HISTORIQUE

La décision de constituer un tel système d'information historique est, compte tenu de la rapidité de l'évolution technologique, très ancienne, puisqu'elle remonte au début des années 1970 (autant dire l'âge de pierre, lorsqu'il est question d'informatique). Cette ancienneté explique que ce travail ait longtemps été subordonné à des contraintes, dont certaines ont aujourd'hui disparu, qui expliquent sa structure et, dans une certaine mesure, ses lacunes. Ces contraintes ont été essentiellement des contraintes techniques, et des contraintes de temps.

Dés les débuts de ce travail, en effet, il était envisagé de constituer d'une part une base prosopographique et une base texte. Nous avons commencé à travailler à Paris I, au Centre de Calcul, sous la bienveillante autorité d'Edouard Valensi, avec Jean-Paul Trystram et Xavier Debanne (le concepteur du système BDP1, que j'ai surtout utilisé dans ses versions successives, BDP3 et BDP4) au niveau prosopographique, et avec François Hucher et Jacques Mondelli au niveau lexical ; il s'agissait là de développer un logiciel nouveau, adapté à la situation orthographique des textes médiévaux. Ce logiciel, ALINE, a tourné et bien tourné jusqu'à la fin des années 1970[1]. A cette époque, toutefois, nos orientations se sont modifiées : il était impossible de continuer à travailler avec François Hucher et Jacques Mondelli, trop occupés par leurs activités professionnelles, et des progrès rapides étaient d'ailleurs enregistrés dans le domaine de la lexicologie informatisée : c'était d'abord le logiciel belge JEUDEMO, puis, l'apparition des logiciels du Centre de Lexicologie Politique de l'Ecole Normale Supérieure de Saint-Cloud. Nous avons donc abandonné, peu à peu et à regrets, ALINE. Tout en poursuivant l'enregistrement des textes politiques anglais que nous souhaitions étudier sous un format compatible avec les logiciels de Saint-Cloud et plus précisément avec PISTES, un logiciel crée par Pierre Muller et dont j'ai pu avoir connaissance grâce à Philippe Dautrey, j'ai commencé à envisager sérieusement de rédiger un dictionnaire prosopographique en langage naturel bien tout en étant exploitable automatiquement et j'ai donc concentré mon travail sur la conception et la réalisation de PROSOP.

Le problème du prosopographe est en effet le suivant : s'il veut pouvoir exploiter commodément les informations dont il dispose, il lui faut utiliser un logiciel de base de données, où les informations seront plus ou moins codées[2]. Avant 1980, autant dire qu'elles étaient codées numériquement, ce qui rendait inconsultables les informations, et donc incontrôlables les sources des tableaux de fréquence et de pourcentages sur lesquels allait s'appuyer la réflexion historique. Précisons une bonne fois pour toutes qu'"incontrôlables" ne signifie pas qu'un quelconque soupçon pèse sur la qualité ou la véracité de ces données. Pouvoir contrôler des

[1] Voir J. Ph. Genet, "Un programme de traitement automatique des textes : ALINE" (en collaboration avec F. Hucher, J. Mondelli, et E. Valensi), *Bulletin du Centre d'Analyse du Discours de l'Université de Lille III*, 1974, pp.96-121 ; "Ordinateur, Lexique, Contexte", in L. Fossier, A. Vauchez et C. Violante, *Informatique et Histoire Médiévale*, Rome, 1977, pp.299-317, et "Un exemple de programme de traitement de texte: ALINE", *Le Médiéviste et l'Ordinateur*, I, 1979, pp.4-9.

[2] Cf. J. Ph. Genet, "Histoire Sociale et Ordinateur", in L. Fossier, A. Vauchez et C. Violante, *Informatique et Histoire Médiévale*, Rome, 1977, pp.231-237.

données, cela veut dire d'une part en savoir assez sur elles pour être capable de modifier les codages, et d'autre part pouvoir disposer de l'information pour refaire le traitement (par exemple, avec d'autres méthodes statistiques) : c'est donc la condition nécessaire pour se trouver en situation expérimentale, au sens que nos collègues scientifiques donnent à ce terme ; une expérience scientifique doit pouvoir être reproduite, et chacune de ses phases doit pouvoir être testée. J'ajouterai que, dans le cas qui nous occupe, l'impossibilité matérielle de retourner à tous les textes contenus dans la base ou de remonter aux sources initiales des informations, jointe à l'abondance des publications et des recherches nouvelles, fait que la base doit nécessairement être mise à jour à intervalle régulier et améliorée sans cesse ; les résultats obtenus à partir d'elles doivent donc être tenus pour "provisoires en permanence", et la mise à jour des données et des informations est l'un des paramètres normaux de ce système, qui est donc un système ouvert, conçu pour n'être jamais "fini", à tous les sens du terme. Il y a là une différence fondamentale avec un travail imprimé classique, et même avec ce que nous pourrions appeler les métasources de la première génération, qui étaient constituées de bases de données codées très difficiles à mettre à jour et à réexploiter en permanence en tenant compte des changements apportés.

Revenons à l'élaboration de la base. Plutôt donc que de coder, puis, le codage et les traitements faits, passer à la rédaction du dictionnaire, il paraissait plus raisonnable de rédiger un dictionnaire en langage naturel, parfaitement accessible à un lecteur, et, à partir de ce dictionnaire, de procéder, si possible automatiquement, à un codage. Tout en travaillant à la conception d'un logiciel capable d'effectuer ce travail, j'ai donc commencé à rédiger le dictionnaire, à partir des fiches manuelles de dépouillement que j'avais réalisées à la British Library, à Londres. La première mise en forme a été faite avec le logiciel APPLEWRITER, sur la machine dont je disposais alors, un APPLE doté de 48 K de mémoire. Le logiciel n'autorisait que quatre caractères accentués ! Quelques fautes d'accent, remontant à cette première saisie, traînent encore çà et là dans HISPOL ... Une première version du dictionnaire (portant sur 745 historiens) a ainsi été réalisée, mais il fallait la transférer au CIRCE, le centre de calcul du CNRS à Orsay, pour envisager un traitement. Le logiciel de saisie retenu était SPF (devenu SPFPC pour les applications sur PC), un logiciel qui avait à mes yeux deux avantages. Tout d'abord, comme ce n'est pas un traitement de texte, le texte qu'il contient est parfaitement standard, donc transférable sans aucun problème d'un support à un autre : il l'est d'ailleurs toujours, même si j'utilise, selon les machines, divers logiciels (et surtout WORD) pour les impressions. Ensuite, il permettait de travailler simultanément sur PC et sur le gros ordinateur du CIRCE à Orsay (sur place, ou via les terminaux du Centre de Calcul de Paris I et de l'EHESS).

La programmation du logiciel, baptisé PROSOP (puisque son objectif premier est le traitement prosopographique) permettant de traiter le dictionnaire a commencé au début des années 1980, grâce à Michael Hainsworth, alors qu'il était directeur du LISH (Laboratoire d'Informatique des Sciences de l'Homme), mais la première tentative sur Macintosh a tourné court. Nous sommes alors repassés sur PC, avec Saleh Dabjen, puis, dans le cadre de l'U.R.A. 1004, le programme a été entièrement repensé et réécrit par Manuel Ornato qui y a travaillé jusqu'en 1995[3]. Il a cependant fallu se rendre à l'évidence : les moyens mis à notre disposition

[3] On suivra les étapes de ce travail dans les nombreuses présentations que j'ai été amené à faire au fil des colloques et des congrès, notamment ceux organisés dans le cadre des activités de l'*International Association for History and Computing* : "The PROSOP system and the problem of Standardization and Exchange of Data", in F. Hausmann, R. Hartel, I.H. Kropac et P.Becker, *Datennetze fur die Historischen Wissenschaften?*, Graz, 1987, p.82-87 ; "The

par le CNRS étaient décidément trop faibles pour mener à bien cet ambitieux projet, et le logiciel PROSOP n'a pas été achevé ; il tourne et je m'en sers, mais il reste difficile à manier et pourvu de fonctions par trop limitées pour permettre une exploitation complète du dictionnaire. Pourtant, l'exploitation des informations contenues dans le dictionnaire reste possible, grâce notamment au transfert des données, sous forme plus ou moins automatique, du format SPFPC vers le format DBase, grâce à des routines écrites en TURBOPASCAL, une procédure à laquelle m'a initié Arlette Faugères. Aussi ai-je réutilisé ce format dans le cadre d'un nouveau projet prosopographique, entrepris dans le cadre de mon enseignement d'histoire et d'informatique à la Sorbonne, celui du dictionnaire des étudiants et des maîtres de l'Université de Paris[4].

