et commente sur et à propos de quelque chose. En d'autres termes, tout discours procède d'une réduction des éléments (acteurs, procès, domaines) que son sujet juge suffisants pour la représentation qu'il entend construire.

En premier lieu, il s'agit pour le sujet énonciateur, de s'assurer la maîtrise d'une progression des effets de sens, et cela au travers d'une détermination progressive des objets et situations dont il construit représentation[20]. (En second lieu, les significations que le discours véhicule doivent être perçues en état d'incomplétude, comme l'est un *schéma.* Cela, parce que le discours doit se donner comme « recherche » communément engagée avec son interlocuteur ou lecteur afin de mieux l'impliquer, et qu'il lui faut en conséquence, maintenir apparemment ouvert un « éventail des possibles » et par suite, une « labilité » des significations, favorable à leur convergence progressive dans le projet de sens qu'il souhaite imposer. La cohérence du schéma discursif assure alors une complétude apparente du parcours énonciatif en même temps qu'elle se voit contrebalancée – en vérité stratégiquement complétée – par cette incomplétude des déterminations successivement affectées aux objets et situations du discours. Ainsi, il y a nécessairement inhérente à tout discours, une *fonction schématisante* qui va opérer de la façon suivante:

- choix des objets que le discours veut traiter et la manière dont ils sont présentés : *mise en collection* ;
- « lecture » par le sujet énonciateur de ces objets sous la forme des modalités d'attribution de propriétés et des déterminations d'existence qu'il va leur affecter : *légitimation d'une collection* ;

construction en conséquence de schémas locaux, qui seront autant de filtrages, de réductions, de schématisations, de collections donc, des objets ainsi référés par le discours.

- Et cela, selon un processus de constructions successives opérant sur le rapport discours-« réalité », par le renvoi de chaque objet à une catégorie ou collection d'appartenance, et celui de chaque situation à un mode de catégorisation de son statut d'existence.

Le processus de la schématisation discursive est donc indissociable d'une catégorisation permanente des objets du monde dans une sorte de « relecture » incessante de leurs propriétés, et dans une non moins incessante réaffectation de leurs *indexations* à l'intérieur de collections, de catégories de faits, de situations ou de « mondes ». Cette dynamique de la catégorisation, que permet ainsi le discours, est essentielle à la régulation de nos jugements,

[20] Grize, J.B. *Logique et Langage*, Paris-Gap, Ophrys, 1990.

de nos conduites et de nos activités. Elle traduit bien l'intrication absolue entre nos opérations de pensée et nos opérations expressives et symboliques. Cela pour autant, ne suffit pas à dissiper les équivoques quant aux types de rapports entre états ou processus cognitifs et les formes d'interaction que ces états ou processus peuvent connaître avec nos activités de langage. Les premiers sont-ils tributaires des secondes ? Celles-ci sont-elles simplement résultantes de ceux-là ? On retrouve ici tous les problèmes évoqués précédemment. Et cela est illustré encore, au travers des études contemporaines sur la catégorisation.

4.5 Les études sur la catégorisation: un retour au « mentalisme »?

Ces études prennent historiquement origine à partir des travaux des anthropologues, lesquels avaient dégagé l'existence de régularités interculturelles dans les structurations de catégories d'objets universellement présents dans l'environnement : couleurs, animaux ou plantes[21]. Ces observations inspirèrent les psychologues, qui à leur tour, s'efforcèrent de vérifier l'existence de tels principes universels appliqués à nos catégorisations d'objets simples ou familiers. Leurs analyses, initiées plus particulièrement par E. Rosch[22], ont effectivement montré que nos processus de catégorisation d'objets témoignaient de la mise en œuvre de *principes* d'organisation des connaissances, motivés par la nécessité usuelle de réduire les complexités de l'environnement.

Deux « principes » furent ainsi dégagés : le premier faisant l'hypothèse que nous sommes capables de percevoir et de mémoriser des régularités sinon des invariances au travers des discontinuités entre éléments du monde; le second postulant que cette organisation des connaissances en mémoire témoigne de l'existence d'une structure propre au monde. Ainsi, nos processus de catégorisation traduiraient-ils à la fois, l'existence d'attributs constitutifs des objets et les formes d'organisation de nos « catégories mentales »[23]. Concrètement, cela signifierait que le niveau de généralité des catégories mémorisées refléterait le niveau de perception des similitudes et des discontinuités entre objets de l'environnement; les catégories seraient donc des «éléments primitifs » de la constitution et des fonctionnements de notre système cognitif, et de ce point de vue « mentaliste », il y aurait stabilité du système, attestée à l'intérieur de chaque sujet et entre individus. Cette hypothèse forte d'application directe de notre équipement cognitif aux découpages perceptifs de l'environnement, Rosch et d'autres[24] se sont efforcé de la démontrer, d'abord en tentant la reproductibilité des observations expérimentales – ce qui est loin d'être le cas et demeure tributaire du matériel expérimental choisi –, ensuite en spécifiant un certaine architecturation des phénomènes de catégorisation, destinée à prendre en compte d'une part, leurs variabilités, et d'autre part, leurs engendrements.

A un premier niveau, dit d'*économie cognitive*, il y aurait une sorte de «découpage de base » de la réalité, plus proche du concret que de l'abstrait, et qui fonderait ainsi l'organisation de nos connaissances. Puis interviendraient différents niveaux d'*abstractisation* croissante, allant du plus générique au plus spécifique. Ainsi Berlin[25], à propos d'espèces végétales, distingue-t-il cinq niveaux d'organisation : le règne (animal, végétal) ; la forme de vie (arbre ou fleur) ; le genre (chêne ou pin) ; l'espèce (chêne vert, pin parasol) ; la variété (chêne vert méditerranéen). Ces différents niveaux ne seraient pas de même importance cognitive, mais ils s'organiseraient de façon optimale à un niveau intermédiaire dit *niveau de*

[21] Berlin, B., « Ethnobiological Classification, *in* E.Rosch and B.B. Lloyd (Eds), *Cognition and Categorization*, Hillsdale, L. Erlbaum, 1978.
[22] Rosch, E., Lloyd, B.B., *Cognition and Categorization*, Hillsdale, Lawrence Erlbaum, 1978.
[23] Berlin, B., *op. cit.*
[24] Barsalou, L. W. et Sewell, D. R., « Contrasting the Representation of Scripts and Categories ». *Journal of Memory and Language*, 24, 646-665, 1985.
[25] Berlin, B., *op. cit.*

base, correspondant au mieux, à l'adaptation humaine à l'environnement selon le même principe initial d'économie cognitive.

A ce *niveau de base*, la discrimination des objets du monde en catégories se ferait effectivement en fonction des usages et des activités humaines. La définition du niveau de base dans la hiérarchie catégorielle se modulerait ainsi selon les connaissances et les pratiques des individus, leur expertise[26]. Ce niveau de base correspondrait encore à la « meilleure » coordination entre similitudes de formes et communautés d'attributs entre objets. Ces communautés d'attributs étant donc supposées s'organiser hiérarchiquement, sous forme « conceptuelle » et par suite, arborescente – les propriétés attachées aux concepts d'objets inscrites alors « harmonieusement » aux différents niveaux de la hiérarchie –, on comprend que ce modèle et ses dérivés aient connu un fort succès en intelligence artificielle[27]. Succès d'autant renforcé qu'on avançait encore que c'était à ce niveau de base que convergeait la communication verbale rassemblant les traitements cognitifs antérieurs. A la suite de ces propositions, de nombreux travaux de psychologie cognitive[28] allaient tenter de mettre en évidence que le découpage catégoriel s'opérait sur des configurations de parties à tout, en fonction d'*exemplaires* se distribuant selon un gradient allant du moins bon au meilleur, celui-ci étant supposé *typique* de la catégorie en question, et dès lors, considéré comme *principe organisateur* du rangement catégoriel considéré.

Toutes ces approches impliquent d'adhérer à quelques postulats dont rien ne prouve la véracité : tout d'abord, que le monde physique serait d'emblée porteur d'une organisation perceptible, et que le sujet humain, ce faisant, ne ferait que redécouvrir des « lois du monde » ; ensuite que ces « lois » s'organiseraient de façon analogue à ce que l'histoire de nos connaissances nous a accoutumés à ranger en *espèces* – il n'est pas innocent qu'autant d'expérimentations portent sur des classifications empruntées à la systématique botanique ou animale – ; enfin, et ce n'est pas toujours compatible avec ce qui précède, le système cognitif humain serait supposé « comprendre » et opérer de façon volontariste sur le monde, bien qu'étant reconnu en interaction avec son environnement. Rajoutons – et ce n'est pas mineur – que le rôle du langage – dans toutes ces approches – est majoritairement cantonné au domaine lexical ; ce qui tendrait à faire conclure que le langage ne serait que traducteur d'états de pensée réduits à des dénominations, et que le système cognitif – « la pensée », sous-entendu « le cerveau » – serait directement organisateur de nos expressions symboliques du monde.

