Le traitement informatique des collections de renseignements appliqué à la recherche en sciences humaines

par

R.A. PATESSON

Université Libre de Bruxelles - BELGIQUE

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

La nature des données que le chercheur utilise en sciences humaines détermine le choix des moyens qu'il adoptera pour leur traitement.

Notre propos est de développer ici les résultats d'une expérience portant sur l'étude des conditions du traitement informatique par un logiciel général, d'une catégorie de données que nous avons appelé : "collection de renseignements". Cette étude nous a conduit par ailleurs à développer le logiciel d'application : GITRI.

Lorsqu'on s'intéresse aux travaux de recherche dans des disciplines telles l'histoire, l'archéologie, la papyrologie, la philologie, la linguistique, etc ..., on s'aperçoit que les chercheurs sont en face d'ensemble de données, d'informations, très différentes entre elles. Les conditions du traitement informatique de base de ces données posent des problèmes du fait de leur disparité.

Nous savons que les données peuvent se présenter sous différentes formes :
- des chiffres et des nombres,
- des mots,
- des suites de mots, des textes,
- des symboles, des suites de signes,
- des codes.

Les textes, les nombres et les chiffres ont conduit vers le développement de nombreux outils de traitement appropriés comme par exemple, les logiciels de concordance, d'analyse lexicale, d'analyse des cooccurrences, etc..., et vers des programmes d'analyse statistique (SPSS, BMD, SPAS, etc...) qui constituent des instruments privilégiés et indispensables pour l'étude de données de cette nature.

Mais souvent, et presque toujours l'historien, le papyrologue, le linguiste, le philologue et plus encore l'archéologue sont en possession de données qui ne peuvent être ramenées uniquement à des chiffres, des nombres ou des textes.

Ils possèdent des objets et sur chacun d'eux un ensemble d'informations qui sur le plan formel peuvent se présenter tout à la fois comme des codes, des symboles, des chiffres, des mots, des phrases.
L'exemple d'une fiche bibliographique illustre assez bien ce propos et correspond à cette notion d'ensemble de renseignements, on y trouve des noms (auteur, éditeur), des chiffres (date, volume, pages, etc...), des sigles, des phrases (titre).

Un archéologue possèdera différents renseignements sur un objet : sa nature, sa situation, les inscriptions qui y figurent, des dates, des identificateurs de forme, d'orientation, de couleur, de contenu, de matière, etc...
Le philologue s'occupant de tradition manuscrite et comparant des copies l'une à l'autre érigera chaque variante comme un objet avec un identificateur, une valeur, le nombre d'attestations de la variante, etc.

Nous avons adopté d'appeler ces données, des collections de renseignements sur un objet ; elles correspondent donc aux informations que l'on inscrit le plus souvent sur des fiches.
Sur ces collections de renseignements, le chercheur est amené à réaliser un certain nombre de manipulations et de traitements de base qui sont autant d'opérations élémentaires et répétitives préalables à des

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

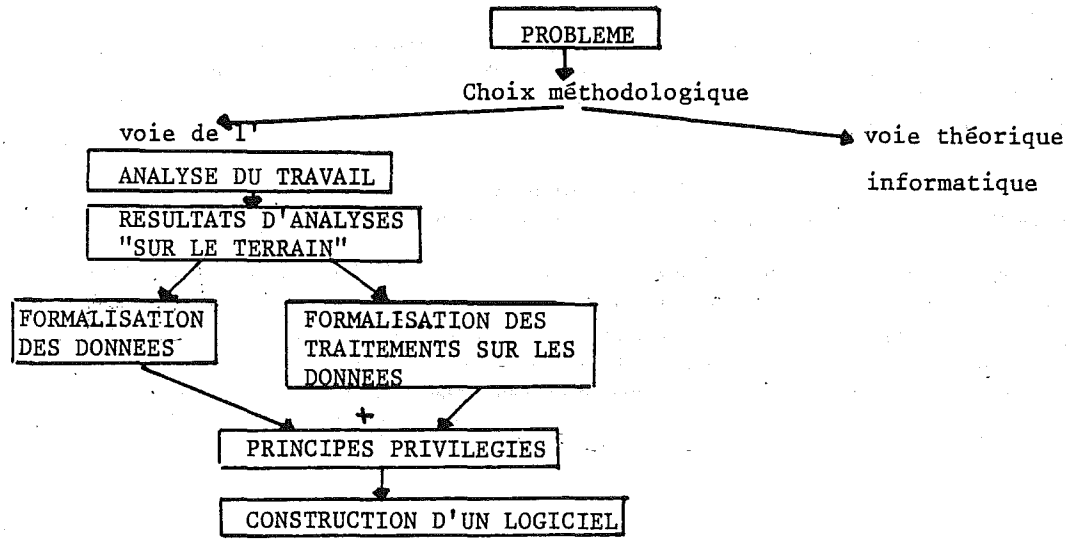
traitements plus élaborés par des moyens statistiques ou multivariés par exemple.

