M. BORILLO-J. VIRBEL

STATUT SCIENTIFIQUE DE L'ARCHEOLOGIE ET FORMALISATION DE L'ANALYSE DES TEXTES

EXEMPLE D'UN METALANGAGE D'ANALYSE DU CORPUS DES INSCRIPTIONS LATINES

1. Introduction

Donnons immédiatement les éléments à partir desquels nous souhaiterions articuler notre réflexion: 1) la recherche dans tous les secteurs des sciences sociales, mais de manière peut-être encore plus vive dans les disciplines historiques, d'un statut plus conforme aux exigences de la pensée scientifique; 2) l'importance nouvelle accordée dans cette perspective aux sources textuelles et à leur intégration dans les constructions historiques; 3) l'émergence de techniques de traitement de l'information —que l'on ne peut plus tout à fait qualifier de nouvelles— dont l'impact réel sur le point 2 (et par ricochet sur le point 1) n'a pas encore été mesuré dans toute son importance. Notre propos sera précisément de contribuer à mettre en lumière l'interdépendance de ces trois éléments et à montrer en particulier comment les contraintes imposées par le recours aux méthodes informatiques pour le traitement des textes induisent des modifications dans le statut même de l'information mise en jeu, et de ce fait comment l'intelligence de la technique retentit en dernière analyse sur la nature des constructions scientifiques auxquelles elle concourt.

Le moment est désormais révolu où l'inventaire des recherches archéologiques mobilisant les moyens de l'informatique se ramenait à une énumération —brève— de singularités. Pour certains secteurs de l'archéologie et/ou pour certains pays, les techniques de calcul font désormais partie de l'instrumentation banale de l'archéologie, au même titre que l'appareillage stratigraphique ou le laboratoire du céramiste ou du palynologue. Le problème majeur nait aujourd'hui de ce que l'on pourrait appeler la prolifération électronique, dans la mesure où elle s'effectue dans la plus grande confusion conceptuelle et qu'elle

engendre de ce fait un certain nombre de mystifications touchant aussi bien la portée scientifique véritable que le bon usage instrumental de ces moyens. A cela des raisons historiques, et en premier lieu ce fait que dans la plupart des premières expériences le recours à l'informatique était d'abord fondé sur des facteurs quantitatifs: des fouilles sans cesse plus nombreuses et étendues, interrogées par des moyens livrant une information toujours plus vaste et plus complexe, appelaient «naturellement» l'utilisation des moyens de traitement électronique des données. Ce n'est que de manière beaucoup plus détournée, par les voies plus ardues de l'analyse, que l'impact théorique de la technologie s'est progressivement dessiné, en particulier à travers les questions de nature sémiologique et linguistique, puis mathématique et logique, qui sont intrinsèquement liées à l'usage cohérent des automates. A partir de là, et au terme *provisoire* de cette évolution, c'est une conception nouvelle de l'archéologie elle-même qui se dessine, dans ses méthodes comme dans ses objectifs.

Dans ses *méthodes*: il s'agit d'une approche dans laquelle la *démonstration* relaie l'intuition et la complète, dans laquelle chaque proposition n'est tenue pour valide que si elle est accompagnée de toutes les données dont elle procède et des calculs qui la justifient et qui permettent à tout savant d'apprécier cette justification puisqu'il possède réellement les éléments qui ont fondé la décision. Tout aussi particulière sinon originale, dans ses objectifs [1], puisque la détermination de produits traditionnels comme les typologies, les chronologies, pour ne citer que ceux-ci, est maintenant entendue comme la recherche des relations qui peuvent exister entre les données —telles que des associations significatives entre certains traits ou des filiations démontrables entre certains autres— de manière à aboutir à la construction de modèles formels qui rendent compte des phénomènes observés et que l'on pourrait appeler par analogie avec des disciplines voisines, *modèles structurels*. Ainsi l'entreprise est-elle toute entière placée sous le signe de l'explicite et du vérifiable dans la recherche de régularités susceptibles de contribuer à la compréhension historique et anthropologique des faits observés. Il s'agit, en bref, de définir une démarche que l'on

[1] Remise en question des méthodes et remise en question des objectifs sont inséparables dans la problématique de l'archéologie moderne. Nous pouvons cependant mieux préciser ici notre distinction entre la *manière* de parvenir à certains résultats (méthodes), et la *nature* de ces résultats (objectifs). Ce dernier terme peut en effet avoir deux sens et, selon le contexte, désigner soit le *contenu* lui-même des résultats visés par l'archéologie (une organisation sociale déterminée, les modalités d'utilisation de telle ou telle ressource naturelle, la séquence des états morphologiques d'un type d'objet, etc.), soit le *statut* de ce résultat sur le plan scientifique, sa valeur logique et sa nature abstraite. (Quelles sont ses relations avec les observations de base qui ont servi à l'établir? S'agit-il d'une tendance statistique? d'une relation déterministe? A-t-il été inféré? Déduit? Est-ce une pure projection subjective de l'archéologue?, etc...)

De ces trois éléments: méthodes, objectifs/contenu, objectifs/statut nos remarques visent le premier et le troisième, qui sont bien entendu interdépendants. Pour la facilité de l'exposé nous ferons volontiers référence à des «objectifs/contenu» parfaitement classiques, étant cependant entendu que la subversion des deux autres ne peut s'accomplir, dans le réel, qu'en gagnant aussi ce dernier.

puisse qualifier de scientifique, dans le sens que ce terme a pris dans la pensée contemporaine classique (cf., par exemple, HEMPEL et OPPENHEIM, 1948). A ce niveau d'abstraction, un tel projet n'a encore qu'une signification limitée puisqu'il se borne à énoncer la nécessité de fonder l'archéologie sur certains principes universels qui, s'ils marquent une rupture avec quelques traits discutables de la démarche traditionnelle ne constituent pas pour autant les propositions par lesquelles les exigences énoncées plus haut pourraient être réellement satisfaites. C'est seulement par la définition précise des moyens que l'on entend mettre en oeuvre et par leur expérimentation qu'une véritable méthode peut être élaborée et un contenu effectif attribué aux principes.