En vérité, d'une part, l'organisation de nos catégorisations du monde est toujours tributaire de nos régulations pour l'action, et d'autre part, *nous sommes toujours en situation d'argumenter à la fois pratiquement et contextuellement, pour réordonner ces catégorisations qui n'ont d'autre fondement que téléonomique, et les ajuster discursivement aux contraintes de chaque situation.* Pour reprendre ici, un exemple de J.-F. Le Ny[29], entre Jean qui vit à Grenoble et Marie qui vit à Rennes, sans doute existe-t-il une représentation pour chacun de ces enfants, de ce qu'est « un chien » à partir de ce qu'il perçoit et de ce dont on lui parle, mais « il n'existe, en principe, aucune intersection entre l'ensemble fini des chiens particuliers réellement vus par Jean et celui des chiens vus par Marie ».

Cela se construit comme toute catégorisation visant à la collection, entre effectivement des perceptions plus ou moins abstractisées, mais toujours révisables, et des discours rapportés, des « jeux de langage » contribuant à cette révision même, et témoignant à chaque

---

[26] Dubois, D., Mazet, C., Fleury, D., « Catégorisation et interprétation de scènes visuelles : le cas de l'environnement urbain et routier », *Psychologie française*, n° spécial, « La psychologie de l'environnement en France », 1988 et Dubois, D. (éditeur), *Sémantique et Cognition*, Paris, CNRS, 1991.

[27] Quillian, M.R., « Semantic Memory », *in* Minsky M. (Ed), *Semantic Information Processing*, Cambridge, Mass., MIT Press, 1968.

[28] Tversky, B. et Hemenway, K., « Categories of Environmental Scenes », *Cognitive Psychology*, 15, 1983, 121-149.

[29] Le Ny, J.-F., *Science cognitive et compréhension du langage*, Paris, PUF, 1989.

fois, de la condition du discours comme lieu où un exemple viendra à expliciter puis affirmer « un exemplaire », lequel prendra statut du «typique », au gré des circonstances, des moyens et des objectifs. Que cela vienne ensuite comme fondation de catégories du monde, cela effectivement s'opère selon des mécanismes où *l'abduction* le plus souvent, s'offre comme condition favorable à la déduction et à l'inférence; la notion d'abduction étant prise ici au sens *peircien*[30], c'est-à-dire ce raisonnement ordinaire, qui s'appuyant d'un cas observé et y ajoutant une règle, en infère une déduction : «Ces haricots que je vois ici épars, sont blancs ; or tous les haricots de ce sac sont blancs; donc ces haricots proviennent (probablement) du sac. » Processus de parcours inférentiel, mais aussi et surtout, de schématisation organisant une catégorisation où à chaque fois, comme le remarque J.-F. Le Ny, tout schéma va se présenter comme « une structure d'information composée d'information constante et d'information variable », jouant d'attributs repérés et de valeurs affectées à ces attributs. Et cela témoigne une fois de plus, à l'évidence, de l'intrication étroite entre nos opérations cognitives – les types de traitement que nous appliquons au monde – et nos opérations langagières – les formes d'agencement et de retraitement du sens que le discours nous permet sans cesse, dans *le travail récurrent du discours sur les discours, du texte sur les textes : l'hypertextualisation.*

5. L'hypertexte, collection de collections

5.1. Les origines

On pourrait faire remonter l'origine de l'hypertexte à celle des bibliothèques. Une bibliothèque est d'une certaine façon, une immense base de données consultable dans laquelle peut « naviguer » un lecteur. L'autre analogie possible, on l'a vu, est celle de l'*Encyclopédie* telle qu'elle se conçoit au XVIII^{e} siècle : vaste base de données et organisation visant à l'exhaustivité du savoir.

En 1936, l'écrivain H.G. Wells propose l'idée d'une encyclopédie mondiale, qu'il imagine sous la forme dun réseau nerveux tissant des liens entre les travailleurs intellectuels du monde grâce à un média d'expression commun et grâce à l'unité produite par la coopération à la réalisation de ce projet commun[30].

De cette préhistoire de l'hypertexte, on pourrait conclure hâtivement ce rapport : la bibliothèque est à l'hypertexte ce que l'encyclopédie est à l'hypertexte pédagogique[31]. Mais les comparaisons s'arrêtent là car en vérité, l'histoire de l'hypertexte coïncide avec les développements technologiques qui le supportent.

Vannevar Bush

En 1941, Bush est directeur de l'office américain de recherche et de développement scientifique. Constatant que les informations et les rapports de recherche augmentent rapidement, il propose un moyen d'automatiser la collecte et la consultation de la documentation technique : le MEMEX, une machine multimédia à base de microfilms dont le nom fait penser à MEMoire et à indEX. Bush pensait à une machine capable d'entreposer les livres et les notes de chacun et à un mode mécanique de consultation rapide et flexible de toutes ces informations. Sans le nommer, il imaginait là déjà, l'hypertexte.

---

[30] Peirce, C. S., *Collected Papers*, vol. I à vol. VI, éd. par C. Hartshorme et P. Weiss, Harvard Univ. Press, 1931-1935; vol. VII et VIII, ed. par W. Burks, Harvard Univ. Press, 1958, II, 619-644.

[30] McKnight, C., Dillon, A., Richardson, J., Hypertext in Context, Cambridge University Press ; 1991.

[31] Rhéaume, J., « Hypermédias et stratégies pédagogiques », *in* de la Passardière, B. et Baron, G.-L. (éd.), *Hypermédias et apprentissages*, Paris, MASI, INRP, 1991 et Rhéaume, J., « L'enseignement des hypermédias pédagogiques », *in* Baudé., J. (éd.), *Deuxièmes journées francophones Hypermédias et apprentissages*, Paris, EPI, 1993.

Ted Nelson

Nelson participe lui, de l'ère des ordinateurs. C'est à lui que l'on doit l'invention du terme « hypertexte » et la conceptualisation associée. Nelson est un original. Dans ses conférences il se décrit comme un «computopien», un utopiste de l'ordinateur. Le terme «hypertexte» porte trace du computopien. Dans cette lignée, surgissent toute une série de termes ordinaires: « hypermédia, hyperdocument, hyperbase, hypergraphique, hyperespace », etc. Nelson a même suggéré le terme «hypergramme» pour désigner un graphique à l'ordinateur dont les parties exécutent des animations lorsqu'elles sont activées[32].

Pour nommer son projet, l'original Nelson se rappela du poème de Coleridge où la coupole du plaisir est appelée Kubla Kha. De ce nom, il dériva «Xanadu», un grand projet hypertextuel dont l'objectif était de créer une structure permettant de relier toute la littérature du monde dans « un réseau de publication hypertextué universel et instantané »[33]. Concrètement, Nelson voulait créer une nouvelle encyclopédie[34].

Douglas Engelbart

Tandis que Nelson a des visions, l'ingénieur Engelbart songe à construire de vrais environnements d'hypertextes à l'Institut de recherche de Stanford. Engelbart est principalement connu pour le développement de ses interfaces, notamment de la fameuse souris qui accompagne maintenant tous les ordinateurs. En 1968, il présente le premier système informatique fonctionnant sous mode d'hypertexte, le NLS, pour « oN Line System », une sorte de base de données qui facilite le travail en collaboration puisque tous les intervenants sont reliés en réseau à l'ordinateur.

Engelbart veut amplifier l'intelligence humaine, ce qui suggéra le titre de son deuxième projet : *Augment*, développé au Centre de recherche pour l'augmentation de l'intellect humain qu'il a fondé à Stanford. *Augment*, commercialisé par McDonnell-Douglas, est un environnement en réseau, de traitement de textes et de gestion d'idées, qui permet la collecte des documents, des notes et des rapports de recherche tout en fournissant des moyens de planification, d'analyse et de communication. Engelbart fournit donc les premiers outils de l'hypertexte, qui, selon ses ambitions, ne limitent ni ne contraignent les gens les plus habiles. Il souhaite ainsi encourager la performance et l'excellence[35].

Bill Atkinson

Atkinson est un des personnages légendaires d'Apple, qui a indirectement aidé à populariser l'hypertexte. Il a d'abord conçu les premiers éditeurs graphiques puis Hypercard, un logiciel qui permet d'en bâtir d'autres, comme il le disait lui-même. Ce logiciel n'était pas spécifiquement conçu pour bâtir des hypertextes. Pourtant, sa distribution gratuite et sa facilité d'utilisation tendait à populariser les hypertextes. Ainsi, à mesure que les concepts et les usages, notamment pédagogiques, se précisaient, le style hypertexte s'affirmait comme une nouvelle technologie intellectuelle qui, en retour, exigeait une technologie informatique appropriée.

