# INFORMATIQUE ET HISTOIRE

**Jean-Philippe GENET**

Le but de la présente recension est de faire ressortir l'originalité (très partielle, on le verra) du recours à l'informatique chez les historiens : à partir de là, de faire des propositions concrètes au niveau des programmes d'enseignement de l'informatique, et enfin de proposer une bibliographie rapide des travaux susceptibles d'aider dans leur travail nos collègues formateurs.

## A/ HISTOIRE ET INFORMATIQUE

Les historiens ont découvert l'ordinateur un peu tard, vers la fin des années cinquante aux U.S.A., une dizaine d'années plus tard en Europe. Depuis, un travail énorme a été accompli, mais il est assez difficile de s'en faire une idée, car ce développement a été surtout l'œuvre d'individualités immergées dans le long travail de la thèse d'Etat, ou de laboratoires et de groupes de recherches liés, d'une façon ou de l'autre, au C.N.R.S., dont les centres de calcul (notamment le C.I.R.C.E. à Orsay) ont été des outils essentiels. L'université, en tant que telle, est restée en dehors de ce développement, même si ses membres y participaient largement : d'où l'absence de manuels et de vues d'ensemble commodes.

Les historiens ont surtout été attirés par deux des apports les plus évidents de l'ordinateur : l'accès aux méthodes statistiques de l'analyse des données et la possibilité de manier des masses d'informations considérables. Il convient cependant de remarquer que, jusqu'à une date très récente (avec la mise au point du logiciel CLIO par Manfred Thaller au Max Planck Institut à Göttingen), les historiens se sont contentés d'utiliser les programmes et les méthodes mises au point dans le cadre d'autres disciplines : toute approche de l'informatique pour l'histoire est donc nécessairement pluridisciplinaire, ce qui serait beaucoup moins vrai, par exemple, pour la sociologie et la linguistique. L'historien "informatisant" a continuellement emprunté au linguiste et au

lexicologue pour le traitement de texte, au sociologue pour l'histoire sociale et politique ou pour la prosopographie, à l'économiste pour l'histoire économique. Seules peut-être se distinguent deux disciplines que l'on ne peut pas pleinement rattacher à l'histoire, dans la mesure où elles sont enseignées de manière autonome au niveau de l'enseignement supérieur, à savoir l'archéologie et la démographie : or, elles ont l'une et l'autre fait preuve d'un très grand dynamisme dans ce domaine nouveau.

Cependant il est assez facile de distinguer un certain nombre de domaines dans lesquels les historiens ont été plus particulièrement actifs :

- la constitution des bases de données, qu'elles soient à fins documentaires ou heuristiques ;
- l'exploitation statistique des banques de données et des séries chronologiques, avec tous les problèmes d'analyse des données qui s'y relient ;
- le traitement des données textuelles et les applications de la lexicologie quantitative au discours (en l'occurrence, surtout politique).

Puis, comme nous l'avons dit, il faut faire une place à part à la démographie et à l'archéologie.

Le premier axe de réflexion qui s'est donc développé plus spécifiquement chez les historiens concerne les données. La construction des grands systèmes de données, chronologiquement l'une des voies qui a le plus attiré les historiens vers l'informatique, est aujourd'hui passée au second plan des préoccupations. La lourdeur des gros systèmes documentaires, comme le système SATIN qui eut son heure de gloire chez les archéologues, effraie et ce sont plutôt des administrations (Les Archives Nationales, l'Inventaire des Richesses Artistiques de la France et les services du Patrimoine, voire certains gros laboratoires du C.N.R.S. comme l'I.R.H.T. à Paris ou l'A.R.T.E.M. à Nancy) qui ont entrepris de créer des banques de données pouvant intéresser les historiens. Il pourrait d'ailleurs être utile de demander à ces institutions des échantillons et des descriptifs de leurs banques qui, pour certaines formations historiques (historiens de l'art, branches spécialisées vers la documentation) sont essentielles.