2. LES MOMENTS PRINCIPAUX D'UNE DÉMARCHE SCIENTIFIQUE EN ARCHÉOLOGIE

Comme la plupart des démarches à visée scientifique, l'archéologie se construit autour des deux moments de la *collecte* de l'information, par l'exploration du domaine qui est l'objet de l'étude, puis du *traitement* et de l'*intégration* de cette information en une connaissance qui organise les éléments à un niveau supérieur de cohérence. La distinction de deux phases séparées dans un processus de ce type peut n'être qu'une approximation plus ou moins conventionnelle; ainsi dans les travaux de l'archéologie traditionnelle où l'exposé de l'information se confond souvent avec les constructions qu'elle devrait justifier, bien que de tels raccourcis ne soient légitimes que si le but visé est purement descriptif ou si la présentation adoptée constitue en elle-même une justification évidente. A l'inverse, la complexité croissante des hypothèses exige que la justification du résultat soit donnée explicitement, ce qui ne va pas sans une véritable analyse du problème, de manière à isoler et définir chacun de ses éléments avant de mettre en évidence les relations qui les unissent. Il s'agit donc, si l'on veut, de s'appuyer sur une séparation provisoire entre les parties d'un problème pour en établir plus sûrement l'interdépendance.

Dans cette situation, quatre moments principaux peuvent être définis; leur présentation séparée et linéaire telle que nous la donnons ci-dessous ne doit pas masquer les multiples inter-relations qu'ils entretiennent nécessairement dans un processus réel d'analyse, qu'il s'agisse de l'interconnexion des différentes phases entre elles ou du caractère réitératif de l'ensemble.

1. *La délimitation de la classe de phénomènes sur lesquels va porter l'expérience.*—Cette phase correspond concrètement à la constitution du corpus de textes, d'objets, de complexes culturels et naturels, etc., dont on connait ou dont on postule la pertinence. Cette délimitation n'a évidemment de sens que par rapport à une *visée* elle-même bien définie. Elle constitue une condition nécessaire à la légitimité de toute démarche ultérieure de généralisation. Fonctionnellement, il s'agira de définir des *critères de sélection* déterminant le corpus: critères internes (sémantiques et/ou formels dans le cas de textes; mor-

phologiques dans le cas d'objets); critères externes (spatio-temporels, sociaux, contextuels dans une situation de fouille, etc.). Le rôle de ces critères est de permettre de juger sans ambiguité de l'appartenance ou non de tout «phénomène» ou «document» à la classe examinée.

2. *La définition de procédures de réduction et de représentation des phénomènes.*—Ce point sera développé plus loin. Contentons-nous d'indiquer ici que nous faisons référence à deux opérations essentielles dans toute démarche scientifique:

— La définition des systèmes de représentation: dispositifs sémiologiques («codes» pour l'information perceptuelle, métalangages pour l'analyse des données textuelles, etc.).

— La représentation des phénomènes par des systèmes symboliques: passage de la perception et de l'intuition globales à des représentations analytiques dont les constituants appartiennent aux systèmes de représentation. Ainsi des matrices descriptives pour les objets ou l'iconographie, des traductions en termes de langages formels pour les textes, etc. La propriété fondamentale de ces représentations est de mettre en jeu une information qui est non seulement *explicite* mais aussi *régulière* et de ce fait apte *opératoirement* et *logiquement* au calcul.

3. *La construction de modèles.*—C'est la phase proprement structurelle du raisonnement, qu'il s'agisse de:

a) Mettre en évidence et structurer des *régularités formelles* au sein de ces données, de telle manière que l'ensemble soit cohérent par rapport à différents types de critères internes découlant, par exemple, des démarches mathématiques ou logiques mises en jeu, ou de

b) Vérifier l'effectivité de régularités connues (conjecturées).

Exemples

— Typologie, classification: pour des critères abstraits (logiques ou mathématiques), mise en évidence des sous-ensembles d'objets homogènes, selon leurs aspects morphologiques, fonctionnels, etc.
— Chronologie relative des éléments d'un corpus: constitution d'une série par la détermination de la séquence qui vérifie au mieux certains «axiomes» définissant un ordre.
— Structures significatives de genres, de styles, etc. (problème réciproque du premier: quels sont les éléments différentiels d'une classe, d'un type...).
— Modèles de structures particulières: relations spatiales (niveaux d'échanges, structures urbaines).

4. *La validation logique et la détermination de la signification du modèle.*—La validation logique peut être incluse dans la manipulation formelle des données évoquées plus haut (ex.: validation d'hypothèses statistiques), ce

qui ramène au point 3 ci-dessus. Dans le cas plus général, en se référant au cadre des sciences empiriques, elle doit cependant être envisagée à travers la mise en correspondance des structures formelles obtenues avec une évidence externe. Outre le dépistage des circularités, cette approche contribue aussi à fixer la signification du modèle par le rapprochement sensforme qu'elle implique. La systématisation du retour au sens liée en général à la validation des structures formelles rétablit la rigueur que trop de recherches conduites par des méthodes formelles abdiquent dans la phase dite «d'interprétation».