5.2. L'histoire de l'hypertexte par les logiciels

Si l'histoire de l'hypertexte a commencé par des hommes, elle se décline ensuite à travers de multiples logiciels et projets de développement qui valorisent tous un aspect ou l'autre du concept: askSam, Black Magic, Document, Examiner, gIBIS, Glasgow On-Line, Guide, Hypercard, Hyperlog, HyperTIES, Intermédia, KMS, KnowledgePro, Linkway, NaviText,

[32] McKnight, C., Dillon, A., Richardson, J., Hypertext in Context, Cambridge University Press ; 1991.
[33] Nelson, T., *Literary Machines*, Swathmore, Pa., 1981.
[34] McKnight, C. *et al*., *op. cit*.
[35] Englebart, D., « Authorship provisions in AUGMENT », *IEEE Comp-Con Proceedings*, Spring, 1984.

Neptune, NoteCards, StrathTutor, SuperBook, SuperCard, Toolbook, Thoth-II, WE, Writing Environment, etc. Toutes ces applications ont eu pour but, d'abord de gérer des masses de données et ensuite d'aider à baliser des navigations capables de transformer ces données en informations structurées puis en connaissances significatives.

Cela implique au plan cognitif, de construire des réseaux qui donnent sens à des «objets cognitifs », qui sont au départ, des données en vrac, lesquelles données deviennent des informations lorsqu'elles sont structurées et des connaissances lorsqu'elles sont humainement assimilées et rendues significatives pour la compréhension humaine.

5.3 Qu'est-ce que l'hypertexte ?

Le terme «hypertexte » désigne un texte électronique composé de blocs de textes liés entre eux de manière non séquentielle. Le Web en est un exemple. Ce type de présentation d'un ensemble de textes constitue une rupture avec les présentations textuelles traditionnelles de l'information. Elle permet à l'utilisateur de choisir un «parcours » dans un ensemble de données (texte, image ou son). On pourrait aussi parler de « méta-texte », dans le sens où une nouvelle dimension est ajoutée au texte imprimé.

Contrairement en effet, au texte imprimé qui est paginé de manière linéaire et conçu pour être lu dans cet ordre, l'hypertexte se présente comme des pages ou écrans accessibles selon toutes sortes de relations ou de séquences pertinentes pour le lecteur. Tout lecteur a la liberté de lire un texte ordinaire sur papier de façon linéaire ou non-linéaire, c'est-à-dire en sautant directement aux passages pertinents. Le lecteur d'hypertexte conserve cette liberté mais, contrairement au livre, la lecture linéaire, d'écran à écran, n'y est pas synonyme de structure ou de suite. Le lecteur d'hypertexte est constamment appelé à voyager jusqu'à un autre nœud à cause d'un type particulier de relation et non parce que c'est la page suivante. Le lecteur d'un hypertexte est donc interactivement invité à se transformer en «auteur » à chaque fois qu'il doit relier entre eux, de manière significative, des éléments d'information. Le parcours d'un hypertexte est plus exigeant que la lecture d'un livre linéaire parce que la question de la pertinence de ce qui est lu est sans cesse remise en cause.

Nelson a construit le terme hypertexte pour parler d'une organisation non-linéaire de l'information[36]. Il pensait à une information sous forme linguistique. Lorsque l'idée fut popularisée, on a vu apparaître aussi le terme hypermédia qui correspond à la même définition sauf qu'il précise que les informations peuvent emprunter divers supports ou médias comme les graphiques, les images numérisées, les animations, les séquences vidéo, les séquences audio, les animations d'objets réels externes ou robots, etc.

À l'intérieur d'un hypertexte, les unités d'information sont appelées nœuds et correspondent à un écran, à une page ou à des fenêtres sur un écran. Chaque nœud peut en principe être relié à une multitude d'autres nœuds par des liens. Les nœuds et les liens sont les éléments constituants des hypertextes.

« Dans cette acception, un ensemble de nœuds s'appelle un réseau ou une base de données; un jeu de liens s'appelle une navigation si l'objectif recherché est précis, un tour guidé si le cheminement est proposé par un tuteur et un broutage ou butinage, si le lecteur évalue chaque îlot d'information à son mérite. »[37]

---

[36] Shneiderman, B. et Kearsley, G., *Hypertext Hands-on !*, Reading, Ma., Addison-Wesley Publishing, 1989.

[37] Rhéaume, J., «Hypermédias et stratégies pédagogiques », *in* de la Passardière, B. et Baron, G.-L. (éd.), *Hypermédias et apprentissages*, Paris, MASI, INRP, 1991 et Rhéaume, J., « L'enseignement des hypermédias pédagogiques », *in* Baudé., J. (éd.), *Deuxièmes journées francophones Hypermédias et apprentissages*, Paris, EPI, 1993.

5.4 Le nœud : unité d'information

Le nœud est l'unité minimale d'information dans un hypertexte. On parle aussi de bloc, d'îlot ou de « frame » ou de « script », si on se réfère à diverses théories cognitives. Dans un nœud, l'information est modularisée ; dans un texte, elle est linéarisée. Pour préciser la grosseur d'un module ou la quantité d'information d'un module, on parle de granularité.

Chaque module ou nœud comprend idéalement une seule « idée », concept, ou sujet qui peut s'accrocher à d'autres (par des liens) qui lui sont naturellement connexes ou à d'autres qui dépendent du choix de l'usager. Les nœuds connexes peuvent être des exemples, des élaborations ou des idées nouvelles. Le support d'un nœud d'information peut être une page, un écran, une carte, une partie d'écran appelée fenêtre, si l'information est textuelle. Si l'information n'est pas uniquement textuelle, le support d'un nœud peut être un graphique, une animation, une image, une séquence de vidéo ou d'audio ou un autre élément externe comme une maquette, etc. L'information contenue dans un nœud peut être modifiée la plupart du temps. Les nœuds d'information peuvent être de divers types: définition, attributs, références, notes, illustrations, exemples, etc. Les idées sont dans les nœuds. Un ensemble de nœuds forme un réseau. Ce réseau correspond à la structure de la matière ou au réseau sémantique de l'usager. L'ensemble des nœuds forme une base de données emmagasinée dans la mémoire de l'ordinateur.

5.5 Les liens entre nœuds

Si on considère l'information, le lien serait le passage à d'autres informations connexes. L'ensemble des liens fournit alors les structures du document. Le type de relation entre des nœuds est souvent indiqué textuellement ou iconiquement: théorie de, exemple de, partie de, vient de, aller à... Comme l'usager est maître des liens qu'il active par la souris, l'écran tactile, etc., il contrôle ainsi la séquence de l'information qui lui est présentée. Des liens peuvent aussi faire le pont entre des documents, soit d'autres hypertextes, soit des nœuds externes comme une image par exemple.

Tout document est structuré selon au moins deux types de liens : les liens référentiels et les liens organisationnels.

• Le *lien référentiel* uni ou bidirectionnel est celui qui établit la relation entre un élément inscrit dans un nœud et un élément de référence inscrit dans un nœud destinataire.

• Le *lien organisationnel*, comme son nom l'indique, touche la structure ou la hiérarchie d'un hypertexte construit sous forme d'arbre : le nœud parent (une définition) est relié par lien organisationnel à un nœud enfant (un exemple, une application, etc). Les liens sont la base de la navigation qui est plus pré-organisée ou plus libre, précisément selon le type de liens. A l'écran, les liens sont indiqués :

par un bouton reconnaissable, avec ou sans icône,

par une marque dans le texte,

par une consigne générale sans signe particulier.

5.6 La « navigation » : cheminement, sentier, tour guidé

Le cheminement est une séquence de nœuds d'information pertinents à un objectif de navigation. Tout comme l'auteur d'un livre suggère de lire son ouvrage de la première à la dernière page, l'auteur d'un hypertexte peut suggérer sous forme de menu ou de carte des itinéraires convenant à telle ou telle circonstance. Pour l'internaute qui n'a pas d'objectifs de navigation en tête, un tour guidé peut être offert en guise de sentier tutoriel, par exemple. Le tour guidé est aussi appelé le cheminement par défaut. Les sentiers sont des adaptations de l'information aux besoins ou aux caractéristiques des usagers. Par exemple, des sentiers plus graphiques peuvent être offerts aux usagers qui apprennent mieux visuellement. Le cheminement peut aussi désigner le parcours de navigation effectivement suivi par un usager

à travers un hypertexte, de manière à pouvoir retourner à des nœuds vus antérieurement. Cette trace du cheminement est très utile, pour l'auteur, au moment de la construction d'un hypertexte; pour l'enseignant, au moment d'évaluer le cheminement d'un apprenant ou pour tout usager, simplement comme mode personnel de navigation. Le cheminement peut être volontairement marqué par des signets qui permettent à l'usager de retourner à des endroits spécifiques.

