Méthode d'exploitation sur ordinateur des sources diplomatiques médiévales

C'est au titre de machine à traiter l'information logique que l'ordinateur a fait son entrée dans le domaine des Sciences Humaines. Certaines disciplines, la linguistique par exemple, en font à présent un usage courant; d'autres, et c'est le cas de l'histoire, restent plus réservées. Pourtant tout chercheur amené à puiser, pour ses propres travaux, aux sources archivistiques, à consigner leur contenu sur d'innombrables fiches qu'il lui faut sans cesse brasser, a rêvé de remettre à la machine le soin d'accomplir les tâches purement matérielles qui freinent tant de ses projets. Mais l'historien est le plus souvent un chercheur individuel; il recourt à des sources très variées sans vraiment les étudier; ce qu'il veut surtout enregistrer, c'est le contenu de ses documents; et traiter sur ordinateur un "signifié" est délicat parce que subjectif. Le manque d'homogénéité des sources, le manque d'objectivité et d'exhaustivité dans le traitement constituent d'assez sérieux barrages. Seuls les utilisateurs de documents statistiques (registres d'état-civil etc...), de plus en plus nombreux, ignorent ces inconvénients.

Toutefois les documents statistiques n'existent en nombre suffisant que pour les tous derniers siècles de notre histoire. En ce qui concerne le Moyen-Age proprement dit, l'essentiel de la documentation écrite réside dans les sources "diplomatiques" qu'il s'agisse de diplômes proprement

dits ou d'actes de la pratique. C'est de leur contenu qu'il faut pouvoir tirer parti.

*

* *

En tout premier lieu, si restreintes soient ces sources, il ne nous semble pas qu'on puisse leur appliquer un traitement uniforme : pour le Haut-Moyen-Âge, une étude attentive et complète des actes diplomatiques s'impose. A cette époque en effet, le scribe chargé de leur rédaction se voit dans l'obligation d'exprimer à l'aide d'un vocabulaire vétuste et qu'il manie le plus souvent très mal des institutions encore mouvantes et hétérogènes; la langue est donc incertaine, variable selon les lieux et les temps; on ne peut fixer l'apparition et le sens d'un mot qu'en en relevant minutieusement tous les emplois; et c'est après qu'on aura procédé à ces relevés que l'on pourra songer à connaître avec plus de certitude les notions que les mots recouvrent.

Un inventaire complet de la langue diplomatique – comme de la langue littéraire – s'impose; mais réduit aux procédés manuels, l'historien n'a pu jusqu'ici dépasser le stade de l'échantillonage. Si l'on enregistre ces textes en mémoire, on se donne la possibilité de constituer avec une grande rapidité des relevés d'occurrences, des index, des concordances; on peut tout aussi bien interroger l'ordinateur, occasionnellement, sur l'utilisation d'un terme donné dont on veut étudier les significations diverses. C'est ce qu'avait si bien signalé dès 1963 Monsieur Léopold Genicot, professeur à

l'Université de Louvain, dans une communication faite à l'Académie royale de Belgique; en 1966, Monsieur Jean Schneider, doyen de la Faculté des Lettres de Nancy, manifestait le même souci en créant au sein du CRAL de Nancy une section des "textes non littéraires" qui m'est à présent confiée.

Je n'insisterai pas davantage sur ce travail qui ne fait pas l'objet-même de l'exposé; à l'assistance, je n'aurai rien à apprendre sur la constitution des "index verborum". Monsieur Evrard m'a prouvé l'an dernier, en tentant l'expérience sur l'un de mes actes, que la lemmatisation automatique, même appliquée à un mauvais latin, n'était pas utopique, et je lui en sais gré. Malheureusement, les circonstances ne m'ont pas permis d'aller encore aussi loin dans cette exploitation. Mais avant de quitter ce sujet, j'insisterai cependant sur deux points essentiels, me semble-t-il, du traitement : tout d'abord, il est évident qu'un relevé global de vocabulaire n'aurait aucun sens : le corpus réuni ne présente ni unité de lieu ni unité de temps; il a donc paru indispensable de signaler approximativement dans la référence chiffrée qui accompagne chacun des mots, la région dont l'acte qui le contient émane, la date de sa création. D'autre part, l'acte diplomatique est truffé de formules ou de citations tirées de documents du Bas-Empire, ou empruntées à des chancelleries prestigieuses. Nous nous trouvons donc à ce moment en face d'un vocabulaire artificiel, anachronique, qu'il faut pouvoir distinguer du reste de l'acte si l'on veut éviter les erreurs. Nous avons tenu compte de ce fait dans l'étude du vocabulaire, en indiquant si un terme est ou non compris dans une formule.

*

* *

C'est un tout autre genre de méthode que je crois devoir utiliser pour des documents plus récents. La masse des "actes" va croissant, à partir du XIIe siècle et l'on ne pourrait guère envisager un enregistrement des textes in-extenso; au demeurant, l'acte s'allonge d'interminables formules dont l'intérêt diminue parce qu'elles se figent et s'uniformisent.