Afin de permettre la meilleure utilisation possible du dictionnaire, l'introduction qui suit présentera brièvement les principes généraux qui ont conduit à sa réalisation (*I. Une métasource pour une histoire culturelle quantitative*). Les règles d'écriture dans le dictionnaire ont beaucoup évolué dans le temps mais elles sont décrites ailleurs, dans la mesure où leur connaissance reste indispensable à une utilisation optimale du dictionnaire HISPOL. Le logiciel PROSOP est décrit (*II. Le logiciel PROSOP*), en distinguant entre les parties programmées et celles qu'il est envisagé d'écrire un jour. Ce sont ensuite les autres composantes éléments du système d'information qui seront ensuite décrites, HP, HPNUM, OPUS et MEDITEXT etc. (*III. Un système d'information*). En annexe, figure la liste des abréviations utilisées dans le dictionnaire, et notamment dans la bibliographie.

I. Une métasource pour une histoire culturelle quantitative

Les données qui sont rassemblées ici font partie d'une base de données heuristique, c'est-à-dire d'une base conçue spécialement pour tester une hypothèse scientifique. Tel est l'objectif, même si avec le temps la base a grossi de façon un peu démesurée : il n'a nullement été dans mon intention de me transformer en bibliographe ou en documentaliste, si utile et si honorable que soit ce type d'activité ; c'est au fur et à mesure du travail que je me suis rendu compte que, pour pouvoir tester l'hypothèse, il me fallait étendre l'emprise de l'architecture de la base. Elle ne constitue donc pas une descendance des nombreux répertoires qui, depuis Bale, Tanner et Leland, sont l'une des gloires de l'historiographie anglaise et dont Richard Sharpe[5] vient il y a quelques mois de nous offrir un dernier exemple. Du moins est-ce le dernier ouvrage de ce genre que j'ai pu utiliser ; mais déjà se profile à l'horizon celui de Joan Greatrex[6]. La base, du strict point de vue documentaire, présente un certain nombre de lacunes fâcheuses), notamment

---

PROSOP system", in P. Denley et D. Hopkin, *History and Computing*, Manchester, 1987, p.191-198 ; 53. "L'informatique au service de la prosopographie: PROSOP", in *Mélanges de l'Ecole Française de Rome (Moyen Age Temps Modernes)*, C, 1988, p.247-263.

[4] J.Ph. Genet, "Le répertoire informatisé provisoire de l'Université de Paris", in M. Cocaud, éd., *Histoire et Informatique. Base de données, recherche documentaire multimédia. Actes du 1er Congrès de l'Association française pour l'Histoire et Informatique, Rennes, 1994*, Rennes, 1995, 109-126 et "A Provisional Computerized Prosopographical register of Medieval Paris University", in P. Denley, éd., *Computing Techniques and the History of Universities,* Saint-Katharinen, 1996, pp.1-14.

[5] Richard Sharpe, *A Handlist of the Latin Writers of Great Britain and Ireland before 1540*, (Publications of the Journal of Medieval Latin, I), Turnhout, 1997 ; *Additions and Corrections*, 2001.

[6] J.G. Greatrex, *Biographical Register of the English Cathedral Priories of the Province of Canterbury, c.1066-1540*, Oxford, 1997 ; Richard Sharpe ayant eu accès à cet ouvrage avant sa publication, quelques unes des informations qu'il contient ont cependant pu être utilisées.

au niveau de la mise à jour[7] : j'en ai fait deux complètes, mais étant donné mes occupations d'enseignement, il faut compter au moins deux ans pour "suivre" toute la base, délai aggravé par le fait que seule les bibliothèques anglaises permettent de faire ce travail (encore que les fonds de la British Library ne soient pas au-dessus de tout reproche pour les publications américaines).

Donc, tester une hypothèse de départ. Elle est détaillée ailleurs, et je ne ferai ici que la résumer brièvement : le développement d'une société politique dans le cadre de l'Etat moderne s'accompagne (terme volontairement neutre) d'une transformation du contenu, des acteurs et des modalités de la communication qui comprend la conquête des champs de l'historique (l'ensemble des paramètres qui conditionnent la perception et la représentation du temps, de l'espace et de la mémoire dans une société donnée) et du politique (l'ensemble des concepts, des pratiques et des ajustements qui permettent à une société de régler les tensions et les conflits survenus dans la sphère du public - sans laquelle il n'est pas de politique possible autre que celle des rapports de domination directs). Une fois donc définies et explorées la configuration des deux champs en relation avec tous les autres, l'étude de l'hypothèse implique l'étude de l'un des groupes privilégiés d'acteurs, les "auteurs", et celle de leur production textuelle ; comme il n'était pas question d'aborder celle-ci dans sa totalité, on s'est contenté d'une bibliographie générale (liste des oeuvres avec leurs principales caractéristiques de diffusion et de circulation), et, pour le contenu des textes afin de pouvoir atteindre la langue du politique, les textes politiques de la période médiévale.

1° Une base de donnée heuristique

Une base de données heuristique, quand il s'agit d'une base prosopographique, est une base construite à partir d'une double série d'hypothèses fondées sur une réflexion théorique. Pour schématiser à l'extrême, je dirais qu'il y a une théorie générale qui justifie la définition de la population étudiée, et des hypothèses locales, qui correspondent à chacune des variables. Dans le cas présent, la population est définie et délimitée grâce à une réflexion théorique centrée sur la production du texte, et non, ce qui est en général plus simple, sur l'appartenance des individus à telle institution ou à tel groupe ; le "groupe" ainsi constitué est inséparable de cette théorie, il est impensable sans elle, et c'est d'ailleurs en cela même que la base est une base heuristique, et non une base documentaire. Le choix qui a été fait ici s'inspire des travaux de Pierre Bourdieu sur les champs. Le champ peut être décrit schématiquement comme un système d'institutions, ce terme étant pris en son sens le plus large : cela va des institutions au sens propre, par exemple les institutions académiques, au marché, qui agit comme un champ de forces plus ou moins autonome pour inciter à la production des biens culturels, ici, en l'occurrence des textes, établir leur valeur, et assurer leur diffusion (ou l'empêcher). La période qui commence en 1300 offre une excellente situation de test pour cette méthode, puisqu'il existe d'un côté une série d'institutions qui contrôlent et dans une certaine mesure assurent elles-mêmes la production des textes, et d'un autre côté, grâce aux progrès de l'éducation et à la rapide expansion des pratiques de l'écriture et de la lecture personnelles, un marché. On a pu ainsi définir plusieurs champs, qui sont décrits dans le chapitre III des Prolégomènes qui accompagnent ce dictionnaire :

---

[7] Un outil indispensable de mise à jour est la bibliographie publiée annuellement dans les *Publications of the Modern Language Association of America*.

1- le champ du religieux ;

2-5 Les quatre champs liés aux Facultés des Arts et de Médecine :

-le champ des disciplines de la philosophie
-le champ des disciplines du texte et du discours : grammaire, rhétorique, philologie.
-le champ des disciplines scientifiques dépendant de la Faculté des Arts médiévales
-le champ du médical.

6 -le champ du "littéraire".

7 -le champ du juridique

8 -le champ des textes "pratiques"

9 -le champ de la musique.

Enfin, restent deux champs, considérés comme des champs composites : le champs du politique, et le champ de l'historique, qui sont ceux qui nous occupent plus précisément.

Chacun des textes est réparti dans les champs ainsi définis ; deux autres attributs sont également mentionnés, en observant des classifications qui sont, comme je l'ai dit plus haut, issues de théories "locales", variable par variable. Les deux systèmes de classification portent respectivement sur les caractères de langue et de genre d'une part, et sur les caractères de contenu de l'autre : ils sont exprimés dans les colonnes 3 et 4 de la ligne titre de chaque oeuvre. Je n'ai fait ce travail que pour les deux champs composites que j'ai étudiés, mais il va de soi qu'il devrait être fait pour chacun des champs définis ci-dessus. Le résultat n'est d'ailleurs pas parfait, car la langue est minorée, puisque cette catégorie distinctive pourrait s'appliquer à toutes les formes de texte qui ont été énumérées, ce qui n'a pas été fait ; en outre, en dehors du latin et de l'anglais, il n'y a qu'une catégorie "langue vernaculaire" prévue, quelle qu'elle soit. Cela est assez correct du point de vue théorique que j'ai adopté, encore que le statut de chaque langue soit différent et varie avec le temps : si le français du début du XIV^{e} siècle équivaut à peu près à l'anglais du XVI^{e} siècle au moins pour les élites aristocratiques, il a une toute autre signification au XVI^{e} siècle ; la même réflexion pourrait s'appliquer au gallois, au gaélique, ou à l'italien et à l'espagnol du XVI^{e} siècle, langues éminemment catholiques. Et cela peut gêner un utilisateur de la base qui cherche des textes en flamand ou en italien, et ne veut pas s'encombrer de textes en français. Ces deux jeux de classification doivent être établis champs par champs.