Exemples

— Vérification de la consistance de la définition d'une classe soit par productions soumises à des jugements empiriques (dans les domaines où un tel jugement vérifie des conditions minimales de reproductivité), soit par application à des phénomènes de même catégorie qui n'auront pas été intégrés dans la construction du modèle, soit par confrontation des résultats formels avec des catégories d'information indépendante (BORILLO, 1970; BORILLO, 1971; BORILLO *et al.*, 1973).

Le schéma ci-dessus n'a d'autre fonction que le rappel de quelques-unes des conditions auxquelles l'archéologie doit satisfaire dans la construction de ses raisonnements si elle entend accéder au statut scientifique dont elle se réclame explicitement. Nous entendons aussi quant à nous y faire référence pour définir la nature et la fonction des systèmes de représentation et plus spécialement de ces métalangages d'analyse des données textuelles que sont les langages documentaires.

Les systèmes de représentation

Eléments qui fondent la *nécessité* et la *nature* des systèmes de représentation:

1. Tout travail scientifique impose une *représentation,* quelle qu'elle soit, des phénomènes sur lesquels porte l'étude, représentation qui au sens le plus général implique une *réduction* de ces phénomènes. Il est donc tout à fait clair que l'alternative dans le choix d'une méthode de travail ne réside pas entre approche réductrice ou approche non-réductrice, mais bien entre réduction *explicite,* ou réduction *implicite.* Il est à remarquer que, de ce point de vue, la quasi-totalité des démarches traditionnelles aussi bien que modernistes se caractérise par l'aspect fondamentalement implicite des réductions opérées. Or, non seulement la réflexion et les opérations du raisonnement scientifique portent sur des représentations inévitablement réduites, mais encore faut-il souligner qu'en toute rigueur ces représentations doivent aussi être *régulières.* Nous entendons par là que la correspondance entre les phénomènes et les systèmes symboliques chargés de les représenter doit être telle que deux phénomènes identiques doivent nécessairement avoir la même représentation et qu'à

deux représentations identiques doivent correspondre des phénomènes identiques ou *équivalents* au regard des critères de l'étude. Cette condition minimale de l'analyse scientifique n'a évidemment aucune chance d'être réalisée en l'absence de critère d'explicitation.

2. Dans le champ de la recherche archéologique les systèmes de représentation (SR) sont constitués par des outils tels que les codes descriptifs, les langages documentaires, les grammaires formelles, etc., dont la conception et la nature sont étroitement dépendantes des éléments spécifiques évoqués ci-dessus: 1) la nature du champ considéré; 2) la visée propre suivie; 3) les propriétés linguistiques du matériau s'il s'agit de textes; 4) les modes de validation des résultats de l'analyse.

C'est ainsi que pour un même champ des visées différentes renverront, pour des raisons de pertinence, à des représentations différentes des phénomènes soumis à l'analyse; et par conséquent, induiront non seulement des différences du contenu, au sens sémantique et relationnel, mais aussi des différences de type structurel au sein des SR définis pour ces visées (l'examen, par exemple, d'un corpus épigraphique en vue soit d'une étude stylistique, soit pour en extraire des données d'ordre historique).

Inversement, une même visée, par exemple l'étude de structures sociales, impose une différentiation au sein des SR selon que l'on s'intéresse à des mythes ou à des archives d'état-civil.

Dans les deux cas, enfin, le rôle des propriétés linguistiques des textes analysés retentit sur l'ensemble de la problématique (cf. § 5).

3. La forme et le contenu des SR est étroitement tributaire des modalités de l'expérience, construction du modèle et validation envisagée. En effet, le SR est sollicité de deux façons: *a)* comme univers de référence *et* moyen d'exprimer les données; *b)* comme contexte expérimental, où les produits de l'analyse (les représentations des données *stricto sensu*) *et* un travail interprétatif fondé sur des *opérations formelles* effectuées sur ces représentations, sont confrontés avec (projetés sur) un univers externe (hypothèses pré-établies, etc.).

4. Du point de vue de ses «documents», l'archéologie présente la particularité remarquable de faire appel aussi bien à des *textes* (dans le cas de l'archéologie historique) qu'à des *objets*. La spécificité propre à chacun de ces supports impose qu'à partir d'ici le problème de la définition de SR doive être considérée séparément dans l'un et l'autre cas.

Nous présenterons donc d'abord assez brièvement les éléments les plus importants d'une démarche descriptive dans le cas de «documents perceptuels» («objets»), avant de présenter plus en détail une expérience précise conduite dans le cas de textes rédigés dans une langue naturelle: le C. I. L.

Analyse descriptive et problèmes sémiologiques

Les sources où puise l'archéologue sont évidemment nombreuses et de nature fort variée, qu'il s'agisse d'objets —poteries, outils, armes, etc.— de