5.7 Base de données

La base de données est le lieu informatique où est emmagasinée l'information dans le but d'y accéder facilement. Tout hypertexte commence par une base de données. Cependant l'hypertexte dépasse la base de données en ce qu'il autorise une représentation de l'information de multiples dimensions. C'est-à-dire que les liens ne sont pas limités aux structures bi-dimensionnelles de la base de données.

Le concept de base de données s'est répandu avec l'ordinateur mais il était déjà connu à travers les dictionnaires, les annuaires, les encyclopédies. Chaque type de base de données est organisé différemment selon le genre de recherche à effectuer. Par exemple, les pages blanches et jaunes de l'annuaire du téléphone proposent deux voies pour retrouver la même information. Une base de données comprend souvent un schéma de base pour regrouper l'information de même nature et faciliter le repérage. Pensons à une fiche de bibliothèque avec le titre du livre, le nom de l'auteur, une cote, etc. A un niveau de structure plus élevé, on peut penser aux fichiers-titres et aux fichiers-auteurs.

5.8 Quatre points de vue

De façon générale, les définitions de l'hypertextualité combinent quatre points de vue :
- l'exploration d'une vaste base de données,
- l'accès à une information enrichie sur un sujet donné,
- la personnalisation d'une base de données et
- la construction d'une base de données.

Cette classification est très proche de l'outil cognitif qu'est l'hypermédia[38].

L'exploration d'une vaste base de données

L'exploration de type hypertexte induit un problème inhérent au système. La simple reconnaissance de liens isolés suffit-elle à promouvoir un mode de pensée non linéaire. Un nœud ou îlot d'information, aussi pertinent soit-il, sera-t-il immédiatement replacé dans un contexte propice à la compréhension d'un sens plus général ? Le problème est donc celui de l'image d'ensemble qui serait dégagée des relations relevées. Le sens est probablement la qualité d'un hypertexte qu'il faut surveiller le plus. Dans un livre ordinaire, il est facile de situer un paragraphe, une phrase ; dans un hypermédia, c'est le lien entre les nœuds qui établit la pertinence et qui fait ressortir le sens. Par contre, à défaut de révéler un sens, le lecteur n'a pas beaucoup de moyens pour s'orienter. La désorientation alors conduirait à la perte du sens.

Les techniques qui s'intéressent à l'analyse, aux liens et à l'organisation de la connaissance comme l'information mapping abordent cette question[39].

Un autre problème majeur est celui de la décontextualisation que peut opérer l'hypertexte soit parce que l'information désirée n'est pas trouvée ou que l'usager ne peut replacer les items d'information dans un tout reconnaissable. Les questions des usagers sont alors

---

[38] Duffy, T.M. et Knuth, R.A., « Hypermedia and instruction: where is the match ? », *in* Jonassen, D. and Mandl, H. (eds), *Designing Hypermedia for Learning*, Heidelberg, Springer-Verlag, 1990 et Rhéaume, J., « Hypermédias et stratégies pédagogiques », *in* de la Passardière, B. et Baron, G.-L. (éd.), *Hypermédias et apprentissages*, Paris, MASI, INRP, 1991.

[39] Horn, R., *Mapping Hypertext*, The Lexington Institute, Ma., 1989.

primordiales: d'où est-ce que je viens, où suis-je, où est-ce que je vais? Une foule de techniques peuvent atténuer ces difficultés: des signets, des cartes, des retours au point de départ, des listes de liens, des marche-arrière[40].
On peut penser que le morcellement de l'information dans les hypermédias conduit à une modification culturelle du mode de pensée. La désarticulation des messages et leur multiplication dans le temps et l'espace créerait une illusion de connaissance qui occulterait la compréhension d'ensemble. Il n'y a pas de solution technique à ce problème; la solution réside chez l'usager qui consulte l'hypermédia pour répondre à un objectif qui lui est propre.

L'accès à une information spécifique sous forme de base de données repose la question de la pertinence de l'hypertexte en termes familiers. L'utilisateur doit acquérir une information de base. Si elle est adéquate, il poursuit sa démarche, sinon, il demande une série d'élaborations sous forme d'exemples ou d'explications. L'hypertexte peut donc répondre aux exigences personnelles des utilisateurs quand il s'agit de comprendre des concepts ou des relations entres concepts dans un domaine bien déterminé.

La personnalisation d'une base de données fait ressortir la dimension utilitaire de l'hypermédia. L'outil permet de refaçonner une base de données existante pour répondre aux besoins spécifiques d'un usager. Par la classification, la compilation, l'analyse des données, la représentation visuelle et l'enchaînement des données, l'hypermédia devient porteur de sens, du moins pour le bricoleur lui-même. Si on admet que le travail avec l'information peut susciter l'apprentissage, on peut dire que l'hypertexte peut contribuer à une pédagogie de la construction, de la réparation, de l'innovation, de l'ajout. Cela s'effectue de trois manières:
• par la possibilité de juxtaposer des îlots d'information,
• par la possibilité d'annoter la base de données en y ajoutant des commentaires personnels à la manière du souligné dans l'imprimé et
• par la possibilité de créer des liens personnalisés.

A partir de la base de données, une nouvelle structure peut être construite pour répondre à un objectif très particulier, externe à l'hypertexte de base, comme celui d'écrire un article par exemple[41].

Cette approche fait en sorte que l'hypermédia n'est jamais un produit terminé, mais demeure un lieu d'expression, de mémoire et de communication en constante évolution.

La construction d'une base données selon les caractéristiques de l'hypermédia fait de tout usager un auteur. L'hypertexte convient bien pour la rédaction de documents complexes, le traitement d'idées, surtout si la tâche s'effectue en groupe. La collaboration peut s'effectuer entre le professeur et les étudiants : le guide et l'apprenant juxtaposent des points de vue qui bâtissent une situation d'enseignement/apprentissage[42].

L'apprentissage par la construction d'un hypermédia est un domaine neuf dont tous les principes n'ont pas encore été élaborés. En l'absence de tradition, une telle construction peut s'effectuer d'abord par imitation. Les lieux à considérer pour des emprunts éventuels sont les situations d'enseignement individualisé par l'ordinateur et les autres médias audiovisuels, les

[40] Horn, R., *Mapping Hypertext*, The Lexington Institute, Ma., 1989; Shneiderman, B. et Kearsley, G., *Hypertext Hands-on !*, Reading, Ma., Addison-Wesley Publishing, 1989.
[41] Duffy, T.M. et Knuth, R.A., « Hypermedia and instruction: where is the match ? », *in* Jonassen, D. and Mandl, H. (eds), *Designing Hypermedia for Learning*, Heidelberg, Springer-Verlag, 1990.
[42] Collins, A., Brown, J.S., & Newman, S.E., « Cognitive apprenticeship : teaching the craft of reading, writing, and mathematics », *in* L.B. Resnick (Ed.), *Cognition and Instruction: Issues and Agendas*, Hillsdale, NJ, Erlbaum., 1988.

manuels, les bases de données. Par exemple, les appendices, les bibliographies, les index, les schémas peuvent être interactivement exploités dans un hypermédia destiné à l'enseignement.

Il faut aussi considérer le passage de la pensée à la structure du document. Le passage de la pensée à l'hypertexte devrait s'effectuer naturellement car il semble y avoir une parenté entre le mode de pensée d'un humain et l'hypertexte. En effet, nous apprenons en replaçant mentalement toute nouvelle information près des idées que nous possédons déjà dans un domaine analogue. L'apprentissage comme la pensée se font par des relations significatives ou associations entre les idées. C'est ce qu'on appelle un réseau sémantique. L'hypermédia peut devenir un outil de structuration de pensée au même titre que la langue.

La structure d'un hypermédia est plus qu'une imitation du mode de penser et plus qu'une application des principes de la planification de l'enseignement, c'est aussi un respect de la matière envisagée. Si on considère *l'information mapping*, l'une des rares méthodes d'hyperécriture, il y a analogie entre la carte géographique qui suit le contour d'un terrain et la structure de l'hypermédia qui suit le contour de la matière décrite[43]. Dans sa perspective, Horn appelle auteur un analyste qui hiérarchise et classifie les nœuds d'information d'après leurs ressemblances et leurs différences. Cette technique a fait ses preuves bien avant l'arrivée des systèmes informatiques. Il faut néanmoins considérer les caractéristiques du système utilisé. Composer un hypermédia, c'est créer des nœuds et des liens. Ces nœuds doivent contenir chacun une seule idée bien identifiée par un titre. La grosseur d'un nœud devrait correspondre à l'espace de la mémoire à court terme selon la technique de *l'information mapping*.