Néanmoins, la tendance est aujourd'hui à la constitution des bases de données de type heuristique, c'est-à-dire constituées non pas en fonction de la documentation engrangée "au kilomètre", mais en fonction

d'une problématique historique précise. Ce sont d'ailleurs deux grandes bases de ce type qui en France ont donné les résultats les plus spectaculaires : l'enquête qu'a dirigée Emmanuel Leroy-Ladurie sur les conscrits français sous la Restauration d'une part, et celle qu'ont réalisée en association Christiane Klapisch et David Herlihy à partir des données exceptionnelles du Catasto florentin de 1427. Mais de telles bases peuvent être beaucoup plus réduites : c'est le cas de la plupart de celles que les chercheurs individuels construisent pour leurs thèses. Dans ces conditions se posent plusieurs problèmes : y a-t-il des règles à fixer puis à suivre pour la documentation des données ? Que faire lorsque les données sont lacunaires, ou peu fiables, donc incertaines ? Jusqu'à quel point la "métasource" fait-elle écran entre la source initiale, la donnée construite et le résultat du traitement informatique (autrement dit, le résultat d'un traitement historique dépend-il, et pourquoi, des solutions adoptées dans la construction de la base ?) ?

Doit-on envisager a priori des saisies en clair, ou peut-on impunément avoir recours aux diverses techniques de codage, ou encore à des métalangages plus ou moins ésotériques ? Si l'on opte pour le codage, comment construire un codage de données historiques ?

Ces problèmes prennent aujourd'hui une acuité nouvelle avec d'une part le développement de la micro-informatique, qui permet à certains de saisir leurs données et de constituer leurs bases directement dans les fonds d'archives ou les bibliothèques, et d'autre part les progrès de l'intelligence artificielle qui risque de démoder rapidement même des bases bien faites. Un atelier international s'est d'ailleurs constitué en 1985 à l'initiative du Max Planck Institut fur Geschichte à Göttigen pour fixer des normes européennes de standardisation des bases de données historiques (il s'est réuni à nouveau en juin 1986 à Graz). Il importe donc de sensibiliser les formateurs et les étudiants aux problèmes liés à la constitution des bases de données, à la construction et à la gestion de la métasource, et enfin à la nécessité de consulter, cumuler ou extraire des bases de données historiques.

Je serai beaucoup plus rapide sur les autres points dans la mesure où la réflexion est ici moins avancée dans le milieu historien. En ce qui concerne l'analyse des données et les applications statistiques, il faut observer que les historiens français ont suivi une route qui diverge de celle de leurs collègues américains. D'une façon générale cette différence --sur laquelle on ne s'est pas encore systématiquement interrogé-- est double. D'une part, les Américains accordent une large place à la

corrélation et à l'analyse de régression, alors que les Français pratiquent surtout l'analyse factorielle : encore que les Américains, quand ils ont recours à l'analyse factorielle, choisissent plutôt l'analyse en composantes principales tandis que les Français préfèrent l'analyse factorielle des correspondances. D'autre part, la plupart des historiens (à la différence d'ailleurs des économistes) français récusent l'économétrie rétrospective, alors que cette pratique a donné naissance aux U.S.A. à la New Economic History.