documents iconographiques —peintures, monnaies, documents glytographiques, etc.— de monuments et constructions diverses, de configurations spatiales, d'éléments du milieu naturel... L'objectif, tout au moins sur le plan scientifique, demeure dans chaque cas l'extraction d'une information dont ces «documents» sont porteurs, selon des modalités liées nécessairement à la visée poursuivie et aux moyens techniques utilisés. Mais tandis que certains des problèmes les plus difficiles posés par l'analyse textuelle tiennent à la nature même des langues naturelles il s'agira, dans le cas des objets, d'assurer une représentation symbolique régulière des données perceptuelles, associées le cas échéant à des informations contextuelles: origine, association avec d'autres objets, etc. Pour ces dernières, compte tenu, par exemple, de l'histoire de l'objet depuis son invention, il s'avère parfois difficile de distinguer nettement ce qui relève de l'observation immédiate et ce qui procède déjà d'une démarche interprétative (ainsi de l'association spatiale et des associations «culturelles» qui lui sont parfois implicitement substituées). Les ambiguités de ce type signalent un des problèmes les plus délicats du raisonnement archéologique: en tant que document, l'objet est une source d'information dans la recherche de la connaissance archéologique (les données recueillies sur l'objet contribuent à justifier les constructions intellectuelles); en retour, un certain état des connaissances se reflète explicitement ou implicitement dans la «lecture» de l'objet, sans que cette dépendance, d'ailleurs inévitable, soit entachée d'aucune sorte de vice intrinsèque puisqu'elle est tout à la fois la conséquence et la condition du dépassement scientifique. Dans la pratique archéologique, ce que l'on pourrait appeler les deux positions de l'objet dans l'organisation du champ des connaissances ne sont pas toujours perçues comme distinctes. En conséquence, le rôle de l'information au moment de la formulation des hypothèses et à celui de leur vérification peut n'être pas distingué, au risque d'entacher la validité du raisonnement d'un certain nombre de failles logiques dont la circularité est la plus grave, sinon la plus fréquente.

Dans sa relation scientifique à l'objet, l'archéologue se trouve donc confronté à deux problèmes: l'un, qui pourrait être qualifié de sémiologique, consiste à étudier les moyens par lesquels peuvent être obtenues des représentations régulières; le deuxième porte sur la fonction logique et la valeur sémantique des éléments ainsi représentés, par référence au raisonnement dans son ensemble. Résoudre simultanément ces deux problèmes constitue l'une des difficultés majeures sur lesquelles achoppent la plupart des recherches: où le fil d'argumentation est très serré, les rapports entre les matériaux et leur représentation n'offrent pas de garantie de régularité (c'est le cas de la meilleure archéologie nord-americaine); et lorsque toutes les précautions sont prises pour que le système de représentation soit régulier, sa définition ne prend pas en compte les questions scientifiques spécifiques à la résolution desquelles on entend le faire servir (cf. les travaux français sur l'analyse descriptive). Il faut cependant souligner que cette conception en quelque sorte «universelle» des SR s'adapte naturellement à la perspective documentaire et gestionnaire des tra-

vaux mentionnés. Toutefois, le passage à une approche dans laquelle la mise en oeuvre de la condition de régularité des descriptions est spécifiée par la construction d'une argumentation archéologique constitue un pas méthodologique important (GUÉNOCHE, TCHERNIA, 1974; FARINAS DEL CERRO *et al.*, 1974).

Que cette intégration puisse porter à une conception nouvelle du discours scientifique dans le champ de l'archéologie, c'est en dernière analyse l'exigence épistémologique des archéologues eux-mêmes qui en décidera. Dans les limites d'un bref rappel, nous nous bornerons à isoler les opérations principales constitutives des systèmes de représentation dans le cas des objets.

Le système qu'il s'agit de définir n'a évidemment d'intérêt que s'il permet de passer sans ambiguité des objets aux symboles qui seront chargés de les représenter. Ceci suppose que les règles opératoires sont définies avec suffisamment de précision pour que le même objet donne toujours lieu à la même description, quel qu'en soit l'auteur, et en sens inverse que des descriptions identiques ne puissent correspondre qu'à des objets identiques ou équivalents au regard des critères de l'analyse. Les exemples sont aujourd'hui très nombreux de la mise sur pied et de l'utilisation de tels systèmes et on ne répètera pas la présentation systématique des méthodes d'analyse descriptive (GARDIN, 1968). Les trois moments essentiels sont celui de l'*orientation* qui fixe les conventions déterminant la position relative de l'observateur et de l'objet; celui de la *segmentation* qui définit un découpage des composantes morphologiques, selon des critères tout à la fois pratiques et pragmatiques, enfin la *différentiation* qui fixe les valeurs distinctes que pourront prendre les parties définies par la segmentation. A chacune de ces étapes se présentent des difficultés pratiques (perçues comme telles par l'archéologue) mais dont la nature véritable renvoie à des questions plus fondamentales: une segmentation trop «interprétative», outre qu'elle risque d'être mal définie, pourrait introduire des circularités logiques; à l'inverse, si elle est trop «métrique», elle risque de masquer des homologies intéressantes; la différentiation, quant à elle, revient souvent à quantifier de manière plus ou moins arbitraire un continuum [2], etc. L'expérience montre que ces obstacles sont surmontables et qu'il est possible de construire des systèmes conventionnels de représentation, ou *codes*, pour les principaux types de documents non verbaux de l'archéologie [3]. Il faut souligner que la conception de la plupart de ces codes descriptifs procède essentiellement de préoccupations concernant aussi bien la gestion des collections que l'établissement d'une communication de type «scientifique» entre archéologues (scientifique en ce

[2] Ces problèmes sont abordés de manière précise dans *Description des outils (mathématiques, linguistiques et informatiques) impliqués par la construction d'une chaîne automatique intégrée de traitement de l'information textuelle et graphique*, BORILLO *et al.*, in «Information Storage and Retrieval», *9*, 10, oct. 1973, 527-560.