Cet ensemble (champs, définition d'une classification des formes et des contenus des unités textuelles) forme donc la théorie générale, qui permet de délimiter la population. Les nomenclatures de chacune des variables sont quant à elles les théories locales ; je ne les détaillerai pas ici (elles le sont plus loin, dans la présentation de la base HISPOL), d'autant que les nomenclatures varient pour chacune des bases dérivées du dictionnaire bio-bibliographique (HP, HPNUM) ; on en trouvera le détail ci-dessous puisqu'elles sont indispensables à la compréhension du dictionnaire et des traitements qui en sont issus ; mais leurs particularités sont décrites dans le cadre de ces traitements[8], car elles conditionnent de façon trés étroite la lecture et la compréhension des résultats.

[8] Voir *infra*, pp.

2° Une métasource

Une base de données prosopographique et bio-bibliographique doit remplir, même si elle n'est qu'heuristique et non documentaire, une certaine fonction documentaire, puisqu'elle doit permettre à l'historien utilisateur de ce travail (ou à tout autre lecteur ou utilisateur) de pouvoir contrôler les processus de comptage et de mesure. Il faut disposer en clair de la source des mesures et des comptages qui sont effectués à partir de la métasource. Le présent dictionnaire fait en effet partie de ce que j'ai défini comme une métasource[9]. La métasource est l'ensemble structuré des informations mises en formes et transmises à l'ordinateur et traitées par lui. Tout autant sinon plus que le traitement lui-même, la constitution d'une métasource distingue le travail informatisé de celui qui ne l'est pas : elle implique, sur le plan scientifique, un saut qualitatif évident, dans la mesure où toute l'architecture d'informations et de données qui sous-tend l'interprétation historique est désormais visible, alors que dans un travail classique, elle n'est apparente qu'à travers un double système de citations et de références. La métasource, même si elle est constituée d'images ou de sons, ou d'un mélange de ces éléments, qu'elle ait la forme d'une base de données ou qu'elle soit constituée de textes au sens le plus classique et habituel du terme, est en elle-même un texte d'un type particulier, pourvu d'une structure créée par son "auteur", et qui, dans tous les cas, diffère du texte de la source, même si l'on a affaire à un enregistrement en *full-text* de la source originelle. Il doit être bien clair que les traitements statistiques (ou autres) n'ont d'autre fondement que la métasource, et que c'est de la qualité de celle-ci que dépend avant tout la qualité des résultats de l'ensemble du travail.

Or, la réalisation de la métasource pose un certain nombre de problèmes. Tout d'abord, il y a le problème du statut critique de la métasource. A la limite, s'il ne s'agissait que de contrôler les informations qui servent de base au traitement informatique (statistique ou autre), le problème serait relativement simple : il s'agirait de fournir simplement les informations en clair, pour étayer les comptages. Mais étant donné l'importance proprement scientifique de la métasource que nous venons de souligner, l'idéal serait de faire de la métasource un répertoire critique des informations utilisées. Autrement dit, toute information devrait être pourvue de la mention de sa source et d'une discussion de la fiabilité de cette source ; et tout texte enregistré devrait être pourvu de ses variantes et de tout l'appareil d'annotation que contient une édition critique. Or, jusqu'ici, à ma connaissance, aucune base heuristique ne présente ce genre de qualité, qui supposerait pour chaque champs au moins une indication de source. Si l'on se représente classiquement une base de données comme une matrice, on devrait donc disposer d'une matrice double, une matrice contenant les données et, associée, une matrice contenant les références ou les "preuves". D'une certaine façon, et me semble-t-il très légitimement, la relative défiance des historiens à l'égard des travaux informatisés s'explique et se justifie par cette carence.

Car pour le moment du moins, carence il y a. En tant que métasource, le dictionnaire souffre de trois déficiences majeures, dont la description s'impose ; le statut scientifique de ce dictionnaire dépend en effet non pas de sa précision fallacieuse, mais de la précision avec laquelle il est possible de cerner son degré de fiabilité. Il n'est d'ailleurs pas fait pour être "terminé" et figé sous la forme immuable du livre imprimé, mais pour être perpétuellement

---

[9] Essentiellement dans deux articles, J.Ph. Genet, "Histoire, Informatique, Mesure", *Histoire et Mesure*, I (1), 1986, 7-18 et "Source, métasource, texte, histoire", in F. Bocchi et P. Denley, éd., *Storia & Multimedia, Atti del Settimo Congresso Internazionale, Association for History and Computing*, Bologna, 1994, pp.3-17.

réinscriptible et perfectible sous la forme dynamique du texte électronique. Ses déficiences principales sont au nombre de trois :

- l'essentiel de l'apport informatif le plus important est relativement peu visible et il n'est pas totalement fiable. Ce sont les caractères marginaux qui classent chaque oeuvre, et qui précisent la position de l'"auteur" par rapport à l'unité textuelle (terme préférable à texte : à un même texte peuvent correspondre plusieurs unités textuelles : un texte peut avoir un auteur au sens habituel, être traduit par un autre individu, commenté ou abrégé par un autre : il apparaîtra sous la forme de trois unités textuelles). Au jour où j'écris ces lignes, il y a 12537 unités textuelles dans la base. Outre le fait que j'ai pu commettre des erreurs d'analyse, je n'ai pas vu toutes ces unités ; cela est d'ailleurs impossible, dans la mesure où de nombreux textes ont disparu et ne sont connus que par les titres ou les descriptions laissées par les bibliographes et les catalographes, et que d'autres sont ce que j'appelle des "textes virtuels"[10] ; par ailleurs, le fonds sur lequel j'ai travaillé est celui de la British Library ; si j'ai aussi examiné pas mal de manuscrits à Oxford et quelques uns à Paris et à Cambridge, je n'ai examiné que les imprimés de Londres ; or, si riche soit-elle, la British Library ne possède pas, il s'en faut, tous ces textes. Quand je n'ai pas vu le texte, je l'ai classé à partir de connaissances indirectes ou, si elles étaient inexistantes, à partir de son titre. Il y a donc là une source d'erreur.
- le dictionnaire est à la fois exhaustif et incomplet. Il est exhaustif en ce sens que j'ai essayé de repérer tous les producteurs d'unités textuelles dans les deux champs que j'ai définis ; s'il en manque que je n'ai pas su débusquer, d'autres les trouveront, là n'est pas le problème. Le problème est : peut-on interpréter les résultats à partir l'examen de deux champs seulement ? Il me semble que oui : parce qu'il s'agit de deux champs transversaux, qui recoupent tous les autres (et donc apportent des informations sur eux) ; mais à condition de ne pas oublier que l'image que nous avons est celle de ces deux champs seulement, et non celle de l'ensemble de la production textuelle d'une époque et d'une société. Notamment, ces 12537 unités ne peuvent en aucun cas se voir attribuer le statut statistique d'échantillon : ils ne sont nullement représentatifs, statistiquement parlant. Enfin, une quantité non négligeable (mais très variable selon les genres) d'unités textuelles ne figure pas dans le dictionnaire bio-bibliographique, et pour cause : ce sont les textes anonymes, que l'on ne retrouvera que dans la base de textes (MEDITEXT : voir *infra*).
- enfin, au niveau de la présentation, la précision est insuffisante. Prenons l'une de mes fiches, la première par exemple. Le lieu de naissance de George Abott, les principaux faits de sa carrière sont indiqués. Pour que le dictionnaire reste lisible (et pour ne pas passer trop de temps à sa réalisation ...), ces informations ne sont pas à chaque fois pourvues de la mention de la source d'origine (voir la discussion ci-dessus). L'avancement des technologies informatiques rend un tel développement tout à fait envisageable désormais, grâce aux hypertextes, qui permettent de créer des liens entre une information visible et une autre, qui n'apparaît à l'écran que lorsque besoin est, par activation d'un bouton.

[10] Voir J.P. Genet, "Matrices, genres, champs: une approche sur le long terme", in A. Vaillant, éd., *Mesure(s) du livre. Colloque organisé par la Bibliothèque Nationale et la société des études romantiques*, Bibliothèque Nationale, Paris, 1992, p.57-74, notamment p.60.