Il y a évidemment dans ces divergences un parallélisme. Cela suppose pour tous ceux qui veulent avoir accès à l'une et l'autre historiographies une connaissance de l'ensemble des méthodes. Or, ici les présupposés statistiques et mathématiques sont intimement liés aux procédures informatiques : ce sont les ordinateurs, et plus précisément les grands packages américains (S.A.S., S.P.S.S., B.M.D.P., OSIRIS) qui ont mis à la portée des utilisateurs ces outils statistiques sophistiqués dont le maniement est désormais relativement facile ; qui plus est certains de ces logiciels sont depuis peu disponibles (aux U.S.A. en tout cas) sur les micro-ordinateurs compatibles I.B.M.-P.C.. Leur emploi suppose donc une formation complexe qui unisse formation mathématique et statistique (corrélation, khi2, matrices et vecteurs propres) qui aille au-delà de la statistique descriptive classiquement inscrite au programme des D.E.U.G. de Sciences Humaines (et encore, pas toujours). Il faut d'autre part sensibiliser les uns et les autres aux difficultés de l'interprétation des résultats de ces méthodes dans le cadre de la recherche historique. Enfin, il serait souhaitable que les formateurs, au moins, soient à même de suivre les progrès réalisés dans ces domaines dans les travaux récents : ces progrès sont rapides, si l'on en juge par l'apparition dans la panoplie de travail des historiens français de l'analyse canonique (André Strauss), de l'analyse d'implication (Jean-Louis Robert), de l'analyse spatiale (Bernard Lepetit) et de la méthode TRI 2 (créée par Philippe Cibois).

Troisième grande orientation : les traitements lexicologiques et l'analyse du discours, lesquels amènent d'ailleurs à des réflexions spécifiques sur les données textuelles d'une part, sur les méthodes statistiques propres au texte (voir notamment sur ce point le livre essentiel de Pierre Lafon). Le rôle pionnier revient ici à Régine Robin qui, si elle n'a personnellement guère utilisé l'informatique, familiarisait dès 1973 dans son "Histoire et Linguistique" le public des historiens avec la lexicologie quantitative. Depuis, l'intérêt des historiens pour la sociolinguistique et l'analyse quantitative du discours s'est affirmé :

mentionnons ici les travaux portant sur l'histoire de la Révolution Française ou sur le Congrès de Tours. Les historiens n'ont certes pas développé de méthodes spécifiques : au moins ont-ils su adopter celle des linguistes, et notamment celles du groupe de lexicologie politique qu'anime, à l'Ecole Normale Supérieure de Saint-Cloud, Maurice Tournier : la revue "Mots" fait d'ailleurs une large place à ces travaux aux confins de l'Histoire et de la Linguistique (voir bibliographie, infra). Une notion reste cependant délicate et importante au niveau de la formation : c'est la notion de "corpus" et celle qui lui est liée, de "sous-corpus".

Quelques mots, pour terminer, sur la démographie et l'archéologie. Quant à la première, elle a certes bénéficié de la capacité des ordinateurs à traiter de grandes quantités de données et de l'accès à des méthodes statistiques nouvelles (voir notamment les processus de simulation et de modélisation dont Hervé Le Bras donne d'intéressants exemples dans la revue Population). Mais le problème majeur reste la reconstitution des familles qui, malgré le progrès que représente un logiciel comme CASOAR, reste encore difficile au-delà du milieu du XIXe siècle. La difficulté majeure est la résolution automatique des problèmes, de "record-linkage", aggravés par les flottements de l'onomastique, plus on remonte dans le temps. Ces problèmes existent bien sûr dans d'autres domaines de l'histoire, mais c'est là qu'ils sont les plus vifs. Quant à l'archéologie, elle est surtout confrontée à trois exigences : engranger d'énormes quantités de données, lier la saisie des données et la conduite de la fouille, et développer enfin des procédures de classification automatique performantes et fiables.

Bien sûr, tout cela évolue très vite. Les deux facteurs d'évolution les plus importants me paraissent être la micro-informatique d'une part, et l'apparition de l'intelligence artificielle de l'autre. La première a permis à beaucoup de nos collègues de se familiariser avec l'informatique, et de rencontrer – voire résoudre – les problèmes signalés plus haut. La seconde apparaît comme susceptible de transformer profondément l'utilisation de l'informatique dans nos disciplines : des travaux pionniers ont déjà été réalisés en archéologie (M. Renaud et M. S. Lagrange), mais de jeunes programmeurs se tournent déjà vers les micro-ordinateurs (exploitation du dictionnaire du Mouvement Ouvrier de Jean Maitron sur micro-ordinateur IBM PC).