[3] En témoignent, pour ne citer qu'eux, les codes élaborés au Centre d'Analyse Documentaire pour l'Archéologie ou sous sa responsabilité. Il faut ajouter que la mise au point de ces systèmes descriptifs est également réalisable pour le milieu naturel ainsi que dans d'autres secteurs des sciences de la nature: zoologie, géologie, etc.

qu'elle s'effectue au moyen d'un langage régulier soumis à des normes précises d'utilisation). Mais les codes ne sont pas par eux-mêmes scientifiques, dans la mesure où ils n'expriment aucune relation entre leurs éléments qui ne se déduise de l'observation des données perceptuelles, que ces relations soient perçues immédiatement ou à travers un appareil physique ou culturel d'analyse (et qui représentent, dans ce dernier cas, le résultat d'un état *antérieur* de la connaissance). Encore faut-il impérativement que ces relations soient explicitement notées: puisque chaque objet, quel qu'il soit, ne peut être confondu avec une juxtaposition quelconque de ses éléments apparents, sa représentation symbolique se trouvera singulièrement appauvrie si elle ne comporte pas l'indication des relations qui articulent les diverses parties (la composante «syntaxique» des codes). En fait, la recherche et la mise en évidence de nouvelles relations de type structurel constituent précisément un autre objectif, de nature non documentaire, pour lequel l'analyse descriptive est également mobilisée. Sommairement, deux visées principales seront donc assignées à l'analyse descriptive, l'une de type documentaire —constitution de catalogues, par exemple—, l'autre de type cognitif, dans laquelle les produits de la description sont utilisés dans un calcul visant à étayer une démonstration. Bien entendu, les situations que l'on rencontre dans les deux cas sont suffisamment diversifiées pour que certaines des applications rendant cette distinction assez arbitraire: la recherche de parallèles (calcul) aboutit par exemple à la constitution d'index (dans des catalogues).

3. Situation de l'analyse des données textuelles dans le contexte de la documentation automatique

La documentation, au sens général, est l'ensemble des activités par lesquelles les documents sont rassemblés et l'information utile qu'ils véhiculent enregistrée de manière à pouvoir être retrouvée et diffusée vers les utilisateurs.

Il est compréhensible que les théoriciens, et notamment les linguistes, n'aient encore accordé que peu d'attention aux méthodes de l'analyse documentaire: celles-ci se sont en effet longtemps réduites aux opérations empiriques par où l'on attribue à un texte quelconque un ou plusieurs «mots-vedette» destinés à faciliter le repérage du document lors de recherches portant sur un thème donné.

Or, le passage d'un texte original à l'ensemble de ces termes de représentation constitue de toute évidence une opération sémantique, même s'il est vrai qu'elle n'obéit souvent à aucune espèce de règle précise, et que chaque organisme de documentation, chaque analyste même, se borne à viser en l'occurrence une certaine régularité interne, fondée sur l'expérience ou l'habitude plutôt que sur une procéduce explicite.

L'intervention de l'informatique a profondément modifié les conditions dans lesquelles s'exerce l'activité documentaire (banques de données vs. bibliothè-

ques), du point de vue des problèmes à résoudre comme des services à attendre, à la faveur de l'automatisation des travaux documentaires, et ceci pour deux raisons. En premier lieu, il apparut que l'emploi de machines pour les recherches bibliographiques permettait d'étendre considérablement le nombre des informations retenues à propos de chaque document; au lieu de se limiter à trois ou quatre mots-vedette, on pouvait désormais envisager des représentations beaucoup plus riches, allant à la limite jusqu'à la mise en mémoire d'une paraphrase complète des énoncés originaux, dans un langage symbolique adéquat. En second lieu, la nature même de cette paraphrase commençait à faire l'objet d'études sérieuses, où l'on cherchait à la rendre justiciable à son tour de la mécanisation.

Ces deux aspects de l'évolution récente des techniques documentaires tendent à relier celles-ci au domaine de la linguistique; et l'on comprend que les méthodes élaborées dans ce domaine, quel qu'en soit le caractère éminemment utilitaire, aient quelque rapport avec le problème général de l'analyse sémantique, tel qu'il se pose sous diverses formes dans les sciences humaines.

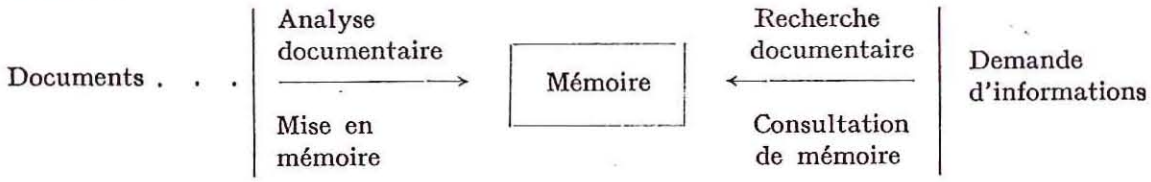
Les *systèmes documentaires,* qui tendent à remplir les fonctions documentaires attachées à un champ scientifique donné, relient ainsi, aussi bien sous leurs aspects techniques que théoriques, différents problèmes méthodologiques concernant la construction et l'utilisation d'outils d'analyse sémantique.