Selon cette technique, quatre principes devraient caractériser ces nœuds. D'abord, l'information doit être partagée en petites unités ou blocs, ce qui correspond à un nœud ; ensuite, un nœud ne doit contenir que l'information relative à un aspect de la question ; puis dans un sujet donné, les blocs d'information doivent présenter une certaine similitude quant aux mots, aux titres, aux formats et aux séquences; enfin chaque nœud doit être étiqueté selon des critères spécifiques[44]. D'autre part, les liens doivent établir des relations pertinentes entre les nœuds: unité de classe et de genre. Les liens doivent donc en quelque sorte établir le réseau ou tableau d'ensemble qui montre le contour d'une question.

5.9 Quelques problèmes liés à l'hypertexte

Désorientation cognitive : Une autre limite importante de l'hypertexte, c'est qu'il n'y a pas encore de « grammaire » qui permette de saisir d'un seul coup d'œil les différentes formes de continuités et d'enchaînements qu'un lien nous apportera si on l'active. Si on active, par exemple, un lien sur un mot, est-ce qu'on aboutira à une définition de trois lignes ou à une thèse de doctorat portant sur ce thème? Le lien ne nous le précise pas. Ceci amène deux types de désorientation cognitive : le problème du « musée d'art » où on voit tellement d'informations et d'images qu'on ne sait plus ce qui est relié et on ne retient rien. La richesse de la représentation non linéaire hypertextuelle porte en elle le risque d'une « indigestion intellectuelle ». Le Crosnier nomme le deuxième type de désorientation, les « digressions imbriquées » où on perd de vue le but de sa recherche en suivant les chemins de traverses offerts par l'hypertexte[45].

---

[43] Horn, R., *Mapping Hypertext*, The Lexington Institute, Ma., 1989.

[44] Horn, R., *op. cit.*

[45] Le Crosnier, H., « L'hypertexte en réseau : repenser la bibliothèque », *Bulletin des bibliothèques de France*, 40, n° 2, 1995, p. 23-31 ; Jonassen, D.H., *Hypertext/Hypermedia*, Englewood Cliffs, New Jersey, Educational Technology Publications,1989.

5.10 Conclusion : des défis passionnants

L'intrusion des grands serveurs commerciaux dans l'Internet risque de poser de nombreux défis aux bibliothécaires, documentalistes et spécialistes de l'information. La NASA distribue des fonds importants pour le « Public Use of Earth and Space Science Data Over the Internet » visant à développer des technologies pour la bibliothèque numérique (voir par exemple <http://dlt.gsfc.nasa.gov/>). Des projets analogues se développent à Stanford, Berkeley, Carnegie Mellon. Il faut espérer que cette initiative de la NASA se répande et que les gouvernements auront assez de vision pour l'appuyer. Actuellement, il semble que ce soit le cas : des bibliothécaires figurent au sein des conseils d'administration de plusieurs groupes de travail sur les autoroutes de l'information (aux États-Unis, au Canada). Devant la complexité des services sur l'Internet, il y aura en effet, un besoin de plus en plus grand de *cyberthécaires* pour effectuer, à l'échelle de la planète, le travail effectué depuis toujours dans nos bibliothèques: acquisition, sélection, conservation, diffusion, catalogage, indexation. Ces besoins sont décuplés à cause de la masse d'information électronique qui déferle sur les usagers désorientés. D'où la nécessité de concevoir des outils de catalogage, d'indexation et de résumé automatique et, d'une manière générale, de développer les méthodes de traitement automatique du langage naturel.

Autres problèmes : qui s'assurera de la conservation des documents électroniques ? Comment assurer l'intégrité de données si facilement modifiables? Comment s'assurer de localiser de façon permanente un document hypertexte? L'URL (Uniform Ressource Locator) permet de localiser un document à un endroit mais qu'arrive-t-il quand ce document est déplacé, effacé ou modifié? Où sont les objets perdus sur Internet? On parle d'allouer des URN (Universal Resource Name), une localisation qui serait permanente, un peu comme les ISBN. Comment citer un document hypertexte?

D'autre part, dans l'Internet, le texte le plus génial peut côtoyer la sottise la plus pure. Comment valider et sélectionner ces textes et établir des liens avec d'autres hypermédias, et sur quels critères? Autrefois, la tâche était relativement simple lorsque le savoir était « stabilisé » dans un livre. Maintenant, nous devons apprendre à travailler avec un savoir mouvant, nous devons gérer des flux d'information.

La technologie hypertextuelle doit être vue dans un continuum auquel les spécialistes de l'information ont toujours participé. La connaissance de cette technologie doit également permettre de comprendre ce qu'il y a de révolutionnaire dans l'hypertexte et d'élargir nos référents (le livre notamment) pour entrevoir les potentialités de ce nouveau médium de diffusion de l'information.

C'est dans ce contexte que nous avons développé l'hypertexte CoLiSciences.

6. CoLiSciences : une collection historique, un hypertexte de référence

6.1. Le projet CoLiSciences

En 1999, devant le constat d'une quasi absence de ressources en ligne offrant des corpus francophones traitant de sciences – alors que les sites de ressources textuelles en littérature sont fort nombreux –, l'équipe « Hypertextes et textualité électronique » du Laboratoire Communication et Politique (Directeur : Georges Vignaux) a conçu un projet visant à proposer un COrpus de LIttérature Scientifique de langue française *(CoLiSciences)* – en l'occurrence, la biologie. Un site prototype fut alors élaboré, consacré à un ouvrage de Claude Bernard, l'*Introduction à l'étude de la médecine expérimentale* (IEME). Cette étape permit : (i) de dresser l'inventaire des diverses difficultés pratiques (éditoriales et informatiques) qui accompagnent tout développement d'un projet de ce type ; (ii) de mettre en œuvre les idées théoriques sur l'hypertextualité produites au sein du LCP.

Des logiciels, des interfaces utilisateurs, des modes d'accès aux textes, etc., furent évalués, ce qui nous autorise maintenant à pérenniser des choix technologiques précis (XML pour les textes en ligne, logiciels libres [Apache comme serveur Web, php comme langage de scripts et mySQL comme base de données] pour l'environnement informatique, etc.), grâce auxquels nous développons dorénavant le programme Colisciences :

<http://www.colisciences.net>.

2. Les ambitions de Colisciences

Colisciences répond à quatre grandes ambitions générales :

• *Culturelles et patrimoniales :*

Il s'agit de collecter, numériser et mettre à disposition un grand corpus des ouvrages et articles de biologistes et naturalistes du 19e siècle, en langue française (près de 6 000 pages sont déjà offertes en mode texte et en fac-similé) : Claude Bernard, Armand de Quatrefages, Isidore et Etienne Geoffroy-Saint- Hilaire, George Romanes, etc.

• *Intellectuelles :*

Le choix de ces textes permet de retracer, en partie, une « histoire des idées », à savoir le développement durant cette période, d'une science moderne du vivant.

• *Scientifiques au sens de la modélisation sémantique :*

L'architecture de ce site traduit les réflexions de l'équipe centrées sur la problématique de l'hypertexte. Par des liens hypertextuels, le lecteur peut, à partir du texte, accéder à : 1) un glossaire des termes scientifiques et techniques, 2) un répertoire des notions, 3) un index des noms de personnes et des ouvrages cités dans chaque texte. Des parcours de lecture, surtout, lui sont proposés grâce à l'établissement de liens hypertextuels exprimant les relations sémantiques et conceptuelles que les notions entretiennent entre elles au travers des textes.

• *Cognitives et pédagogiques:*

Une de nos problématiques centrales est celle de la lecture et de la navigation dans une double perspective : 1) Les modalités de la lecture vont-elles radicalement changer avec le support électronique ? 2) Réciproquement, comment spécifier ces nouvelles conditions de l'offre de lecture pour l'apprentissage ? Comment se construit le sens dans un hypertexte électronique (*dans* et *à partir de*) ?

De ces objectifs résultent plusieurs axes de travail :

- la constitution de ce corpus (collection?) raisonné et annoté, permettant la mise en valeur scientifique et patrimoniale de certains états de la pensée (au sens d'« histoire des idées ») au sein d'un domaine – la biologie – constamment traversé par des controverses et des innovations méthodologiques et conceptuelles ;

- la prise en compte de la spécificité de l'hypertextualisation quand il s'agit de tester sa pertinence comme outil pour aborder les questions de la « navigation» dans une masse documentaire particulièrement étendue, profuse et diverse ;

- le questionnement sur les processus d'acquisition de connaissances via ce dispositif particulier.

Concrètement, ont été mises en ligne environ 6 000 pages empruntées à des éditions originales des textes de Claude Bernard, Ludwig Büchner, Étienne et Isidore Geoffroy Saint-Hilaire, Armand de Quatrefages, G.J. Romanes, Oscar Schmidt. Par la suite, nous nous donnonspour but de proposer, toujours en accès libre, une édition complète des livres de Claude Bernard (une quinzaine de titres), plus une sélection de ses articles, soit environ 5 000 pages supplémentaires.