S'il veut obtenir des traitements automatiques simples sur ces fichiers de données brutes, le chercheur est confronté d'une manière générale à l'alternative suivante :

- Il peut construire un logiciel ad hoc. Dans ce cas, il est conduit à investir en programmation et cela peut être un investissement important mais qui peut se justifier par exemple par la taille du corpus ou par le degré de raffinement souhaité dans l'exploitation des informations. Cette éventualité implique que le chercheur ait un bagage informatique suffisant pour se prendre en charge ou bien qu'il puisse avoir recours aux services d'un informaticien.
- L'autre possibilité de l'alternative est de recourir pour les dépouillements à des programmes existants. Il se fait que la nature des données elle-même et les traitements élémentaires souhaités rendent ces programmes très rares et que bien souvent ils sont construits pour des cas particuliers.
  L'utilisateur doit alors, soit adapter ses données et ses ambitions aux programmes qu'il souhaite utiliser soit transformer le programme pour l'adapter à son problème ce qui requiert les mêmes moyens techniques que pour la solution précédente.

L'expérience dont il est question ici n'est pas indépendante de ces choix et c'est confrontés à de multiples demandes venant de chercheurs de disciplines diverses, que par manque de moyens et par souci d'économie et de généralisation nous avons décidé de notre démarche.

Le problème et le cadre de référence étant posés, notre cheminement a été le suivant :
Nous avons développé une méthodologie d'approche du problème du traitement des collections de renseignements comme celle que les psychologues industriels adoptent pour l'analyse du travail en industrie ou en administration afin d'y introduire des améliorations (1).



(1) FAVERGE, J.M. (1977), L'analyse du travail, in *Traité de Psychologie appliquée, 3*, 6-60, P.U.F.,Paris.

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

Nous avons donc entrepris une analyse du travail auprès de gens concernés : les moyens de cette analyse étaient habituels :

- On leur faisait expliciter le problème de la gestion des données dans leur domaine de recherche ; comment travaillaient-ils ; comment utilisaient-ils leurs sources documentaires ; par quelles opérations arrivaient-ils à trouver les réponses aux questions qu'ils se posaient dans leurs travaux ; quelles étaient les différentes étapes concrètes dans les manipulations qu'ils entreprenaient sur leurs documents ?
- On leur faisait également parler des données elles-mêmes ; de la part d'information pertinente qu'elles contenaient par rapport à l'ensemble des informations disponibles; des transformations et des codages qu'ils opéraient sur les documents avant de faire les analyses.

Nous avons isolé ainsi les opérations qui sont effectuées le plus couramment sur ce type de données brutes. Nous en donnons ici la liste :

- Retranscrire les données à partir des documents originaux ;
- Coder certaines informations des données d'origine ;
- Contrôler les informations retranscrites ;
- Corriger les informations erronées ;
- Ajouter ou éliminer des informations ;
- Etablir des listes, des relevés ;
- Collectionner des informations ;
- Comparer des renseignements ;
- Etablir des classements complets ou partiels ;
- Rechercher une information ;
- Réaliser des dénombrements, des comptages d'occurrences ;
- Retranscrire des informations et des résultats.

Une constatation importante est la suivante :
Alors que ces disciplines sont très différentes entre elles et que les objets qu'ils étudient sont sans rapports directs entre eux, sur le plan formel on s'aperçoit que les opérations des traitements de base sont limitées,comparables d'une discipline à l'autre, que les données elles-mêmes présentent des caractéristiques formelles qui les rendent également comparables entre elles.