## B/ PROPOSITIONS DE PROGRAMME SPÉCIFIQUE "HISTOIRE"

Les quelques indications qui suivent sont destinées à servir de base pour une réflexion sur ce que pourrait être la partie spécialisée (c'est-à-dire correspondante à la matière dominante) d'une U.V. de D.E.U.G. Sciences Sociales/Sciences Humaines. Il va sans dire que, comme ce programme ne concerne qu'une petite partie des vingt heures d'enseignement dont bénéficiera chaque étudiant, il ne peut être question de l'enseigner en entier ; outre le fait qu'une modulation (dont une esquisse est donnée ci-dessous) doit être faite discipline par discipline, trois principes devraient, à mon sens, guider les enseignants concernés :

- privilégier la réalisation en commun avec les étudiants de l'outil pédagogique (une base de donnée, un texte) qui servira de base aux démonstrations ultérieures.
- privilégier les aspects du programme qui peuvent être liés à des enseignements donnés dans le cadre du même D.E.U.G.. C'est en particulier dans ce cadre qu'il convient d'entreprendre avec les étudiants une réflexion approfondie sur la nature de la donnée historique : problèmes induits par la dimension chronologique, incertitude dans les données, caractères lacunaires, nécessité de définir des références (cotes d'archives etc.).
- lier le niveau de l'enseignement au niveau réel en informatique des étudiants, qui peut être extrêmement variable selon qu'ils ont ou non suivi un enseignement d'informatique dans le secondaire, ou qu'ils disposent à domicile d'un équipement.

De toute façon, l'énumération qui suit doit être comprise comme un menu dans lequel étudiants et enseignants pourraient puiser ad libitum : seuls les centres d'appui devraient d'abord rassembler des logiciels couvrant l'ensemble du menu, et être ensuite capables d'en assurer la diffusion et la maintenance.

### I- Base de données et gestion de fichiers

Construction d'une mini-base de données relative à la discipline dominante (exemple : la base qu'a fait construire à ses étudiants André Zysberg à partir du dictionnaire biographique des généraux de l'Empire, ou celles que j'ai fait construire aux miens successivement sur les membres de l'Académie Française, les députés élus en 1981, les radios libres en 1984 etc.). La base peut être de type bibliographique,

documentaire, textuelle, ou codée à structure matricielle (type dépouillement d'enquête).

A partir de ce travail, plusieurs lignes de réflexion peuvent être développées :

- qu'est-ce qu'implique la création de données (notion de méta-source) ?
- quels sont les objectifs d'une recherche informatisée: situer les perspectives d'exploitation documentaire, statistique, lexicale, bibliographique ?
- réexaminer, à partir de cette expérience informatique, les notions de questionnaire, de codage et de définition de variables.

**II- Dépouillement d'enquêtes - Traitements statistiques**

Cette partie de l'enseignement doit être étroitement dépendante du programme de l'U.V. de Statistiques qui figure dans de nombreux organigrammes de D.E.U.G. : l'éventail des points abordés pourrait être :

- notion de tri (tri plat, tri croisé) et tests statistiques associés aux tris, notamment khi 2 et phi 2.
- statistiques descriptives : moyenne, médiane, écart-type.
- traitement des séries chronologiques.
- corrélations et régressions (simples et multiples).
- analyse factorielle (en composantes principales et surtout analyse factorielle des correspondances).
- analyse hiérarchique et classification automatique.

**III- Traitements graphiques**

La liste qui suit n'est pas limitative :

- construction de pyramide des âges.
- établissement de courbes et construction d'histogrammes ; lissage de courbes.
- cartographie automatique.
- construction de schémas divers (stemmes, généalogies et structures familiales, etc.).

- construction de caractères nouveaux à l'écran.

-initiation au traitement de l'image.