Le dictionnaire bio-bibliographique, malgré ses déficiences, a cependant sous sa forme électronique, deux avantages essentiels qui me paraissent conditionner toute la démarche entreprise ici : il est indéfiniment perfectible, je l'ai déjà souligné. Et surtout il a une double valeur de preuve (mais cela n'est pas une nouveauté) et d'instrument expérimental : les informations étant en clair, il est possible d'en repartir pour tester une théorie nouvelle, différente de celle qui est formulée ici. L'un des reproches formulés à l'égard de ce genre de travail est que le résultat est surdéterminé à la fois par la théorie et la méthode. Cela est évident, et c'est d'ailleurs une démarche normale dans la plupart des sciences : une théorie permet d'atteindre un résultat après vérification expérimentale, jusqu'à ce qu'une nouvelle théorie aboutisse à la formulation de nouvelles hypothèses, à leur tour testées empiriquement pour produire des résultats qui infirment ou confortent ceux précédemment obtenus. Mais aussi bien le statut épistémologique des sources que le poids matériel du travail qu'elles impliquent a jusqu'ici pratiquement interdit une telle démarche aux historiens : les outils électroniques, tel ce dictionnaire et les différents fichiers (notamment MEDITEXT) qui l'accompagnent, l'autorisent et la facilitent désormais. Surtout, ils permettent l'introduction de la mesure et des méthodes statistiques là où elles n'avaient pas cours auparavant.

3° La visée quantitative

L'un des objectifs majeurs de ce travail est d'introduire la mesure dans l'histoire culturelle, dont elle reste en général absente[11]. Pour quoi faire ? Pour dénombrer d'abord, pour résumer ensuite, pour interpréter enfin. Mais l'idée même de dénombrer n'est-elle pas absurde ? Quelle espèce d'équivalence peut-il bien exister entre l'unité textuelle *Macbeth* et l'unité textuelle *Sermones ad populum* de l'un des nombreux Carmes auxquels John Bale a attribué un tel titre ? D'équivalence littéraire, il n'y en a - et il ne peut y en avoir - aucune. Il y a par contre, et c'est ce qui m'intéresse, deux indications de même ordre d'une relation entre un individu (l'auteur) et des destinataires (éventuellement, un public). Cette relation peut être comptée, et elle peut être qualifiée : *Macbeth* est une pièce de théâtre, destinée à être jouée dans un théâtre, et son sujet est emprunté à l'histoire (mais version légendaire) des Iles Britanniques ; les *Sermones* sont des sermons, prononcés en français ou en anglais (question de date) dans une église ... La diffusion et l'impact superficiel du texte peut à son tour être qualifié : en comptant le nombre des manuscrits, des éditions, des représentations, par exemple, ou celui des citations - directes ou indirectes -d'une oeuvre donnée dans les autres unités textuelles du corpus. Toutes ces qualifications peuvent être dénombrées, et nous disposons donc à travers cet ensemble de dénombrements d'un instrument de mesure des relations établies au sein d'une société donnée à travers les textes. C'est tout, c'est peu et c'est beaucoup à la fois. C'est peu, sur le plan littéraire ; mais c'est beaucoup, et c'est ce qui m'importe, pour la connaissance du champ. Les textes ouvrent à leur tour l'accès à une autre séries de mesures, par l'intermédiaire de la lexicologie quantitative ; mais ceci est une autre histoire.

Les méthodes statistiques employées pour résumer l'information sont simples. Ce sont des dénombrements, des tris croisés et des pourcentages. Je n'ai pas eu recours aux méthodes

[11] Des efforts considérables sont faits aujourd'hui : voir par exemple A. Vaillant, éd., *Mesure(s) du livre. Colloque organisé par la Bibliothèque Nationale et la société des études romantiques*, Bibliothèque Nationale, Paris, 1992 ; et de nombreux articles sur ce thème ont été publiés dans *Histoire & Mesure* ; voir J.P. Genet, "Pour l'informatisation des dictionnaires biographiques : une expérience", *Histoire & Mesure*, I (2), 1986, p.99-110. et "La mesure et les champs culturels", *Histoire & Mesure,* II (1), 1987, p.137-153, à propos du remarquable livre d'Alain Viala.

couramment utilisées par les sociologues en général mais aussi par des historiens américains : les analyses de corrélation, de régression et de causalité. Ce n'est pas le lieu d'en discuter ici la pertinence ; il me semble simplement qu'il y aurait en quelque sorte une différence de statut entre le type de résultat auquel conduisent ces méthodes et la nature même de mes informations, leur instabilité et leur fragilité mêmes. Les tableaux de tris croisés, sans élaboration supplémentaire, me paraissent en tous cas suffisants pour représenter et pour résumer.

Reste l'interprétation. Dans ce domaine, l'analyse factorielle des correspondances m'a paru apporter exactement ce dont j'avais besoin : on trouvera ailleurs une description détaillée de la méthode et de la façon dont les processus d'interprétation fonctionnent à partir du résultat des calculs. Disons ici qu'elle est à la fois une méthode d'investigation et une méthode de description. Une méthode d'investigation parce que, dans une matrice quelle qu'elle soit, elle signale, en les hiérarchisant par ordre d'importance, des liaisons, des proximités qu'il faut ensuite expliquer, comprendre, en retournant aux données. C'est donc un guide pour la lecture de données structurées mais entre lesquelles les liaisons sont si nombreuses qu'elles ne peuvent être toutes explorées. Et c'est en même temps une méthode de description parce que chaque liaison qui est signalée est une liaison entre deux profils, profils d'individu ou/et profils de variable : à chaque pas, on tient ainsi compte de la totalité des relations existant à l'intérieur de la métasource.

4° De la base de données au système d'information

Le dictionnaire HISPOL a été conçu comme une base de données autonome (les trois bases HP, HPNUM et OPUS n'en sont que des produits dérivés) ; MEDITEXT représente, de son côté, un corpus de textes médiévaux tout aussi autonome. A l'origine, il n'y avait pas de liaison entre ces deux métasources, si ce n'est qu'elles s'intéressaient, vaguement, au même terrain historique. Les changements rapides de la technologie m'ont progressivement amené à réviser certaines des options prises initialement. La rapidité et les capacités accrues de la transmission de l'information, la miniaturisation des données stockées sur support électronique, l'augmentation rapide de la taille et la baisse vertigineuse du prix des mémoires de masse font qu'il est désormais possible de disposer de quantités pratiquement illimitées d'informations ou de "données". Il est donc apparu possible de relier ces deux ensemble initialement séparés. C'est vers ces développements que s'est aujourd'hui infléchi le travail sur ces bases et sur les logiciels qui permettent (ou plus exactement devraient permettre) de les exploiter.

Si l'on pense à l'ensemble de la production textuelle médiévale et moderne, il y a deux façons d'atteindre le texte: soit mener la recherche par des critères de recherche internes, c'est-à-dire par le contenu du texte, qu'il s'agisse de son titre, d'une citation, voire d'un simple mot, soit la mener par des critères de recherches externes, que l'on parte d'informations chronologiques, géographiques, biographiques ou ayant trait aux systèmes de production et de diffusion des textes. C'est dans cette seconde optique, jusqu'ici tout à fait absente du "marché" logiciel, qu'il m'a semblé que PROSOP avait sa place ; cela suppose donc d'intégrer les textes dans le système (cela est par exemple possible pour les auteurs actifs dans le domaine politique en Angleterre, ou un corpus de plusieurs centaines de textes est dores et déjà disponible), et éventuellement d'introduire les outils permettant de les traiter, ce qui n'exclut d'ailleurs nullement le fait que ces traitements puissent être faits par interfaçage avec un logiciel de traitement lexical, type LEXIS, HYPERBASE, TACT, OCP ou "St-Cloud" (cf. PISTES, LEXICO, ou SYNCHIEF) : de tous ces logiciels, c'est HYPERBASE tel qu'il est développé par Etienne Brunet qui paraît pour le

moment le plus proche de ce dont nous aurions besoin, mais une nouvelle version de PISTES est actuellement en cours de préparation.

Les dictionnaires (ceci vaut en effet non seulement pour HISPOL, mais aussi pour les fichiers PARIS, sur les maîtres et étudiants de l'Université de Paris au Moyen Age) ont jusqu'à présent été réalisés, nous l'avons dit, avec le logiciel SPF, un éditeur standard (non un traitement de texte). Elles sont totalement standardisées [SDF] et transférables d'un logiciel à l'autre et d'une machine à l'autre sans aucune difficulté. Les dictionnaires sont rédigés en langage naturel, mais néanmoins structurés en fonction d'un protocole, qui est détaillé ci-dessous.