Nous en avons alors tiré une double formalisation, celle des traitements et celle des données.
En ce qui concerne la formalisation des données elles-mêmes, et sans trop nous étendre sur les particularités, nous avons retenu les éléments suivants :

- Les fichiers sont constitués d'un *ensemble d'objets.* Selon les cas, l'objet peut être par exemple un accident du travail, un vase étrusque, un individu, un ouvrage scientifique, etc...
- Pour chacun de ces objets, on possède une collection de renseignements qui sont autant de *variables* à prendre en considération dans les traitements.
- Les renseignements peuvent varier dans leur *nature* : ce peuvent être des codes, des chiffres, des parties de codes, des nombres, des sigles ou des abréviations, des mots, des noms, des phrases.
- Ils peuvent varier dans leur *longueur :* un seul signe comme un numéro de volume par exemple ou une suite de signes comme le titre d'un ouvrage.
- Ils peuvent varier dans leur *valeur* bien sûr, mais aussi dans leur *finalité :* un même renseignement pouvant être utilisé différemment selon le cas, selon différentes clés, à divers usages dans les traitements.

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

- Il peut exister des *emboîtements* entre renseignements et le traitement pourra être différent selon qu'on considère le tout ou la partie, selon que les éléments sont hiérarchisés ou non.
- La comparaison des renseignements entre eux peut se faire selon des *modes numériques* ou selon un *classement alphabétique* dont la clé peut être elle-même une variable.

Dans l'esprit qui est le nôtre, c'est-à-dire celui d'une adéquation de l'outil informatique à l'utilisateur — puisque notre démarche est partie de lui en allant voir dans les détails ce qu'il fait — une de nos options a été de construire un logiciel qui prend en charge les contraintes informatiques des systèmes de langage et de l'ordinateur de telle sorte que l'utilisateur soit amené à le maîtriser dans un dialogue qui lui est adapté. On ne lui demande pas de faire le pas vers l'informatique mais que l'informatique fasse le pas vers l'utilisateur.

La seule chose qui est imposée dans cette maîtrise du logiciel est la connaissance du mode d'utilisation de l'outil, qui se présente par ailleurs comme le serait le mode d'emploi d'une voiture, d'un appareil électro-ménager, d'un enregistreur vidéo.

Les principes sur lesquels le programme est construit sont :

1) permettre les dépouillements et les traitements de base que le chercheur effectue habituellement à la main, mais avec de multiples possibilités qui peuvent entre autres déboucher sur le traitement ultérieur par d'autres programmes d'analyse plus sophistiqués ;
2) permettre le traitement sur des fichiers de données eux-mêmes constitués en transformant le moins possible les informations d'origine ;
3) permettre des mises en page de résultats qui sont claires à lire et à interpréter et qui peuvent éventuellement être utilisées directement dans une publication. Cela donne à l'utilisateur le contrôle des sorties-papier afin qu'il les agence en fonction de son utilisation ultérieure ;
4) offrir un moyen simple de commander le programme par des instructions transparentes, des mots-clés écrits en clair qui correspondent aux opérations à faire entreprendre par le logiciel et isomorphes à celles effectuées manuellement.

Pour y arriver dans le logiciel lui-même, on distingue trois classes d'opérations différentes :
- selon qu'elles portent directement sur le corpus de base et répondent aux besoins de listage, contrôle et transformation des données de base avant un traitement,
- selon qu'elles répondent aux opérations du traitement lui-même devant conduire aux résultats attendus comme les tris, les classements, les dénombrements et enfin
- selon qu'elles concernent l'édition, la présentation et la conservation des résultats.

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

CLASSES D'OPERATIONS

A. SUR LES DONNEES
- transformations — codages — contrôles — listages de base, etc...
- descriptions des contenus

B. DE TRAITEMENT
- tris — classements — sélection d'informations — dénombrements — statistiques de base, etc.

C. D'EDITION
- composition de la présentation
- conservation des résultats
- production de nouveaux fichiers pour les traitements par d'autres logiciels de niveau supérieur.

Le logiciel comporte en outre des possibilités simples pour caractériser les éléments variables dans les données.
Je ne m'étendrai pas sur les autres caractéristiques techniques du programme qui sont sans intérêt pour cet exposé mais néanmoins je vous signale que il a déjà trouvé des applications diverses chez des gens appartenant à des disciplines qui ne sont pas formés à l'informatique et pour qui l'ordinateur est bien souvent vu d'un oeil hautain et méfiant (philologue, historiens, docteurs en droit, etc...).