**IV- Traitement de texte**

Il s'agira là aussi de créer un texte concernant la discipline majeure du D.E.U.G., et d'en assurer l'exploitation avec les étudiants, à des niveaux évidemment très variable ; l'expérience de la création d'un texte peut être combinée le cas échéant avec une comparaison avec des données provenant de texte antérieurement créés.

- indexation alphabétique ou par ordre de fréquence décroissant ; indexation par orthographe inversée ; localisation des références. Il sera souhaitable d'entreprendre à partir de ce premier travail une réflexion sur les implications du choix de l'unité d'indexation (notion de mot, lexème, phonème, etc.).
- calcul des occurrences ; notions de fréquence, fréquence relative : notion de co-occurrence ; étude de la répartition du vocabulaire, notions de rafale, de fonctionnement syntagmatique etc.
- analyse syntaxique et problématique de l'analyse automatique du discours ; problématique de la traduction automatique.
- analyses stylistiques.
- la notion de banque de textes.

**V- Enseignement assisté par ordinateur**

Présentation d'exemples de didacticiels concernant la discipline majeure du D.E.U.G. concerné et initiation à leur utilisation. A ce stade, un certain degré de coordination est souhaitable avec nos collègues de l'enseignement secondaire.

## C/ BIBLIOGRAPHIE

Il n'existe pas de livre d'ensemble consacré à "Histoire et Informatique", si ce n'est l'ouvrage de Edward Shorter, *The Historian and the Computer*, a practical guide, Englewood Cliffs, 1971, excellent en son temps, mais vieilli. Un article de présentation général contenant une importante bibliographie est : J.Ph. Genet, "L'Historien et l'Ordinateur", *Historiens et Géographes*, 270, 1978, p. 125-142. Il est donc indispensable, pour suivre le chemin parcouru, de se reporter aux

périodiques et aux actes de nombreux colloques qui permettent l'échange rapide des informations. Enfin, je mentionnerai quelques ouvrages particulièrement importants.

**1. Périodiques**

En français, on dispose, en dehors de la parution d'*Histoire & Mesure*, depuis Mars 1986 par les éditions du C.N.R.S., essentiellement de deux newsletters, faites par des historiens pour des historiens, et dont le service est assuré gratuitement à ceux qui en font la demande :

- *Le Médiéviste et l'Ordinateur* (publié par l'I.R.H.T., le C.R.H., et l'U.A. - 1004) édité par une équipe animée par Lucie Fossier. Il est possible de se procurer le volume de reprint des 10 premiers numéros (50f) auprès de l'I.R.H.T., 40 Avenue d'Iéna, 75116.
- *H.M.C.I.* (publié par l'I.H.M.C. et le L.I.S.H.) : le responsable de la publication est André Zysberg.

L'une et l'autre sont gratuites et font une large place à la micro-informatique et à ses applications. Elles contiennent de nombreux exemples de traitements historiques, la présentation de méthodes nouvelles, des mises au point sur des problèmes techniques. On peut les compléter par d'autres revues :

- *Archéologues et Ordinateurs*, newsletter publiée par le C.R.A. à Valbonne qui couvre l'archéologie sans distinction de période. Le responsable de la publication est Henri Ducasse.
- *B.M.S.* : une autre newsletter, aussi publiée par le L.I.S.H.. Les responsables de la publication sont Karl Van Meter et Philippe Cibois. La revue est très utile pour tous ceux qui utilisent les statistiques en Sciences Humaines.
- *Informatique et Sciences Humaines*, publiée par Paris IV (Département de Mathématiques) et l'E.H.E.S.S.. Le responsable de la publication est Pierre Weiss. La revue publie des dossiers sur des sujets variés, dont certains touchent à l'histoire.
- *Mots*, publiée par la Fondation Nationale des Sciences Politiques : cette revue de très haut niveau contient d'excellents exemples d'application de l'informatique à l'étude des textes politiques.
- *B.D.S.P.* : une newsletter publiée par le C.N.R.S. et l'Institut d'Etudes Politiques de Grenoble.
- *Bulletin d'Information* du Service Informatique des Archives de France (responsable de la publication : Ivan Cloulas) qui, outre des

informations indispensables sur l'informatisation des archives historiques, contient souvent de très utiles mises au point sur le traitement secondaire.