Le logiciel PROSOP qui traite ces fichiers est donc capable de gérer toutes les informations sur les "auteurs", entendu ici au sens large d'acteurs producteurs dans l'un des champs de production que nous avons étudiés, mais, à condition d'avoir constitué les dictionnaires bio-bibliographiques *ad hoc*, on peut en étendre l'application à tous les champs et à tous les types d'acteurs (par exemple, scribes, libraires, propriétaires de livres, etc...). Les dictionnaires comportent aussi tous les renseignements élémentaires sur les textes eux-mêmes (titre, incipit, date de rédaction), sur leur diffusion, sur les manuscrits et les éditions. Les paragraphes de PROSOP n'ayant pas de taille limite, le dictionnaire peut même inclure des textes : c'est le cas dans le dictionnaire HISPOL qui comprend un grand nombre de textes (ainsi tous les discours parlementaires du Moyen Age, transcrits à partir de l'édition des *Rotuli Parliamentorum*). L'idéal serait cependant d'établir un système de liaison sous PROSOP entre, d'une part, les dictionnaires bio-bibliographiques (par exemple HISPOL) et, d'autre part, une banque de textes (par exemple MEDITEXT).

Il faudrait toutefois pouvoir disposer de l'information contenue dans HISPOL (par exemple, mais cela doit pouvoir s'appliquer à tous les dictionnaires de même structure, à commencer par PARIS, le dictionnaire des maîtres et des étudiants parisiens déjà évoqué) sous trois formes : d'abord, dans la forme d'un dictionnaire biographique (chaque fiche correspondant à un auteur) qui est la sienne actuellement ; ensuite, sous la forme d'un dictionnaire bibliographique (chaque fiche correspondant à une unité textuelle), ce qui est le cas d'OPUS aujourd'hui : mais OPUS ne contient que des informations restreintes et sous une forme codée, ce qui est insuffisant ; par ailleurs, ce dictionnaire bibliographique devra comporter aussi les oeuvres anonymes, pour le moment exclues d'HISPOL. Enfin, un troisième dictionnaire, celui des supports (chaque fiche correspondant à un support lié à une ou à plusieurs unité-textes), qui n'existe pas du tout aujourd'hui, s'avère nécessaire, et contiendrait des données sur chacun des manuscrits et chacune des éditions anciennes mentionnées succinctement dans HISPOL (ou tout autre dictionnaire de même niveau), l'information étant reclassée par manuscrits et par éditions, et enrichie de descriptions (notamment codicologiques) et d'indications sur la date et les circonstances de copie du manuscrit, sur l'impression des livres et sur leurs propriétaires successifs.

A partir de là, il devrait être possible d'établir une liaison avec le corpus MEDITEXT ou tout autre ensemble de textes correspondant au contenu du dictionnaire bio-bibliographique dont on part. Nous n'aurions plus, dés lors, affaire à une base ou à un ensemble de bases de données (situation présente avec les bases HISPOL, HP, HPNUM, OPUS et MEDITEXT, toutes décrites en détail ci-dessous), mais à un véritable SIH, système d'information historique au sens que les géographes donnent à l'expression lorsque l'on parle de SIG. Un tel système d'information permettrait donc de rechercher et d'afficher tout texte contenu dans MEDITEXT, grâce à des

requêtes fondées sur des informations figurant dans les dictionnaires de niveau HISPOL (dictionnaire biographique, dictionnaire bibliographique, dictionnaire des supports. Prenons HISPOL comme métaphore d'un dictionnaire bio-bibliographique de tous les "acteurs" des champs de production textuels des sociétés médiévales et modernes (jusqu'au XVIIe siècle en tous cas), et MEDITEX comme la référence de la grande bibliothèque numérisée du futur (mais d'un futur très proche, puisque la plupart des grandes bibliothèques du monde, à commencer par la Bibliothèque Nationale, travaillent en ce moment même à des projets de ce genre) ; le texte pourrait d'ailleurs être stocké sous plusieurs formes (par exemple, chacun des manuscrits pour un texte médiéval, en même temps que chaque édition ancienne et la ou les éditions critiques modernes). On le voit, les perspectives qui s'ouvrent ici sont immenses, voire démesurées ...

Elles le sont en tout cas par rapport à ce que, étant donné les moyens modestes dont nous avons disposé, nous avons pu faire. Sur le plan du support des données, nous sommes entrain de travailler à la mise sur CD-ROM d'HISPOL et de MEDITEXT, avant d'aborder le passage à INTERNET. Dans les deux cas, le problème du support - gravage sur CD-ROM et transfert en texte HTML - ne présente pas de difficultés majeures. Il en va tout autrement des "moteurs" qui permettraient, dans l'un et l'autre cas, de manier commodément l'information. Grâce à des étudiants de l'I.U.T. de Vélizy et à leurs professeurs, une première tentative de mise dur CD-ROM d'HISPOL est en cours en ce moment[12] ; nous verrons avec intérêt quels résultats peuvent ainsi être obtenus. Nous rejoignons là le problème du logiciel PROSOP qui m'a jusqu'ici permis d'exploiter mes fichiers et dont j'aurai plus loin l'occasion de signaler l'insuffisance (en même temps que les perspectives de développement qu'il offre encore). Mais à partir du moment où l'on parle de système d'information, les exigences sont encore plus poussées, et ceci dans trois domaines au moins.

Premier de ces domaines, celui des mises à jour : une information nouvelle introduite dans l'un des fichiers HP??.DIC devrait pouvoir être répercutée automatiquement et instantanément sur tous les fichiers qu'elle affecte ; par exemple, si un nouveau historique texte est attribué à notre premier auteur, George Abbott, il faut que le nombre d'oeuvres historiques qui figure dans la variable c50 de HP.DBF soit incrémenté de 1, que la nouvelle valeur de la variable TOTAL de cette même base soit calculé, et que si besoin est la classe de la variable QUA dans la base HPNUM soit modifiée, en même temps qu'une ligne nouvelle apparaît dans OPUS, porteuse de toutes les informations requises par la structure de la base et dont je dispose sur ce texte (tous ces noms de bases et de variables sont explicités dans les pages qui suivent). Deuxième domaine, celui de l'interactivité : il faut à la fois que les résultats des requêtes faites au système viennent enrichir celui-ci, et que d'autre part le système soit assez ouvert pour que les suggestions et informations venues de l'extérieur puissent également l'enrichir. Troisième domaine, enfin, celui de la sécurité : si le système est ouvert sur l'extérieur, par exemple par l'intermédiaire d'Internet, il faut néanmoins qu'il soit protégé contre les modifications

[12] Ce travail est accompli dans le cadre de leur travail de fin d'étude par un groupe d'étudiants de l'I.U.T. de Versailles-Saint Quentin en Yvelines (Vélizy), Charlotte Echardour, Frédéric Mayer, Jerôme Moles, Jérôme Suadeau et Karine Ventura. La base est transformée en un lieu physique, que l'utilisateur visite ; les thèmes sont "affectés à des quartiers spécifiques, à l'intérieur d'une ville, représentative des fondations de la base de données". La base restera connectable sur internet pour réactualiser les données. Les liens hypertexte et le moteur de recherche fonctionneront à partir de trois éléments : le vocabulaire cumulé des codes des quatre bases (HISPOL, HP, HPNUM, OPUS), le vocabulaire des textes de MEDITEXT, et les index de noms de personnes, de lieux et d'institutions générés par PROSOP.

inconsidérées, la malveillance et les erreurs de maniement. Aucun de ces problèmes n'a de solution simple ; il faut atteindre à un niveau de modélisation et de structuration très supérieur à celui auquel nous sommes généralement parvenus : mais on sait généralement les résoudre lorsqu'il s'agit de gros systèmes commerciaux, préparés en tenant compte de tous ces paramètres. Il est toutefois évident qu'ils ne pourront être surmontés avec une programmation au coup par coup, étalée sur dix ans, sans l'appui et la collaboration de spécialistes de haut niveau.

Etablir des liens entre tous les fichiers au sein de véritables "systèmes d'information et les faire fonctionner dans des conditions satisfaisantes représente le nouveau défi que doivent affronter les informaticiens en cette fin de siècle. Espérons que les historiens pourront en profiter pour mettre sur pied des "systèmes d'information historiques". Qu'ils y parviennent ou non, il ne leur en restera pas moins , comme aux autres spécialistes des sciences humaines, à structurer et à reformuler les données que le passé nous a léguées de façon à ce que les bases de données et systèmes d'information deviennent de bons instruments de savoir et de recherche, capables d'égaler en fiabilité et en richesses les rayons de nos vieilles bibliothèques d'érudition. HISPOL, MEDITEXT et les bases qui en sont dérivées, que nous allons maintenant décrire dans le détail, se veulent un premier pas dans cette direction.