Chaque auteur est présenté par : une courte notice biographique, une bibliographie *de* et *sur* l'auteur, une présentation ou réflexion sur ses travaux et leur portée.

Ce corpus sera entièrement inter-relié, via les *notions clés*, en tant qu'elles sont des «moteurs sémantiques » permettant de suivre les usages et les transformations du vocabulaire scientifique et philosophique (à vocation cognitive), au sein des différentes représentations du vivant (telles que, par exemple : vie, méthode expérimentale, physiologie, raisonnement, etc.) Ce ne sont donc pas seulement des textes « bruts » qui sont ainsi rendus disponibles, mais des parcours d'exploration et de lecture qui sont proposés au travers de dispositifs de navigation permettant idéalement : (i) de trouver de la façon la plus économique en termes de temps et de « charge mentale » les informations souhaitées ou recherchées ; (ii) de « saisir » les idées contenues dans ces textes par d'autres moyens que les lectures linéaires impliquées par le dispositif « livre» habituel.

6.3 Colisciences : un outil pour la prise en compte de l'« histoire des idées »

Le site CoLiSciences met donc à disposition un ensemble structuré de textes-sources et de textes-commentaires.

Les textes sont dotés d'une « aide à la lecture », classique dans sa forme et dans son usage, mais impérative quand il s'agit de permettre au lecteur non-expert un accès « cognitif » qui ne soit pas rebutant, risque constant en raison de la difficulté même des lexiques engagés par de tels domaines de connaissance. Ainsi, nous réalisons un *glossaire* des termes scientifiques et techniques, glossaire auquel l'internaute peut recourir en cours de lecture. Le lecteur peut, grâce à de courtes définitions, saisir le sens usuel et/ou circonstanciel de tel ou tel terme exigeant un éclairage spécifique. De même est offert pour chaque ouvrage et pour l'ensemble du corpus, un *dictionnaire des savants cités*, constituant un véritable panorama de la vie scientifique d'alors. Ce qui est donc privilégié ici, c'est la possibilité d'une lecture le plus possible circonscrite à l'intérieur du site, sans la nécessité fastidieuse de recourir à des dictionnaires externes.

Les *notions*, quant à elles, sont des termes clés qui condensent la nature problématique des différentes parties d'un texte-source. Outre leur rôle particulier dans l'optique d'une hypertextualisation du corpus qui s'appuie sur elles pour en appréhender la trame conceptuelle, ces notions font l'objet d'articles rédigés par l'équipe « Hypertextes et textualité électronique », en vue d'expliciter ces notions, en tant qu'elles signalent des moments spécifiques ou permanents de l'institution d'un domaine ou d'un questionnement théorique.

Dans la perspective d'une lecture de la portée sémantique du corpus en question, on distingue donc : (i) au niveau lexical, le glossaire et (ii) au niveau des idées, les « notions » constituées en noyaux générateurs de significations plurivoques. Une des premières « leçons » que l'on peut tirer de l'examen d'un corpus scientifique portant sur des auteurs et des sous-domaines variés – mais concourant à l'élaboration d'un domaine scientifique de grande ampleur et tout de même unifié – est bien de montrer que l'abord d'une science, surtout dans une perspective historique, passe par l'exploration des champs sémantiques qu'exhibe le corpus.

Le domaine notionnel

« Tout énoncé est porteur d'une "orientation" déterminée du fait d'une certaine mise en relation qu'il opère entre différents "repères" linguistiques renvoyant à des acteurs, des états, des processus, des situations, des domaines. Ce qui importe donc […], c'est de travailler sur ces types de mises en relation (thématisation, prédication, modalités) grâce auxquels, à chaque

fois, des auteurs vont “tisser” un jeu structuré de références repérant des domaines et des significations. »[46]

6.4 Le mode d’emploi de l’hypertexte CoLiSciences

A la page d’accueil du site, on va trouver cinq entrées :

- présentation du site
- acteurs et soutiens
- accès corpus
- le projet colisciences
- mode d’emploi

Explications :

« présentation du site » résume les quatre grandes ambitions du programme de recherche Colisciences telles qu’exposées précédemment.

« acteurs et soutiens » présente les membres de l’équipe Hypertextes en charge du projet, nos partenaires à titre occasionnel ou permanent et les soutiens institutionnels et financiers dont le projet CoLiSciences a bénéficié ou continue de bénéficier.

« le projet CoLiSciences » : on trouvera là un menu permettant d’accéder à :

• présentation du projet : les différents développements du projet CoLiSciences et ses extensions.

• réflexions sur l'hypertexte : une dimension importante des recherches de l'équipe porte sur l'exploration des formes de l'hypertextualité ; en témoignent les textes et articles ici présentés.

• réflexions sur la numérisation : ici on trouvera les enseignements que nous avons tirés pour numériser les textes et ce qu'il faut entendre aujourd'hui par processus de numérisation (techniques de balisage, DTD, XML, etc.).

• réflexions sur les rapports entre langage et cognition : quels modèles d'analyse sémantique et cognitive sont sous-jacents aux traitements dont les textes ont fait l'objet en vue de leur balisage ? Plusieurs synthèses résument ici nos choix théoriques et méthodologiques.

• projets en réponse aux appels d'offre : au cours du développement de CoLiSciences, plusieurs projets ont été rédigés en réponse avec succès à des appels d'offre ; ils témoignent des évolutions de nos objectifs et de nos inspirations.

« accès corpus » : il s’agit là du cœur du site : on y accède aux textes constitutifs du corpus sous trois entrées pour chaque ouvrage : fac similé, texte, notions et relations sémantiques dans le texte.

Dès l'ouverture de « l'accès corpus » en page d'accueil, quatre entrées s'offrent au lecteur : les auteurs, les disciplines, les domaines, les notions.

Si on clique sur :

• les *auteurs* : on y trouve la liste de tous les ouvrages du corpus rangés par noms d'auteur ;

• les *notions* (cf. leur définition dans le texte « Typologie des notions et relations » sous l'entrée le *projet colisciences*) : en cliquant sur l'onglet *notions,* on va se voir offrir le menu déroulant de la liste des notions et la possibilité de rechercher chacune d'elles comme notion principale en relation avec une notion secondaire ou réciproquement en tant que notion secondaire associée à telle ou telle notion principale en même temps qu'apparaissent à l'écran, si on le souhaite, les textes des paragraphes concernés (sélection de parcours notionnels) ;

[46] Vignaux, G., *Discours, Acteur du monde*, Paris-Gap, Ophrys, 1988.

• les *domaines* : ceux à l'intérieur desquels on peut ranger les ouvrages selon les types d'« objets » dont ils traitent ; exemples : l'évolution, la glycogenèse ;

• les *disciplines* : celles sous lesquelles on peut ranger un ou plusieurs ouvrages selon le contenu et l'ancrage disciplinaire ; exemple : la physiologie cellulaire, l'anatomie.

6.5 Où aller et dans quel ordre :

Sont ici proposés quelques schémas de parcours ; il en existe bien sûr d'autres.

En cliquant sur le titre d'un ouvrage dans la liste des ouvrages par auteur, le texte de cet ouvrage apparaît à l'écran selon trois types d'entrées :

• l'entrée *fac similé* : tout le texte est disponible en mode image, c'est-à-dire l'image de l'édition originale de l'ouvrage ou de l'article ; l'entrée fac similé et l'entrée texte sont synchronisées ;

• l'entrée *texte* : il s'agit du texte disponible en mode texte et découpé en paragraphes ; les termes techniques et scientifiques surlignés sont définis dans un glossaire

• l'entrée *notions et relations* qui offre deux opportunités de recherche : découvrir les notions et relations sémantiques présentes dans chaque paragraphe et dégager une sélection de parcours notionnels, telle qu'explicitée précédemment ; on peut aussi afficher tous les paragraphes où telle ou telle notion apparaît.

• Premier exemple de parcours : une « plongée » dans les textes :

En choisissant l'entrée *notions et relations*, le lecteur aura accès au texte organisé paragraphe par paragraphe, qu'il pourra ainsi faire défiler d'un paragraphe au suivant, après avoir sélectionné une notion dans la liste proposée. Au dessous de chaque paragraphe, seront affichées en tableau les notions et relations sémantiques présentes dans ce paragraphe.

Au niveau des ressources complémentaires offertes par le site, le lecteur peut encore accéder au sommaire de l'ouvrage, aux chapitres et sous-chapitres, aux illustrations.

Il peut surtout accéder au *dictionnaire des savants* mentionnés par les auteurs, chacun doté d'une biographie succincte, à un dictionnaire des termes scientifiques et techniques surlignés dans les textes, et à un dictionnaire des notions recensées.