Quelques exemples d'application :

- gestion de bibliographies,
- gestion de renseignements signalétiques,
- en archéologie :
  - études des tablettes en akkadien,
  - étude de vases grecs et chinois,
  - bibliographie de papyrus,
  - étude de tombes assyriennes,
- en philologie :
  - comparaison des variantes d'une oeuvre littéraire médiévale,
- en histoire :
  - étude des patentables à Bruxelles à la fin du XIXième siècle.

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

Exemple d'utilisation du logiciel GITRI extrait de :
G. KURGAN–VAN HENTENRYK et G. VIRE. Les registres des patentables, sources de l'histoire de Bruxelles à la fin du XIXième siècle, *Histoire et méthodes, IV,* 367-415. Ed. de l'Université Libre de Bruxelles, 1981.

Cartes de paramétrisation nécessaires pour établir la hiérarchie des patentes (cumulées) payées par les patentables

TRI PATENTES	→ titre du tri
10 ZONES	nombre de zones définies sur 1 carte
1 CARTES	nombre de cartes par enregistrement
ZONES	→ définition des zones :
01/S/SECT	— numéro de la section ;
1/01-1/02	
02/S/IDENT	— identificateur au sein de la section ;
1/03-1/06	
03/S/CARTE 1	— adresse, sexe, code professionnel, activités du patentable ;
1/07-1/57	
04/S/1	— activité principale du patentable ;
1/58	
05/S/CL IMP	— classe d'impôt (toutes les patentes payées par un patentable sont cumulées et indiquées dans la classe correspondante du tarif A) ;
1/62-1/64	
06/S/CARTE 2	— autres éléments d'information, relatifs au patentable ;
1/59-1/61	
07/S/N STAT	— numéro statistique de la profession ;
1/65-1/67	
08/S/CARTE 3	— autres éléments d'information relatifs au patentable ;
1/68-1/78	
09/S/SA	— secteur d'activité ;
1/79	
10/S/BA	— branche d'activité.
1/80	
OPTIONS	→ caractéristiques du tri :
TRIS	— ordre dans lequel les zones sont triées ;
10 9 7 5 2	
ÉDITIONS	— ordre dans lequel les zones sont éditées ;
1 2 3 4 6 5 7 8 9 10	
MISE EN PAGE	— position du début de chaque zone éditée sur le listing ;
1 3 7 58 59 62 65 68 79 80	
50 LIGNES PAR PAGE	
4 ZONES DE COMPTAGE	— nombre et définition des zones qui font l'objet d'un comptage (numéros des zones, position sur le listing, nombre de chiffres prévu) ;
5 7 10 9	
80 90 100 110	
5 5 5 5	
FILTRES	— nombre et définition des filtres (numéros des zones servant de filtres, type de filtre, nombre et liste des valeurs des filtres).
2 FILTRES	
03 2 01 1	
01 2 01 3	
FIN	

Remarques :

Les mots en capitales représentent des mots-clé du programme de tri et doivent, dès lors, toujours être orthographiés de la même façon.

Les deux filtres ont pour but de retenir, lors du tri, les patentes payées par les patentables habitant la section 3 (second filtre), et d'éditer uniquement les données relatives à la patente payée pour l'activité exercée à titre principal (premier filtre).

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.

EXEMPLE DE RESULTATS

Remarquons que le choix des codes qui peuvent paraître hermétiques au lecteur non concerné, a été fait par le chercheur uniquement dans un souci d'économie d'écriture dans la retranscription de ses données et qu'il aurait très bien pu adopter un mode d'écriture plus explicite, le programme n'imposant pas arbitrairement l'usage de tels codes.

Résultats obtenus à l'issue du tri des patentes (cumulées) payées par les patentables habitant la section 3.