On aurait pu envisager d'enregistrer et d'exploiter des extractions d'actes, comprenant l'essentiel du contenu. Mais il me faut alors revenir à ma remarque préliminaire : ce qui intéresse l'historien est avant tout le "signifié", la notion; il s'est trouvé que pour des époques anciennes, il a semblé dangereux et inefficace de séparer l'étude du concept de l'étude du mot, du moins pour l'instant. Mais à partir du XIIe siècle, environ, les institutions s'affermissent et il est moins malaisé de les dégager d'un texte. Dès ce moment, l'extraction pure présente au contraire un grand nombre d'inconvénients. A les énumérer, je ne ferai que démarquer les propos de J. C. Gardin, mais il se trouve qu'ils s'appliquent particulièrement bien à une exploitation de documents historiques : en premier lieu, si la question n'est pas formulée dans les termes que le scribe a utilisés dans les documents, ces derniers ne seront pas retenus même s'ils sont pertinents. Pour pallier cette difficulté, il faut recourir à ce que J. C. Gardin appelle "l'indexation à l'envers" qui consiste à amplifier une question en assortissant le terme demandé d'un certain nombre de synonymes; on se doute des difficultés de la méthode quand on se trouve devant un ensemble de textes anciens dont le vocabulaire ne peut être que difficilement et incomplètement recensé. Toutefois, le problème peut être résolu.

Ce n'est pas le cas quand surgissent les difficultés de l'homonymie : il faut alors accepter qu'à l'interrogation, un certain nombre de documents non pertinents — ceux qui précisément contiennent les homonymes — soient retenus; il faudra les écarter à la main.

Mais le défaut le plus grave de la méthode réside dans la polysémie ou plus exactement dans les glissements ou dans les divergences des sens d'un même terme d'une région à l'autre, d'une époque à l'autre.

Aux extractions de texte naturel j'ai donc préféré des condensations indexées des actes médiévaux. Au demeurant, les motifs invoqués ci-dessus n'ont fait que renforcer *a posteriori* une position déjà choisie depuis longtemps : la section de diplomatique de l'Institut de recherche et d'histoire des Textes à laquelle je suis attachée a pour tâche depuis un quart de siècle de recenser et d'analyser une source diplomatique particulièrement intéressante, celle que constituent les "cartulaires" (recueils d'actes) conservés dans les dépôts de France. Plus de 500 cartulaires, représentant environ 150.000 actes, ont été analysés jusqu'à présent. Il m'a semblé que si l'on pouvait enregistrer sur ordinateur, les analyses des actes relevés, leur contenu pourrait être exploité rapidement et exhaustivement. C'est la raison première de mon entreprise et, par là-même, dès l'origine je l'ai conçue comme une exploitation de textes indexés.

Il ne reste guère que le temps d'exposer rapidement les principes essentiels de ce traitement. Je passe sur la nécessité d'uniformiser le vocabulaire utilisé dans les analyses, afin de diminuer les risques d'ambiguité et de confusion; également sur les conditions d'une bonne indexation qui nécessite un travail en commun, et la consultation des spécialistes intéressés par le projet (une "commission" composée d'historiens et de juristes contrôlera périodiquement l'exactitude des descripteurs retenus). Mais je crois devoir insister sur deux aspects essentiels du traitement : un historien s'intéresse rarement à une notion isolée; même s'il en est ainsi il est évident qu'il demandera non seulement les documents contenant cette notion, mais également ceux qui contiennent des termes plus spécifiques se

rattachant au même concept : quand le chercheur s'intéresse par exemple au gibier, il exige qu'on lui fournisse du même coup les documents qui parlent de loups, de perdrix ou de lièvres. Si donc les descripteurs ne sont pas enregistrés isolément en mémoire, mais constituent un lexique organisé en classes, chapitres, champs, il sera possible, par sommation automatique, de fournir à la demande tous les termes spécifiques dépendant d'une même classe générique. Il a paru indispensable de doter les descripteurs de cette organisation paradigmatique.

Le plus souvent, d'autre part, un concept n'intéressera l'historien qu'en fonction d'un autre : s'il étudie par exemple la constitution de rente, il ne lui sera pas indifférent de savoir quelle catégorie d'individus a procédé à l'opération, au profit de quelle autre elle a été réalisée, sur quel genre de bien la rente a été assignée. Les relations occasionnelles entre deux notions ne sont donc pas à négliger et il a paru nécessaire de pourvoir le système d'une organisation syntagmatique assez poussée*; on n'y recourra cependant qu'en cas de nécessité, parce qu'elle est coûteuse.

*

* *

---

* L'exploitation même n'a pas encore débuté. Mais la mise au point du programme, dans son ensemble, a pu être effectuée grâce à la collaboration de l'Institut Universitaire de Calcul Automatique de Nancy et plus particulièrement de Madame Marion Créhange, maître-assistant à la Faculté des Sciences. Sous la direction de cette dernière, la programmation a été entreprise par les étudiants de la section de programmation de l'Institut Universitaire de Technologie de Nancy.

Toutefois, il ne peut être question dans une exploitation de ce genre de négliger le problème du vocabulaire qui reste primordial : car si l'on ne peut raisonnablement penser à étudier une notion à partir d'un mot, on ne peut pas plus se dispenser de soutenir un concept de la notation des termes qui l'expriment. La constitution d'un "thesaurus documentaire", comme l'appelle J. C. Gardin, mettant en relation le descripteur avec les termes correspondants utilisés dans les actes – et pourvus d'indications de lieu et de temps – sera la prochaine étape de cette exploitation. Elle permettra d'améliorer le lexique, d'en corriger peu à peu les erreurs. Surtout, c'est elle qui peut nous mettre sur le chemin d'une exploitation automatique des textes qui reste le but essentiel à atteindre mais dont nous sommes encore pour l'instant très éloignés.

Lucie FOSSIER

Attachée à la Section de Diplomatique
de l'Institut de Recherche et
d'Histoire des Textes

Responsable du "Traitement automatique
des Documents historiques médiévaux"
au C.R.A.L. de Nancy