Tel ou tel mot peut en dernier lieu être retrouvé dans son entourage immédiat, en allant à l'entrée *recherche dans le texte.*

• Deuxième exemple de parcours : naviguer pour explorer

Une entrée dans le site peut se faire selon un mode exploratoire, c'est-à-dire en allant en divers points sans stratégie prédéfinie sinon au hasard, uniquement pour «se faire une idée ». On peut ainsi :

• aller voir par curiosité les fac similé après être entré dans un ouvrage sous un nom d'auteur, et apprécier les qualités formelles des éditions originales,

• on peut de là choisir de regarder le sommaire de l'ouvrage (entrée sommaires) et repérer une notion : celle de *vie* par exemple,

• on peut de là aller au paragraphe concerné par cette partie du sommaire (entrée notions) et demander à voir avec quelles autres notions, cette notion de vie est reliée et sous quelles formes de relations,

• on peut s'intéresser aux explicitations et commentaires de cette notion (dictionnaire des notions),

• et on peut recommencer…

• Troisième exemple de parcours : naviguer pour apprendre

Deux objectifs sont ici possibles, déterminant deux types de stratégies :

• « se faire idée » des éléments de conjoncture scientifique historique (histoire des idées) que représentent ces textes : aller donc voir (au niveau *accès corpus*) la liste des auteurs, celle des ouvrages collectés, les domaines et les disciplines concernés ; en même temps, toujours au même niveau, on pourra trouver des biographies des auteurs et des synthèses sous forme de textes-commentaires sur l'œuvre et les idées et les apports de tel ou tel d'entre eux.

• apprendre à travers l'architecture du site ce que peut être un hypertexte : pour ce faire, aller d'abord en page d'accueil à l'entrée présentation du site, on y trouvera une introduction exposant les ambitions du programme et une explication des cinq entrées disponibles en page d'accueil ; si on va ensuite à l'entrée le *projet CoLiSciences,* on pourra consulter une série de textes de réflexions sur l'hypertexte, la numérisation et les modèles d'analyse conçus et adoptés ; enfin, l'entrée *mode d'emploi* aidera à comprendre comment utiliser le site, le parcourir et en bénéficier.

• Quatrième exemple de parcours : collecter et analyser

Un chercheur, un étudiant, un curieux pourra souhaiter collecter des informations sur telle ou telle notion pour comprendre comment la biologie de cette époque l'abordait et la traitait.

On ira donc pour ce faire, au niveau de l'accès corpus, à l'entrée notions dans le corpus, et à l'intérieur de celle-ci, à l'entrée recherche de parcours notionnels. On pourra ce faisant, à profondeur variable, collecter dans différents textes du corpus, les passages (paragraphes) où cette notion est traitée, et les autres notions avec lesquelles elle entre en relation pour constituer un réseau d'idées.

Cette collecte de différents fragments textuels peut se constituer comme « dossier » épistémologique et sémantique, ou «collection » relative à un réseau de notions, qu'il sera toujours possible de compléter, mais déjà propice à un travail de chercheur ou d'étudiant voire à une réflexion érudite.

7. En guise de conclusion

Le rapprochement entre la problématique de la collection et celle de l'hypertexte n'est ici pas fortuit. En effet, de même que l'hypertexte, la collection n'est jamais ou rarement un produit terminé ; comme lui, elle demeure un espace d'expression, de mémoire et de communication en constante évolution. On retrouve ce faisant, ces réflexions que Buffon inscrit dans son « Premier discours », qui sert d'envoi à son *Histoire générale et particulière* : « Le premier obstacle qui se présente dans l'étude de l'Histoire naturelle vient de cette grande multitude d'objets ; mais la variété de ces mêmes objets, et la difficulté de rassembler les productions diverses des différents climats, forment un autre obstacle à l'avancement de nos connaissances, qui paraît invincible et qu'en effet le travail seul ne peut surmonter ; ce n'est qu'à force de temps, de soins, de dépenses, et souvent par des hasards heureux, qu'on peut se procurer des individus bien conservés de chaque espèce d'animaux, de plantes ou de minéraux, et former une collection bien rangée de tous les ouvrages de la Nature. »[47]

De même encore que dans le cas de l'hypertexte, la problématique de la collection revient de plus en plus à apprendre à travailler avec des savoirs possédant un taux d'expansion jamais vu jusqu'alors, portés par des documents souvent labiles, sans la stabilité que, dans une large mesure, l'imprimé conférait autrefois.

Les technologies de constitution et de valorisation des collections doivent alors être vues comme un continuum auquel doivent participer différents spécialistes, prenant en compte les questions de choix des collections, de numérisation de ces collections et de leur promotion.

[47] Buffon, G.L. de, *Histoire naturelle,* textes choisis et préfacés par Jean Varloot, Folio, Gallimard, 1984.

C’est dire, comme on l’a vu dans ce texte, que les questions d’indexation, de catégorisation, et de validation sémantique et sociale sont essentielles. Tout prendre en compte excède les forces et la durée d’une Action spécifique.

Néanmoins, ces questions retrouvent sens dans un Réseau thématique comme « Document et contenu » qui nous semble devoir dans l’avenir, se reconfigurer comme programme de recherche englobant dès lors, ces perspectives.

Références consultées par Georges Vignaux

Foucault, M., *Les mots et les choses*, Paris, Gallimard, 1966.
Lecointre, G., Le Guyader, H., *Classification phylogénétique du vivant*, Paris, Belin, 2001.
Mullaney, S., "Strange things, gross terms, curious customs : the rehearsal of culture in the late Renaissance", *Représentations*, 1983, 3.
Pomian, K., *Collectionneurs, amateurs et curieux, Paris, Venise : XVIe-XVIIIe siècles*, Paris, Gallimard, 1987.
Dictionnaire des notions philosophiques, sous la direction de Sylvain Auroux, Paris, PUF, 1990.
Dubois, Christine, "L'œuvre-collection : de la taxinomie du visible à l'utopie", *Parachute*, 1989, n° 54.
Grasskamp, Walter, "Les artistes et les autres collectionneurs", Museum by Artists, Toronto, Art Metropole, 1983.
http://collections.ic.gc.ca/parcours/laboratoire/livre/creation.html
« Les papillons de Laurent Schwartz », *La Recherche*, mars 2003, 72-78.
Brian, Eric, « L’ancêtre de l’hypertexte », La Grande Encyclopédie, *Les Cahiers de Science et Vie*, 1998, n° 47, 30-38.
Darnton, Robert, *L’Aventure de l’Encyclopédie. Un best-seller au siècle des Lumières*. Paris, Perrin, 1982.
Martin, Henri-Jean, et Chartier, Roger (sous la direction de), *Histoire de l’édition française*. Tome II : *Le livre triomphant. 1660-1830*. Paris, Promodis, 1984.
Moureau, François, *Le roman vrai de l’Encyclopédie*. Paris, Gallimard, Découvertes, 1990.
Proust, Jacques, *Diderot et l’Encyclopédie*, 2e édition. Paris, Armand Colin, 1967.
Wilson, Arthur M., *Diderot, sa vie et son œuvre*, traduit de l’anglais par Gilles Chahine, Annette Lorenceau, Anne Villelaur. Paris, Laffont/Ramsay, 1985.
Yeo, Richard, « Modèles d’outre-Manche », La Grande Encyclopédie, *Les Cahiers de Science et Vie*, 1998, n° 47, 24-27.
Arnauld, A. et Nicole, P. *La Logique ou l'art de penser.* Ed. critique par H. E. Brekle. Stuttgart/Bad Cannstatt: F. Froman Verlag, 1970.
Auroux, S. "Le paradigme lockien et la philosophie du langage". *Revue Internationale de Philosophie*, 1988 , 42, n° 165, 133-149.
Auroux, S. "Le rationalisme et l'analyse linguistique". *Dialogue*, 1989, XXVIII, 203-233.
Barsalou, L. W., Sewell, D. R. "Contrasting the Representation of Scripts and Categories. *Journal of Memory and Language, 24,* 646-665, 1985.
Berlin, B. *Ethnobiological Classification. In*: E.Rosch and B.B. Lloyd (Eds). *Cognition and Categorization.* Hillsdale: L. Erlbaum, 1978.
Chomsky, N., *Cartesian Linguistics.* New York: Harper and Row, 1966.
Culioli, A., "Notes du séminaire de D.E.A, 1983-1984". Paris: Université Paris 7, Unité de Formation et Recherche en Linguistique, 1985.
Culioli, A. *Pour une linguistique de l’énonciation*, Paris-Gap, Ophrys, 1990.
Dominicy, M. *La naissance de la grammaire moderne.* Liège, Mardaga, 1984.
Dubois, D., Mazet, C., Fleury, D. (1988). "Catégorisation et interprétation de scènes visuelles: le cas de l'environnement urbain et routier". *Psychologie française, n° spécial:" La psychologie de l'environnement en France"*.
Dubois, D. (éditeur*), Sémantique et Cognition*, Paris, CNRS, 1991.
Duchesneau, F. *L'empirisme de Locke.* La Haye: Nijhoff., 1973.
Fodor, J. *The Modularity of Mind.* Cambridge: MIT Press, 1983.
Grize, J.B. *Logique et Langage.* Paris-Gap, Ophrys, 1990.
Jakobson, R. *Essais de linguistique générale.* Paris: Minuit, 1963.
Langacker, R. (1987). *Foundations of Cognitive Grammar. I. Theoretical Prerequisites.* Stanford: Stanford University Press, 1987.
Le Ny, J. F. (1989). *Science cognitive et compréhension du langage*. Paris: Presses Universitaires de France, 1989.
Locke, J. (1690). *An Essay concerning Human Understanding.* Edition Yolton à Londres: 1955.
Locke, J. *De la conduite de l’entendement*, trad. Par Yves Michaud, Paris, Vrin, 1975.
Peirce, C. S. (1931-1935 et 1958). *Collected Papers*, vol. I à vol. VI, éd. par C. Hartshorme et P. Weiss, Harvard Univ. Press, 1931-1935; vol. VII et VIII, ed. par W. Burks, Harvard Univ. Press, 1958.
Quillian, M. R. *Semantic Memory. In:* Minsky (Ed). *Semantic Information Processing.* Cambridge, Mass.: MIT Press, 1968.
Rosch, E., Lloyd, B.B. *Cognition and Categorization.* Hillsdale, New Jersey: Lawrence Erlbaum, 1978.
Tversky, B., Hemenway, K. (1983). "Categories of Environmental Scenes". *Cognitive Psychology*, 15, 121-149.
Vendryes, J. (1923). *Le langage. Introduction linguistique à l'histoire.* Réédition (1968). Paris: Editions Albin Michel.
Vignaux, G. *Le Discours, acteur du monde. Enonciation, argumentation et cognition.* Paris-Gap: Editions Ophrys, 1988.
Vignaux, G. *Le démon du classement*, Paris, Seuil, 1999.
Vignaux, G. *Du signe au virtuel*, Paris, Seuil, 2003.
Balpe, Jean Pierre. Hyperdocuments Hypertextes Hypermedias. *Paris: Eyrolles, 1990.*
Baron Georges-Louis, Bruillard Éric (2001), " Didactique de l’informatique ?", *Revue Française de Pédagogie*, n° 135, 163-172.
Bates, M. (1989), « The design of browsing and berrypicking techniques for online search interface », Online Review, 1989, 13, 5, 407-423.
Benoit, J., Fayol, M. (1989), « La catégorisation des types de textes », Pratiques, *62, 71-85. Beeman, W., Anderson, K., Bader, G., Larkin, J., McClard, A., McQuillan, P. & Shields, M. (1987). Hypertext and pluralism: from lineal to nonlineal thinking,* Hypertext'87 Papers, *1-20.*
Beeman, W., Anderson, K., Bader, G., Larkin, J., McClard, A., McQuillan, P. & Shields, M. (1988). Intermedia: A case Study of Innovation in Higher Education. Providence, RI, Brown University, IRIS.
Bosak Jon, et Bray Tim, « XML and the Second-Generation Web », *Scientific American*, May 1999.
Boutell, Thomas. World Wide Web Frequently Asked Questions. Version mise à jour fréquemment :