Ident. Nom		Adresse	Code prof.	Act.	Cl. imp.	N° stat.	Br. et sect. d'activ.	Comptages		
3 311VANDENVELDE	JMPL	NINOVE	4	ABAT P2211	6N315		2A			
								2	7	32
3 35ANULLEN	EMR	REMP DES MOINES	84	FTBI P8811	7N391/393		2A			
								1	1	33
3 498HENRI	LMR	FABRIQUES	9	CONF P8811	11N391		2A			
3 557LANERES	LMR	FLANDRE	58	PATI P8811	11N391/392		2A			
3 66PILAET	AMR	ANDERLECHT	63	COPA P8821	11N391/392		2A			
31459RICHARD	EBD	ANSPACH	125	PATI P8821	11N391		2A			
31016SIAEMS	EPL	ST GERY	12	CONF P8821	11N391		2A			
31162VAN CRAENEN	VMR	VAN ARTEVELDE	15	PACO P8821	11N391/392		2A			
								6	7	39
3 663BRAYARD	AMR	GRANDE ILE	1	PATI P8811	11N392/391		2A			
3 848LANGHENDRIES	PMR	DEVAUX	21	PATI P8821	11N392		2A			
								2	2	41
3 769BLOEK +	R	JERICHO	2 +	BOUL P8821	8N395		2A			
								1	1	42
31429WERY	LFR	ANDERLECHT	70	BOUL P8821	10N395		2A			
3 303WILLOCX	EMR	BORGVAL	5	BOUL P8821	10N395		2A			
								2	3	44
3 788DEPSAM	CHR	MARCHE POULETS	11	BOUL P8811	11N395/302		2A			
31626SMITS BAERT	LMR	FLANDRE	112	BOUL P8821	11N395		2A			
								2	5	46
3 56BAEYENS	VER	ANDERLECHT	35 +	BOUL P8821	12N395		2A			
3 435DE VULDER	VER	CHARTREUX	66	BOUL P8821	12N395		2A			
3 289HELTZER	CMR	BODEGHEM	28	BOUL P8821	12N395		2A			
31139PLOUM	JMR	SOIGNIES	8	BOUL P8821	12N395		2A			
31041VAN BLADEL	EMR	STE CATHERINE	30	BOUL P8821	12N395		2A			
31916VANLINTHOUT	EMR	SOIGNIES	28 +	BOUL P8821	12N395		2A			
31350VERINCKX	BME	AUX GRAINS	27	BOUL P8821	12N395		2A			
								7	12	53
31055BAEDIS	D R	SENNE	29	BOUL P8821	13N395		2A			
31277BALISTER	HMR	VERDURE	23	BOUL P8821	13N395		2A			
31175BRUGGEMAN	VER	VAN ARTEVELDE	85	BOUL P8821	13N395		2A			
31075DECREE	PMR	SENNE	88	BOUL P8821	13N395		2A			
3 979DEKOSTER	EMR	RICHES CLAIRES	4	BOUL P8821	13N395		2A			
31159DELCAMPE	DMR	VAN ARTEVELDE	3	BOUL P8821	13N395		2A			
3 357DELEEBEECK	LMR	CAMUSEL	59	BOUL P8821	13N395		2A			
31353DEMEULEMEESTER	EME	AUX GRAINS	31	BOUL P8821	13N395		2A			
31970DERRE	CMR	VAUTOUR	18	BOUL P8821	13N395		2A			
3 464DESMET	JMR	CUILLER	12	BOUL P8821	13N395		2A			
31236DESMET	HMR	VAUTOUR	7	BOUL P8821	13N395		2A			
31891FERNON	PMR	SENNE	17	BOUL P8821	13N395		2A			
31239HOEBRECHTS	GMR	VAUTOUR	17	BOUL P8821	13N395		2A			
3 272MORTELMANS	PMR	BODEGHEM	51	BOUL P8821	13N395		2A			
3 150SPILLEBANT	AMR	ANDERLECHT	146	BOUL P8821	13N395		2A			
3 60VAN PEYER	VMR	ANDERLECHT	47	BOUL P8831	13N395		2A			
3 419VANDENEEDE	FMR	CHARTREUX	8	BOUL P8821	13N395		2A			
31495VANDENWYNGAERT	J R	BODEGHEM	35	BOUL P8821	13N395		2A			
3 348VANNURVEL	FMR	CAMUSEL	3	BOUL P8821	13N395		2A			
								19	31	72

Actes du Congrès international informatique et sciences humaines
1981 - L.A.S.L.A. - Université de Liège - Tous droits réservés.