II. Le logiciel PROSOP

1°. Principes généraux

Pour la prosopographie, le médiéviste (l'antiquisant et le moderniste sont d'ailleurs logés à la même enseigne) travaillant sur des sources de qualité irrégulière, est confronté à l'extrême inégalité des volumes d'information dont il dispose selon les individus. Les bases construites en utilisant des systèmes de gestion [SGBD] du commerce, bien que ceux-ci se soient considérablement améliorés au fil des ans (par exemple 4D ou Visual-dBase), sont des bases qui doivent respecter strictement des normes de format, puisque pour chaque variable un espace prédéfini est accordé à priori. Dans notre cas, cet espace risque d'être soit sous-utilisé, lorsque les informations manquent, soit insuffisant, lorsqu'elles abondent. Il paraît plus raisonnable d'enregistrer les informations en format standard, mais en les structurant de telle sorte que les informations soient facilement retrouvées et interprétées, tout en profitant de l'enregistrement pour réaliser une indexation aussi complète que possible, quitte à transférer automatiquement une extraction de la base en format dBase (ou autre) pour profiter des fonctions de tri et de comptage de ce type de logiciel. C'est ainsi que nous avons été amené à concevoir un logiciel spécifique, PROSOP, destiné à gérer des dictionnaires en langage naturel, qui sont à la fois prosopographiques, donc destinés à permettre sur l'ensemble de la population des traitements à la fois sériels et quantitatifs, et bio-bibliographiques, c'est-à-dire qu'ils doivent permettre de maîtriser l'ensemble de la documentation concernant un individu (sa production de textes d'une part, et les textes qui le concernent). Depuis peu, et à cause de l’apparition de médias nouveaux, et notamment du C.D. Rom, il est envisagé que PROSOP gère aussi les textes produits par les auteurs.

Le logiciel, auquel ont travaillé successivement Michael Hainsworth, Saleh Dabjen et maintenant Manuel Ornato qui est pour l'essentiel responsable de l'actuelle configuration de PROSOP, assume dans la version en cours de réalisation actuellement cinq fonctions dont trois seulement sont complètement programmées à l'heure actuelle :

1° sélectionner et compiler les fichiers.
2° Opérer des sélections dans les fichiers.
3° Indexer les fichiers.
4°Transformer les fichiers en fichiers sous d'autres formats
5° Aider l'utilisateur à

En outre, pour pouvoir opérer dans de bonnes considérations de sécurité, PROSOP impose une phase préliminaire de compilation. Nous présenterons ces différents stades avant de présenter les développements potentiels du système.

b. Les cinq fonctions principales de PROSOP

Fonction 1 : sélection des fichiers et compilation:

Le premier écran de PROSOP permet de définir la base sur laquelle on va travailler. Une routine permet de récupérer, n'importe où sur le disque dur (par scanning du directory du DOS) les fichiers de dictionnaire sur lesquels on désire travailler. Une fois la sélection opérée, c'est-à-dire une fois que l'ensemble des fichiers choisis a été constitué en un groupe, la première action qui s'impose est de compiler les fichiers. On clique sur "Action", et parmi les opérations proposées on choisit compilation. La compilation est une opération un peu lente, mais elle ne requiert aucune participation du chercheur. Toutefois, à l'issue de la compilation, un diagnostic est affiché par le logiciel: il indique les fautes repérées, incohérences dans les signes d'indexation et de soulignement, signes de fin ou de début de fichier absent (i.e. paragraphes 1a et C BIBLIOGRAPHIE, 99a ou 99b). Tant que ces fautes, préjudiciables aux autres opérations, n'ont pas été corrigées, la compilation ne peut pas aboutir. Chaque fois qu'un fichier dictionnaire a été modifié, par exemple par l'adjonction de données nouvelles, la compilation doit être recommencée. Une fois tous les fichiers constituant le groupe compilé, il est possible d'entreprendre d'autres actions.

Fonction 2 : la sélection

La sélection s'opère par un langage de commande qui a été mis au point par Manuel Ornato. Il repose sur deux clauses, la clause TROUVER et la clause EXTRAIRE. La clause TROUVER comporte une localisation, à savoir l'indication du ou des paragraphes et des sous-paragraphes dans lesquels on va faire la recherche ; vient ensuite l'indication de ce qui va fonctionner comme critère de sélection, un mot (chaîne de caractère quelconque) ou une date (y compris un intervalle) ; il peut y avoir plusieurs critères de sélection, unis par les opérateurs booléens ET et OU. La clause EXTRAIRE comporte l'indication du ou des paragraphes que l'on veut extraire.

Un certain nombre d'options sont d'autre part disponibles. Dans la clause TROUVER, * indique que l'on ne recherche ni une date ni une chaîne de caractère, mais seulement que l'on teste la présence ou non du paragraphe ; ! indique que l'on arrête la lecture si l'on rencontre le signe ?, indiquant que l'information est incertaine ; et l'expression ENTETE fait référence à la première ligne du chapitre, qui peut se trouver dans un paragraphe différent des paragraphes ou sous-paragraphes indiqués dans la localisation (c'est souvent le cas quand on fait une recherche

sur la production textuelle). Dans la clause EXTRAIRE, * indique que l'on veut extraire toute la fiche.

Exemple: (DANS: 7.m TROUVER: * EXTRAIRE: 7.m) OU (DANS: 7.n TROUVER: * EXTRAIRE: 7.n) EXTRAIRE : 2

Le logiciel va donc chercher pour toutes les fiches celles pour lesquelles il existe un paragraphe 7.m (marchands etc.), puis celles pour lesquelles il existe un paragraphe 7.n (médecins) ; les deux paragraphes sont extraits, et l'on extrait en outre le paragraphe 2 (origine géographique). Le fichier résultant, qui reçoit l'extension SEL comprendra pour chacun des individus sélectionnés:

-le label (nom, description principale, dates)
-le paragraphe 2a
-le paragraphe 7m
(le paragraphe 7n

Le nouveau fichier est en fait un nouveau fichier dictionnaire sur lequel on peut aussi opérer des actions PROSOP, par exemple une nouvelle sélection (sur les dates par exemple) ou une indexation.

Fonction 3 : l'indexation.

Autre action possible, l'indexation. Elle peut être faite à tout moment, sur n'importe quel fichier ou groupe de fichier compilé. Etant donné la taille et le caractère disparate du contenu des fichiers prosopographiques de base, il est déconseillé cependant de faire cette opération sur ces fichiers: ainsi, il ne sert à rien de faire une recherche sur "Paris" qui mélangerait par exemple les paragraphes 5 et 7, où l'on trouvera toutes les mentions de Paris dans le contexte des études et de l'enseignement, les paragraphes 9, où l'on trouvera toutes les mentions de Paris dans le cadre des voyages et des déplacements et d'autre part tous les sous-paragraphes d, où le renvoi à Paris dénote l'édition à Paris d'un volume avant 1700! La démarche normale est donc de faire une sélection des paragraphes ayant un contenu voisin, puis de faire sur eux une indexation. Une indexation sur 5a, 5b, 5c, 7a, 7b donnera ainsi tout ce que la base contient sur les études et l'enseignement.

L'indexation se fait en cliquant sur le type d'indexation désiré: indexation des items marqués par une étoile * (en général, les lieux ou les concepts) ; par une livre sterling £ (un événement, un lieu spécifié) ; ou compris entre deux dollars $$ (les noms de personne). On peut sortir les trois index séparément, ou ensemble. Le fichier de sortie peut-être placé dans le répertoire de son choix (par exemple WORD si on veut tout de suite faire une impression). Le format de sortie est un standard de 80 colonnes, la zone de droite étant fixe et donnant la référence de l'item indexé (numéro de la fiche, nom de la vedette, paragraphe et sous-paragraphe): la zone gauche est donc d'une longueur fixée par avance, et il est donc possible que l'item indexé soit coupé. On dispose donc d'une option qui permet de sortir la zone gauche sans coupure ; mais la zone droite n'est alors plus alignée et il faut faire la mise en page avec un traitement de texte normal.

Deux fonctions sont encore prévues, la fonction 4 (transformation des fichiers) et la fonction 5 (aide à l'utilisateur). Bien qu'elles soient indispensables a une utilisation agréable du logiciel, il a été impossible de les programmer.