<URL: http://sunsite.unc.edu/boutell/faq/ www_faq.html.>
Bruillard Éric (1997), *Les machines à enseigner.* Éditions Hermès, Paris, 320 p.
Bruillard Éric, Grandbastien Monique (eds.) (2001). Éducation et informatique. Hommage à Martial Vivet, Sciences et Techniques éducatives, vol. 7, n° 1, Hermès Science, 300 p.
Bruillard Éric, de La Passardière Brigitte et Baron Georges-Louis (eds.) (1998). Le livre électronique. Sciences et Techniques Éducatives, *vol. 5, n° 4, Hermès Science.*
Bruillard Éric, de La Passardière Brigitte. (1998). « Fonctionnalités hypertextuelles dans les environnements d'apprentissage », in Tricot A. et Rouet J.-F. (dir.), Les hypermédias, approches cognitives et ergonomiques, *Hermès, Paris, p. 95-122 (correspondant à un numéro spécial de la revue* Hypertextes et Hypermédias*).*
Bush, V, (1945). « As we may think », Atlantic Monthly, *July 1945, 176(1).*
Collins, A., Brown, J.S., & Newman, S.E., (1988). Cognitive apprenticeship: teaching the craft of reading, writing, and mathematics. In L.B. Resnick (Ed.), Cognition and Instruction: Issues and Agendas. *Hillsdale, NJ, Erlbaum.*
Davis, Paul J. The World Wide Web : Industry Analysis. *New York: J.P. Morgan Securities Inc., May 5, 1995. <URL: http://www.jpmorgan.com/MarketDataInd/Research/WebReport/page1.html>*
Duffy, T.M., Knuth, R.A. (1990). Hypermedia and instruction: where is the match?, in Jonassen, D. and Mandl, H. (eds). Designing Hypermedia for Learning, *Heidelberg, Springer-Verlag.*
Englebart, D. (1984). Authorship provisions in AUGMENT, *IEEE Comp-Con Proceedings, Spring.*
Guédon, Jean Claude. "Les bibliothèques à l'heure des réseaux télématiques planétaires." Argus 23, no 3 (September1994): 9-14.
Idem, Internet. Le monde en réseau. *Paris, Gallimard, 1996.*
Horn, R., (1989). Mapping Hypertext, *The Lexington Institute, Ma. 289p.*
Jonassen, David H. Hypertext/Hypermedia. *Englewood Cliffs, New Jersey: Educational Technology Publications,1989.*
Jonassen, D., Mandl, H., (1990). Designing Hypermedia for Learning, *Springer-Verlag, Series F, vol. 67.*
Jones, B.F., Pierce, J., & Hunter, B. (1988). Teaching students to construct graphic representations, Educational Leadership, 20-25.
Laufer, Roger, and Domenico Scavetta. Texte, hypertexte, hypermédia. *Que sais-je?, 2629. Paris: Presses Universitaires de France, 1992.*
Lebrave, Jean-Louis, « Réflexions sur l'hypertexte », Culture et recherche, *1995, n°51, p.6.*
Le Crosnier, Hervé. "L'hypertexte en réseau : repenser la bibliothèque." Bulletin des bibliothèques de France, *40, no 2 (1995): 23-31.*
McKnight, C., Dillon, A., Richardson, J., (1991). Hypertext in Context, *Cambridge University Press.*
Maignien, Yannick. "La bibliothèque virtuelle : ou de l'ars memoria à Xanadu.", Bulletin des bibliothèques de France, *40, no 2 (1995): 8-17.*
Nelson, T. (1970). No More Teachers' Dirty Looks, Computer Decisions, September.
Nelson, T. (1981). Literary Machines, *Swathmore, Pa.*
Palmer, J. Duffy, T, Mehlenbacher, B. (1990). A System for Aiding Designers of Online Help, Lotus:acm SIGCHI.
Rada, Roy. Hypertext : from text to expertext. *London: McGraw Hill, 1991.*
Rhéaume, J. (1991). « Hypermédias et stratégies pédagogiques », in de la Passardière, B. et Baron, G.-L., éd. Hypermédias et apprentissages, *Paris, MASI, INRP.*
Rhéaume, J. (1993). L'enseignement des hypermédias pédagogiques, in Baudé., J., éd., Deuxièmes journées francophones Hypermédias et apprentissages , Paris, EPI.
Shaw, Debora. "Libraries of the Future: Glimpses of a Networked, Distributed, Collaborative, Hyper, Virtual World." Libri, 44, no 3 (1994): 206-223.
Shneiderman, B., Kearsley, G., (1989). Hypertext Hands-on!, *Reading, Ma., Addison-Wesley Publishing, 165p.*
Stern, David. "Expert Systems: HTML, the WWW, and the Librarian." Computers in Libraries 15, no 4 (April 1995): 56-58.
Yankelovich, N., (1987). Creating hypermedia material for english students, Sigcue-Outlook, 20.

Bush et le Memex Brève présentation de Vannevar Bush :
http://www.iath.virginia.edu/elab/hfl0034.html
"As we may think" :
http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm www.theatlantic.com/unbound/flashbks/computer/bushf.htm
ou www.isg.sfu.ca/~duchier/misc/vbush/
Une image du Memex : http://www.kerryr.net/pioneers/memex_pic.htm bootstrap.org/engelbart/index.jsp
Ted Nelson et l'hypertexte Brève présentation : http://www.iath.virginia.edu/elab/hfl0155.html
Projet Xanadu : http://xanadu.com
Sa home page : http://www.sfc.keio.ac.jp/~ted/
Ted Nelson et le mot hypertexte en 1965 : http://iberia.vassar.edu/~mijoyce/Ted_sed.html
Memex et au-delà : http://www.cs.brown.edu/memex/