3.1. *Les systèmes documentaires*

Un système documentaire est un ensemble de règles qui autorisent l'enregistrement et la recherche des informations, dans un champ scientifique déterminé; c'est également l'ensemble des procédures qui sont mises en jeu pour réaliser effectivement ces opérations à l'aide de moyens techniques appropriés. La complémentarité de l'approche analytique selon laquelle sera exploré, puis restitué le contenu des documents et de sa concrétisation à travers un appareillage déterminé se reflète dans les deux types d'opérations qui marquent chacun des deux moments essentiels de l'activité documentaire. Dans le premier moment, qui est celui de l'acquisition de l'information, il revient à l'*analyse documentaire* de sélectionner au sein des documents la partie correspondant à la visée scientifique globale poursuivie, pour la transformer et l'exprimer sous une forme qui soit compatible avec les contraintes découlant de la nature du système utilisé. Ce sont les résultats de cette analyse et aux seuls qui sont retenus. L'opération d'*enregistrement,* au plan technique, est celle par laquelle ils viennent occuper la mémoire du dispositif physique, constituant le *stock* d'information où viendra puiser dans un deuxième moment la recherche *documentaire*. Cette phase pose elle-même en premier lieu des problèmes du type de ceux que l'on a rencontrés lors de l'analyse des documents: la demande d'information devra être transformée de manière à s'exprimer sous une forme régulière, c'est-à-dire comparable avec celle de l'information. Le contenu des documents ne pourra être mis en mémoire qu'après une transformation destinée

à le mettre sous une forme normalisée. Cette opération porte précisément le nom d'analyse documentaire, et l'instrument qui permet de l'effectuer, le système sémiologique dans lequel seront exprimés les produits de l'analyse constitue un *système de représentation* dont les termes peuvent ou non être organisés par un ensemble de règles qui expriment les relations existant entre eux. Dans la mesure où la recherche documentaire s'effectue par une *consultation de mémoire,* qui est un calcul strictement formel, il va de soi que les questions elles-même devront subir une transformation identique. La mise au point des algorithmes d'enregistrement et de recherche de l'information soulève d'ailleurs des difficultés techniques qui augmentent rapidement avec le volume des documents recensés et la complexité des critères de représentation, difficultés liées également aux caractéristiques des dispositifs physiques utilisés et des outils de programmation qui leur sont attachés.

Un système documentaire peut être représenté par un schéma analogue à celui-ci :



Ce schéma rend compte des traits essentiels de tout système documentaire, mais chacun de ses moments —analyse documentaire et mise en mémoire, recherche documentaire et consultation de mémoire— ou de ses *outils,* qu'il s'agisse de systèmes de représentation, de programmes de mise en mémoire, de recherche ou d'édition, de dispositif physique proprement dit, est en réalité susceptible de prendre les aspects les plus variés, selon les caractéristiques propres au système envisagé.

3.2. *Les langages documentaires*

L'information emmagasinée dans la mémoire de la machine consiste en un ensemble de configurations particulières d'éléments qui, pour les machines dites digitales, se trouvent dans l'un ou l'autre de deux états physiques possibles : magnétiques, électriques, transparent/opaque, etc. C'est seulement à travers l'utilisation d'un système de procédés opératoires de plus en plus complexes, dont l'organisation constitue précisément ce que l'on appelle le *software,* que l'utilisateur arrive à communiquer avec la machine en termes de mots ou de nombres décimaux. Ces procédés opératoires ne sont rien d'autre qu'une médiation entre la manipulation formelle de symboles et le contrôle des modifications correspondantes de l'état physique de la machine. Leur intérêt est de permettre à l'utilisateur, de la place qui lui est assignée dans le système, d'opérer *comme si* la machine comprenait les mots et les nombres. Cette «compréhension», est-il besoin de le préciser, est purement conventionnelle et strictement limitée à la définition attribuée à chaque terme. En particulier, ces

signes n'auront entre eux aucune relation, aucun rapport qui ne leur ait été explicitement affecté. Il y a là, dans la perspective de l'analyse documentaire, une opposition essentielle avec la richesse, la complexité, mais aussi avec l'imprécision, l'ambiguité du réseau notionnel de l'opérateur humain.

Le propre des langues naturelles dans lesquelles sont rédigés les documents-sources est de présenter en effet un certain nombre de phénomènes qui vont à l'encontre des exigences de l'automate. Le contenu des documents ne pourra donc être utilement mis en mémoire que sous une forme normalisée. Nous appellerons analyse documentaire l'ensemble des opérations qui permettent d'aboutir à cette forme et *langage documentaire* (L. D.) le système de représentation qui en est l'agent.

L'utilisation d'un automate implique par conséquent la construction préalable d'un tel réseau pour le champ scientifique visé, en isolant les concepts que l'on juge utile de retenir et en définissant les relations qui les articulent les uns par rapport aux autres. La mise sur pied de cette trame sémantique fixe exactement la nature de l'information qui sera effectivement traitée par le système documentaire; et cette précision, répétons-le, est tout-à-fait indispensable si l'on entend confier à une machine le soin de conserver, puis de reconnaître et de restituer tout ou partie de cette information.

Il est inutile de souligner que si les deux opérations évoquées ci-dessus sont entendues dans leur acception la plus générale, elles désignent alors ni plus ni moins l'essentiel de la démarche cognitive, et cela seul suffirait à marquer ce que seraient les ambitions d'une démarche documentaire extrême. En fait, chaque système documentaire réel se définit par rapport à une *visée particulière* et par rapport aussi à un *niveau de connaissances*.

Un *lexique documentaire* est l'ensemble des termes qui figurent dans un L. D. C'est par conséquent la *forme minimale* du L. D. Ces termes sont appelés *descripteurs*, ou *mots-clé*. Il peut se présenter en particulier sous la forme de *dictionnaire normalisé* (liste alphabétique des termes à utiliser nécessairement pour l'indexation) ou de *classifications* définies selon des critères sémantiques. Ainsi, par exemple, des classifications qui groupent selon des structures arborescentes les sous-ensembles de termes ordonnés par inclusion (mais on dispose alors d'un outil qui est davantage qu'un lexique puisqu'il comporte des relations entre les termes).