Fonction 4 : la transformation des fichiers

C'est la prochaine étape importante de programmation de PROSOP. "Transformation" est ici entendu en deux sens:

- Transformation d'un fichier "PROSOP" en un fichier "X"

L'objectif de PROSOP n'est pas de tout faire. Une partie de l'exploitation des dictionnaires peut valablement être assurée par d'autres logiciels: les repérages lexicaux peuvent être faits avec WORDCRUNCH, les comptages et les tris croisés peuvent être faits avec DBase ou avec ORACLE, les éditions avec WORD ou WORDPERFECT. La transformation des fichiers est plus ou moins délicate. Pour l'heure, nous utilisons WORD 6 sous WINDOWS en utilisant les macrocommandes pour l'édition et nous disposons d'un module rudimentaire mais néanmoins utile de transfert vers DBase, mis au point il y a dix ans avec Arlette Faugères, et grâce auquel j'ai pu extraire de HISPOL deux bases, HP (prosopographique) et OPUS (bibliographiques) qui tournent bien sous DBase. Un essai de transfert vers ORACLE dans le cadre d'une thèse de l'EHESS a aussi donné des résultats encourageants, mais nous n'avons pas ORACLE à notre disposition, ni d'ailleurs de station de travail permettant de le faire tourner dans de bonnes conditions. Il s'agit pourtant là d'une étape essentielle et qui touche au coeur même de notre projet.

- Génération de fichiers PROSOP de structure différentes de celle du fichier-père

Les fichiers prosopographiques contiennent une foule d'information qui sont susceptibles d'être organisées différemment. Reprenons deux exemples que nous avons déjà évoqués dans la première partie de ce texte : un individu est un "auteur" ; il a écrit des "oeuvres" ; chaque "oeuvre" est contenue dans un ou plusieurs manuscrits.

Nous pouvons d'abord envisager de constituer un fichier oeuvre. Les codes qui désignent les oeuvres sont spécifiques, si on les lie au numéro d'auteur: chaque oeuvre est donc désigné dans la base par un numéro unique. Pour chaque oeuvre, l'on a déjà le titre, la date, le ou les manuscrits et l'incipit pour une oeuvre manuscrite, la ou les éditons anciennes jusqu'en 1700, des indications sur le statut du texte (s'il s'agit d'une édition, d'une traduction, d'un commentaire) etc..., sans parler des éléments de classification contenus dans les codes marginaux, ni des zones libres de commentaire dont la plus importante est en général à la fin du dernier paragraphe concernant l'oeuvre en question et qui peut aller jusqu'à contenir le texte lui-même *in extenso*. Il suffit alors d'ajouter le label de la fiche (contenant le numéro, le nom, les dates et un descriptif sommaire de l'individu) au descriptif de chaque oeuvre pour que l'on ait un fichier oeuvre parfaitement autonome, auquel on pourra dés lors ajouter les oeuvres anonymes, particulièrement nombreuses dans le cas de la base sur l'Université de Paris.

Maintenant, pour chaque oeuvre, on a une liste des manuscrits qui la contiennent. Un même manuscrit peut contenir plusieurs "oeuvres" ; certaines sont dans la base, d'autres non. L'idée est donc de générer un nouveau dictionnaire, mais qui sera cette fois un dictionnaire des

manuscrits: il donnera, sous le label nouveau constitué par la cote du manuscrit, le contenu du manuscrit, avec le titre, la date de composition, l'incipit de l'"oeuvre", et le label de l'"auteur" de l'oeuvre. Ce nouveau dictionnaire devient alors un dictionnaire autonome et il est alors possible de l'enrichir en complétant les informations contenues dans PROSOP, par exemple en complétant la description du manuscrit (critères physiques ; marques de propriétaires ou d'utilisateurs ; colophon etc.) et l'analyse de son contenu (en introduisant les "oeuvres" et les noms de leurs "auteurs" qui ne figurent pas dans le dictionnaire de départ), soit à partir des catalogues imprimés existant, soit à partir de l'examen des manuscrits subsistants ; je me suis livré à un travail de ce genre pour comparer la diffusion des oeuvres d'Ockham, Fitzralph, Burley et Wyclif d'après les manuscrits survivants. Le nouveau fichier peut soit être exploité de façon autonome (par PROSOP), ou utilisé conjointement avec le fichier originel comme "aide" (voir *infra*). La même opération est possible pour les éditions anciennes.

Fonction 5 : la fonction d'aide.

Deux domaines, dans PROSOP, requièrent déjà de tout évidence la mise au point rapide d'un sous-programme d'aide à l'utilisateur. C'est tout d'abord la zone des codes marginaux, dont le nombre est tel qu'il est difficile à un utilisateur même expérimenté de les connaître tous. Par exemple, en cliquant sur 50yt, l'utilisateur doit pouvoir voir apparaître dans un coin de son écran l'information en clair suivante:

50 = Champs historique.
y = édition de document.
t = cartulaire.

La même observation est valable pour la zone 99a, qui contient les références bibliographiques données sous forme résumée. Ainsi, pour GLORIEUX MT, on devra voir apparaître à la demande la référence complète, Mgr. P. GLORIEUX, *Répertoire des maîtres en théologie de Paris au XIII^e^ siècle*, Paris, 2 volumes, 1933 et 1934. Ceci ne devrait pas poser de problèmes particuliers, dans la mesure où les informations nécessaires sont dores et déjà disponibles.

c. Les développements possibles

Mais il est possible d'envisager d'aller plus loin, si l'on a à sa disposition les bases issues du module de transformation (fonction 4) décrite plus haut. On peut ainsi demander, à propos d'un des manuscrits figurant dans la liste des manuscrits d'une oeuvre que l'on étudie, des renseignements supplémentaires en interrogeant le fichier manuscrit (constitué comme dans l'exemple donné ci-dessus). On pourra ainsi savoir par exemple quelles sont les oeuvres voisines de celle que l'on étudiera dans le manuscrit en question. On peut évidemment multiplier de tels fichiers d'aide: le tout est d'avoir le temps de les constituer. Mais on peut aussi construire des relations avec des fichiers ou des bases de données extérieures à PROSOP, par exemple, pour les *incipits*, avec l'*Incipitaire* de l'I.R.H.T. diffusé par Brépols, ou avec des bases en cours de développement aujourd'hui, notamment à la British Library.

Amélioration de l'ergonomie du système

Il s'agit d'abord dans un premier temps de conserver les caractéristiques structurelles essentielles de PROSOP : c'est ue base de données dont le but est d'être un outil de travail pour

les historiens. Sa "philosophie" est de caractériser de toutes les manières possibles un individu, et, quand c'est un auteur, ses textes, de manière à pouvoir faire des entrées, des tris et des classements en fonctions de données non seulement internes aux textes, mais aussi externes. Les données sont organisées dans une (ou plusieurs) base de donnée qui est arborescente, c'est à dire que l'élément principal est l'auteur, lequel compte de nombreux attributs, dont les textes ; ces attributs comptent eux-même des attributs, et ainsi de suite. La richesse de la description permet de faire des tris, indexations, etc... sur tous les types de critères connus par la base.

Pour avoir un outil efficace pour les historiens, et qui puisse être effectivement diffusé auprès d'eux, le système doit permettre des enregistrements complexes, et il doit être capable de manipuler des relations autant que des données. Il doit être parfaitement fiable, et capable d'extraire des données facilement, de faire des tris, des classements, des demandes, de manière conviviale (*user-friendly* disent les informaticiens), ce qui n'est pas encore tout à fait le cas. On doit pouvoir introduire de nouvelles données de manière simple même si elles sont très partielles (avec les problèmes de mise à jour automatique que cela implique, dans le cas où le module de transformation fonctionnant, il existerait des fichiers dépendants. Il devra permettre d'introduire une information supplémentaire liée à toute nouvelle information: une description de l'information elle-même (qui a entré cette donnée, quand, etc...), et une référence détaillée de l'information (jusqu'à présent, les fiches ne comportent qu'une bibliographie détaillée en fin de fiche: il pourrait être utile d'avoir la possibilité d'appeler à l'écran une documentation plus précise sur les sources de telle ou telle information). Le système devra néanmoins rester rapide, et éventuellement être accessible par réseau.

Il faut reprendre la réflexion sur la structure, c'est à dire au type de représentation des connaissances, qui est souhaitable pour cette base de donnée. Il y a deux niveaux : la représentation interne au système, et la représentation accessible à l'utilisateur (historien), dite vue de la base. La démarche à suivre parait être de partir du second niveau pour définir un cahier des charges, et d'en déduire une structure de représentation. On peut supposer, en particulier si on veut pouvoir introduire des informations de nature diverses (images, son, texte, données archéologiques, etc...), que les structures pourraient être le plus décentralisées possible : entité-auteur, entité-texte, entité-faculté, etc... Les relations entre les entités pouvant être de plusieurs type (tel individu appartient à telle université, a écrit ou commenté tel ouvrage) et être elles-mêmes documentées (tel individu appartient à telle université entre telle date et telle date, avec tel grade, etc...).