Enfin il convient de signaler quelques revues étrangères :

Computer and the Humanities est évidemment une revue utile à toutes les Sciences Humaines, mais elle est à prédominance littéraire. L'historien trouvera beaucoup plus de richesses dans la revue allemande *Quantum* et surtout dans l'excellente *Historical Methods Newsletter*, publiée aux U.S.A..

**2. Colloques**

Je signalerai ici quelques titres essentiels seulement :

- Archéologie et histoire ancienne :

J.E. Doran et F.R. Hodson, *Mathématics and Computers in Archaeology*, Edimbourd, 1976.

D.R. Hodson, F.G. Kendall, P. Tautu, *Mathematics in the Archaeological and Historical Sciences*, Edimbourg, 1971.

*Les banques de données en archéologie*, Paris, 1974 (colloque international du C.N.R.S.).

En outre, l'équipe d'Archéologues et Ordinateurs publie tous les ans un bilan des activités informatiques en archéologie.

- Histoire médiévale :

A. Fossier, A. Vauchez et C. Violante, *Informatique et Histoire Médiévale*, Rome, 1977.

B. Gilmour-Bryson, *Computer Applications to Medieval Studies*, Kalamazoo, 1974.

- Toutes périodes historiques :

A. Millet, *Informatique et Prosopographie*, Paris, 1985 (colloque international du C.N.R.S.).

J.M. Clubb et D.K. Scheuch, *Historical Social Research. The use of Historical and Process-Produced Data*, Stuttgart, 1980 (donne un accès fondamental à une importante bibliographie allemande et américaine).

**3. Applications diverses**

Constitution et exploitation de bases de données :

A. Leroy-Ladurie, J.P. Aron, P.Dumont, *Anthropologie du conscrit français*, Paris, 1972.

Ch. Klapisch et D. Herlihy, *Les Toscans et leur famille*, Paris, 1978.

L'un et l'autre de ces articles sont accompagnés d'importants articles parus dans les Annales E.S.C.

Corrélation, analyse de régression multiple, path-analysis :

A. Shorter et Ch. Tilly, *Strikes in France, 1830-1968*, Cambridge (Mass.) ; 1974.

Analyse factorielle :

Ph. Cibois, *L'analyse Factorielle*, Paris, 1983.

Parmi les nombreuses utilisations par des historiens, signalons :

A. Prost, *Le Vocabulaire des proclamations électorales de 1881, 1885 et 1889*, Paris, 1974.

J.L. Robery, *La scission syndicale de 1921. Essai de reconnaissance des formes*, Paris, 1980.

B. Millet, *Les chanoines du chapitre cathédral de Laon, 1272-1422*, Rome, 1982.

Pour comprendre la New Economic History et disposer d'un accès commode à une importante bibliographie :

A. Heffet et R. Andreano, *La nouvelle histoire économique*, 1977.

Pour le traitement des textes :

*Des tracts en Mai 1968*, Presses de la Fondation Nationale des Sciences Politiques, Paris, 1975.

Pour la démographie :

J.P. Barbet et M. Hainsworth, *Le logiciel Casoar*, Paris, 1981.

Et les thèses de J. Dupaquier et J.P. Bardet.

Extrait du *Rapport sur l'utilisation de l'informatique dans les études de Sciences Humaines*, Paris, M.E.N. Secrétariat d'Etat chargé des Universités, 1986, p 31-40.

Jean-Philippe GENET
Université de Panthéon-Sorbonne
(Paris I)