La *syntaxe* du L. D. est l'ensemble des procédés utilisés pour exprimer les relations logiques que l'on a décidé de retenir entre les termes. Chaque document représenté (indexé) en vue de son enregistrement et de son *stockage* dans la mémoire du système physique, se réduira donc à une liste de descripteurs normalisés, articulés entre eux (en général deux à deux) par des relations marquées formellement (au même titre que les opérations arithmétiques ou logiques habituelles).

Dans la mesure où la recherche documentaire s'effectue par une consultation de mémoire qui est un calcul strictement formel, il va de soi que les questions elles-mêmes devront subir une transformation identique. Un *langage do-*

cumentaire est donc un ensemble de termes, *avec* ou *sans* procédés syntaxiques conventionnels, utilisé pour représenter un certain nombre des documents scientifiques, en vue du classement et de la recherche rétrospective d'informations.

3.3. *Les problèmes d'analyse concernant les données textuelles*

Le fait même de substituer aux textes en langage naturel une paraphrase ou «représentation» formulée dans des termes différents suppose l'existence d'un système de représentation autonome, ou sont définis les éléments constitutifs de la paraphrase: unités lexicales d'une part, à savoir les symboles désignant les notions ou concepts élémentaires du métalangage, et conventions syntaxiques d'autre part, pour l'expression des relations logiques observées entre ces concepts dans les chaînes soumises à l'analyse. La raison d'être d'un tel métalangage tient aux anomalies connues du langage naturel, du point de vue sémantique: des termes LN différents sont tenus pour équivalents (synonymies); à un même terme LN sont associés plusieurs sens distincts (homonymies, homographies, polysémies); des tournures syntaxiques différentes sont tenues pour équivalentes quant à la relation logique sous-jacente (allotaxies); à une même tournure LN sont associées des relations logiques distinctes (homotaxies), des équivalences plus complexes sont posées entre mots et phrases d'un même langage (définitions). Il n'est pas utile de s'étendre sur le poids de ces phénomènes, évoqués dans tous les traités de sémantique, comme aussi dans les ouvrages consacrés plus particulièrement aux problèmes documentaires. Bornons-nous à souligner qu'il suffit de les reconnaître pour que soit postulée la réalité de systèmes symboliques étrangers aux langages naturels considerés, où l'irrégularité des correspondances entre signifiants et signifiés fait place à une normalisation des premiers, fondée sur l'invariance relative que l'on prête aux seconds, dans un domaine de référence donné. Que cette invariance soit relative, cela parait aller de soi: les assimilations ou dissimilations sémantiques proposées par chacun varient selon le degré de finesse qu'il assigne à l'analyse, en fonction notamment du champ d'observation concerné. Le découpage du monde empirique n'est évidemment pas le même pour l'astronome, l'archéologue ou le physicien.

Les types de problèmes sur lesquels débouchent ces reconnaissances sémantiques sont à la fois généraux par l'étendue des disciplines où ils se posent (une grande partie des sciences de l'homme) mais également importants ou du moins jugés comme tels puisque c'est précisément à leur résolution que s'attache l'ensemble des procédures connues sous le nom d'analyse de contenu. Cette exploration sémantique mobilise aussi bien des méthodes syntaxiques que mathématiques, par exemple statistiques; leur mise en oeuvre passe le cas échéant par le traitement mécanique des documents. Au niveau le plus modeste, il est facile de vérifier que la simple mise sur pied d'un *lexique* où sont levées les anomalies banales dues aux synonymies et polysémies ne va pas sans problèmes; a fortiori si l'on veut associer à chacun des termes retenus sa ou ses définitions. Il y a là

prétexte à une série de confrontations et de mises au point d'où peuvent sortir des clarifications substantielles pour le domaine considéré.

Ces quelques observations élémentaires montrent comment le jeu propre de l'analyse documentaire engendre au fur et à mesure qu'il se déroule la constitution d'un système de symboles nécessaires pour nommer d'une manière ou d'une autre les produits mêmes du jeu, sur le plan lexical et syntaxique.

Le fonctionnement satisfaisant d'un système documentaire suppose ainsi résolues un ensemble de questions auxquelles les réponses doivent nécessairement être apportées *lors de la construction du système*. Ces questions, pour nous résumer, concernent *conjointement*:

— La délimitation du corpus à partir d'un matériau initial.
— La mise au point d'un système de représentation, constituant la base sur laquelle sont exprimées l'ensemble des décisions concernant la nature et l'étendue de l'information jugée pertinente pour le domaine considéré (en vertu par conséquent d'un corpus donné, pour la génération de produits documentaires donnés).
— La mise au point parallèle d'un système de consultation, exprimant les opérations de traduction du contenu des demandes documentaires dans des termes compatibles avec ceux de l'analyse initiale, et les opérations logiques de sélection des informations recherchées.

Ce sont ces aspects que nous voudrions aborder maintenant, en décrivant assez sommairement les éléments caractéristiques d'un système documentaire particulier dont l'objectif est de mécaniser l'édition et la consultation du Corpus des Inscriptions Latines. Nous nous bornerons seulement à signaler celles de ses propriétés qui nous semblent les plus intéressantes pour notre propos, réservant pour un dernier paragraphe (§ 5), quelques remarques tendant à éclairer mutuellement les points abordés successivement dans ce paragraphe et le précédent (§ 2), du moins pour les plus importants d'entre eux.

4. Exposé sommaire des caractéristiques d'un système documentaire visant l'édition et la consultation automatiques du Corpus des Inscriptions Latines (Système SYCIL)

Le domaine de l'épigraphie latine est caractérisé par une forte tradition d'ordre documentaire. L'édition des textes et la constitution de nombreux indices et tables sont en effet l'objet de soins attentifs et répondent à des règles bien précises, affinées par una longue pratique.