Sur le plan de la représentation des connaissances, les bases de données orientées objet proposent une représentation interne des entités plus souple. Cet aspect est utile en particulier pour ce qui est de la facilité de dialogue avec le système par un non-informaticien, la simplicité d'utilisation.

Le recours à l'intelligence artificielle

Elle peut apporter plusieurs types de choses. Tout d'abord, une fiabilité accrue. Une approche de type *système expert* peut permettre (comme la compilation actuelle de Prosop le fait dans une certaine mesure) de tester la cohérence interne des données. Cela ne concerne pas seulement l'aspect syntaxique ou "fautes de frappe", mais aussi et surtout la cohérence logique à

partir d'une base de règles (tel docteur ne peut pas être à deux endroits en même temps, ni écrire un volume à l'âge de 8 ans ou rencontrer quelqu'un qui est déjà mort, etc...). On aura ainsi une plus grande fiabilité des information, importante si n'importe qui peut introduire des données nouvelles dans la base. Cet aspect peut servir soit à contrôler l'entrée de nouvelles données, soit à constater des contradiction entre diverses sources d'information (sans que l'on doive forcément donner raison à l'une ou l'autre).

Les dictionnaires PROSOP sont rédigés de telle sorte qu'il est possible d'envisager deux utilisations de l'intelligence artificielle:

-création de réseaux: la rédaction des paragraphes concernant les liens de famille et les liens personnels en général est faite selon des règles rédactionnelles strictes (notamment, le fait que dans les relations familiales, EGO correspondant à la vedette de la fiche est toujours le référent: "son" père est X, "sa" mère est Y et ainsi de suite). Il est donc possible de faire des recherches dans la base pour reconstituer les réseaux de famille, de patronage, d'amitié.

-génération d'informations virtuelles: notamment à partir des mentions ayant trait au cursus universitaire, proposition de compléments d'information sur l'âge de la vedette, ou sur les éléments du cursus non attestés. Ces informations peuvent être générés en fonction de plusieurs bases de règles: l'une par exemple établie à partir des statuts des établissements d'enseignement (universités, facultés, collèges) impliqués, l'autre à partir d'une évaluation statistique des éléments attestés dans la base, ce qui permettra d'interroger le système sur la valeur de vérité probable de tel ou tel fait hypothétique.

Cette approche devrait permettre de reconstruire des aspects manquants de la vie d'un auteur, d'un manuscrit, etc, par inférence à partir des faits avérés et des bases d'information dont on dispose.

L'exploitation hypertexte et multimedia de PROSOP

On peut penser accéder au texte dans une bibliothèque électronique de plusieurs façons. Soit, on sait sur quel texte on veut travailler et on demande le texte. Soit on cherche à construire un corpus à partir de critères externes au texte et l'on ne sait pas à l'avance quels textes il s'agit de consulter, et si les textes dont on trouve les titres satisfont aux exigences de constitution du corpus (par exemple). Le module de sélection de PROSOP est de ce point de vue un outil précieux: il permettre par exemple de déterminer aisément tous les textes produits par exemple par les auteurs qui ont été membres du Trinity Hall de Cambridge entre 1580 et 1590, puis de sélectionner là-dedans un genre particulier de production, par exemple les poèmes latins ; ou encore, touts les commentaires des Sentences produits par des maîtres originaires de la France du Nord ayant eu leur maîtrise en théologie entre 1385 et 1395. L'idéal est ensuite de pouvoir afficher les textes à l'écran. Le texte peut être disponible sous plusieurs formes: forme numérisée, le texte ayant été saisi par un procédé quelconque. Pour les textes médiévaux, ce n'est pas encore l'opulence, mais peu à peu cela vient, et notre équipe apporte aussi sa pierre à l'édifice, en ayant constitué un important corpus de textes politiques médiévaux français et anglais. Mais le texte peut aussi être disponible sous forme d'image, permettant d'accéder à l'image du manuscrit ou d'une édition particulière. A partir d'une recherche dans PROSOP, on

devrait ainsi pouvoir accéder à une bibliothèque de textes électroniques (sur disque ou CD Rom) ou à une banque d'images de textes (sur vidéodisque par exemple). C'est donc une double optique qui s'impose: hypertext, pour pouvoir naviguer dans les données de toutes les façons possibles par un jeu de liens et d'appels faciles à mettre en oeuvre, et multimedia, puisqu'il s'agit aussi d'avoir recours à l'image.

A ce stade, il faut aussi prévoir une interface efficace avec un logiciel d'analyse lexicale, ou introduire les modules élémentaires d'analyse lexicale. Ceux dont nous avons besoin sont une possibilité d'indexation, avec comptage et mise en mémoire des références, un générateur de concordances et de listes d'items spécifiques à un texte ou communs aux différents textes d'un corpus, et des modules statistiques élémentaires (calculs de probabilité, analyse factorielle).

III. Vers un système d'information

1° Les bases de données et leurs relations

La table qui suit permet de prendre commodément connaissance de l'enchaînement des différentes bases qui, toutes ensemble, constituent le système d'information dont HISPOL est en quelque sorte le socle. Le contenu principal de chacune des bases de données est en italique ; son nom d'usage en caractères gras, bien que dans le cas de MEDITEXT les textes ne soient pas pour le moment structurés en une unité physique distincte ; par ailleurs, la base de données MS n'est pas (et ne sera peut-être jamais) complétée. OPUS, HP et HPNUM sont connectables par l'intermédiaire du champ "NUM", qui contient le numéro d'individu de l'auteur (champ 1a d'HISPOL) et en format dBase ; toutes trois sont issues des fichiers HP en format PROSOP, qui contiennent toute l'information de base en format texte, y compris une partie des textes (par exemple les discours parlementaires) ; mais ceux-ci comme tous les autres textes existent cependant sous forme d'entités séparées.

LES "AUTEURS" : HISPOL, dictionnaire bio-bibliographique.

- Le dictionnaire est constitué par les 38 fichiers **HP??.DIC** auxquels s'ajoutent le fichier **ANNEXE1.DIC** (fantômes et refusés) ; les fichiers sont en format texte sous SPF, transférables sous WORD6 via une série de Macros. Les fichiers contiennent, en clair, des informations sur :
  - origines et relations sociales.
  - éducation.
  - carrière (laïque ou ecclésiastique).
  - affiliations politiques et religieuses.
  - voyages.
  - écrits classés sommairement selon les champs (champs 20-30).
  - écrits classés finement pour histoire (50) et champ du politique (60).
- **HP.DBF**, résumé codé en format dBase (sauf informations sur les textes), destiné aux interrogations documentaires et aux comptages.
- **HPNUM**, fichier dBase, mise sous forme numérique de **HP.DBF**, destiné à servir de source aux tris croisés et fichiers de données (DAT ou BRT) pour les analyses factorielles.

LES TEXTES : MEDITEXT et OPUS

MEDITEXT : rassemble tous les textes, y compris ceux qui sont déjà dans **HISPOL**. Ils sont enregistrés en full-text, en format "Texte-seul", en principe sous SPF, parfois sous WORD. Il n'y a toutefois pas de formatage SGML. Ces fichiers sont les sources de l'information interne sur tous les textes ; c'est d'eux que dérivent :

les dictionnaires : orthographiques
par ordre de fréquence décroissante
les concordances (mots dans leur contexte).
les graphes sémantiques (voisinages spécifiques de chaque mot).

OPUS.DBF, source de l'information externe (reliable à HP.DBF). OPUS contient un certain nombre d'informations dérivées automatiquement d'HISPOL (titre, nom d'auteur, dates de l'auteur et de l'écriture du texte, forme et contenu, nombre de manuscrits et d'éditions, auxquelles s'ajoutent ou s'ajouteront des informations sur le statut du texte :

- statut : privé et non diffusé, privé mais diffusé avec ou sans l'accord de l'"auteur".
: public mais inédit ou édité.
: document de travail, brouillon, copie propre, édition manuscrite ou imprimée.
- nature : texte original, traduction, commentaire, abrégé, extrait etc...

LES SUPPORTS : MSS (MANUSCRITS) ET EDITIONS (EDD)

Ces deux bases ne sont ni réalisées, ni même en cours de réalisation, à l'exception toutefois d'une base préparatoire, **MSSENGL.DIC**, qui contient une liste d'un peu plus d'un millier de manuscrits contenant des textes politiques anglais, sous format WORD. Les informations suivantes sont répertoriées pour le moment :

cote et localisation actuelle (ville, bibliothèque).
scribe, patron.
dates et localisation de l'écriture.
dates et localisation de la propriété.
autres textes présents dans le manuscrit.

Elle devrait aussi comporter par la suite des informations sur :

la "qualité" d'écriture.
la taille et l'espacement des lettres, la mise en page.
les illuminations, les décorations.