Les raisons qui pouvaient inciter à envisager la mécanisation de cette documentation sont principalement:

a) La très grande masse des matériaux épigraphiques, cet aspect quantitatif jouant négativement sur la publication des instruments documentaires.

b) Le besoin, très mal assuré dans l'état actuel des publications, de dis-

poser de la totalité des informations et des références pour un matériel donné.

c) Une troisième raison tient au *principe* même des indices. Ceux-ci en effet consistent principalement en une édition tabulée et référenciée d'extraits de textes classés selon divers critères (alphabétique, conceptuel, chronologique, géographique, etc.). Or leur nombre étant nécessairement limité, toute recherche recouvrant plusieurs thèmes impose un travail de comparaison extrêmement étendu, et fastidieux.

d) Enfin, un dernier point concerne le *contenu* des indices. Il est clair que, malgré leur orientation encyclopédique, les indices ne peuvent enregistrer la totalité des éléments attestés dans l'ensemble du Corpus. Or en droit, n'importe lequel de ces éléments est susceptible de constituer une source d'information. Il s'agit donc d'envisager de passer à une conception non limitative des indices —que ce soit quant à leur mode d'organisation, ou à leur contenu, et par conséquent à leur nombre—, à condition, naturellement, que les éléments qui doivent y intervenir soient catégorisables du point de vue épigraphique.

Pour ces raisons, le recours à l'ordinateur, en vertu de ses caractéristiques (capacités de mémoires, performances, souplesse d'organisation, etc.) permet d'apporter des solutions satisfaisantes, du point de vue de la quantité de l'information à traiter, de la rapidité d'accès, des facilités d'organisation et de classement, des possibilités d'effectuer rapidement et systématiquement toutes les mises à jour nécessaires [4].

4.1. *La nature des objectifs*

Le projet de reprendre et de continuer par le recours à des moyens informatiques l'édition et la consultation du CIL correspond ainsi assez clairement à une nécessité impérieuse de la recherche, tenant aux conditions mêmes d'accès aux données épigraphiques. Cette constatation initiale peut être alors prolongée par l'examen des objectifs recherchés, c'est-à-dire par l'analyse des impératifs et des centres d'intérêt de la recherche épigraphique, en vue de définir les caractéristiques du système pouvant réaliser ces objectifs.

Ces impératifs peuvent pour l'essentiel être enumérés comme suit:

a) La nécessité fondamentale d'exprimer les diverses interprétations et restitutions apportées à l'état initial du texte. Cette nécessité découle du fait que l'exploitation du matériel épigraphique passe par un très minutieux travail de mise au point qui est déjà, par nature, un aspect de cette exploitation. Une contrainte très raisonnable est donc apportée dans l'exigence de pouvoir toujours utiliser soit *une lecture particulière* de l'inscription, soit sa *forme originelle*. Le projet de produire conjointement les deux états d'un texte est donc

[4] L'étude théorique et expérimentale ayant conduit à la construction de ce système documentaire a été menée conjointement par des techniciens, chercheurs et professeurs de l'Unité de Recherche «Archéologie et Calcul» (Centre de Recherches en Archéologie, CNRS), MM. E. Chouraqui et J. Virbel; de l'Institut d'Archéologie Méditerranéenne (CNRS), M. M. Janon, et de l'Université de Provence, MM. P. Corbier et P. A. Février.

tout-à-fait fondé, et répond bien aux *nécessités internes* de la recherche épigraphique [5].

b) La nécessité de prendre en considération, selon diverses pondérations qui doivent être précisées, les points de vue extrêmement variés à partir desquels le matériel épigraphique constitue une source d'information. Que ce soit en effet à titre de source principale ou auxiliaire, comme données initiales ou de vérification, il apparait à l'examen que les données épigraphiques peuvent intéresser un nombre très important de disciplines: histoire générale ou liée à certaines données, militaires, économiques, sociales, généalogiques, religieuses, etc.; archéologie; géographie humaine; histoire de la langue, de la littérature, etc. La diversité possible de ces points de vue récoupe la nature polyvalente des inscriptions qui, par leur aspect linguistique, intéressent tout ce qui touche à l'histoire de la langue latine, par le contenu qu'elles véhiculent, l'histoire de la civilisation, et par leur aspect matériel, plus particulièrement l'archéologie.

Les produits documentaires visés doivent donc répondre aux préoccupations centrées sur l'un ou l'autre de ces aspects.

c) Cette diversité s'exprime concrètement au niveau même des éléments tenus pour pertinents parmi l'ensemble de ceux que comportent les objets considérés. En effet, suivant les thèmes de recherche projetés, l'exploitation du matériel épigraphique peut selon les cas s'intéresser exclusivement ou, le plus souvent, concurremment, aux aspects suivants:

— Les propriétés matérielles de l'objet (son matériau, ses dimensions, etc.).
— Ses propriétés plus interprétées archéologiquement (le contexte archéologique précis de sa découverte, le support sur lequel réside l'inscription, l'existence d'un décor, la date connue ou supposée, etc.).
— Ses propriétés proprement linguistiques (existences de fautes d'orthographes, de phénomènes syntaxiques ou morphologiques particuliers, etc.).
— Ses propriétés disons «stylistiques» ou «littéraires»: utilisation de formules particulières (par exemple, funéraires, dédicaces, etc.), citations ou expressions littéraires particulières (vers, etc.).
— Ses rapports avec un «genre» épigraphique donné, que ce genre réponde à des régularités couramment observables (comme c'est le cas des inscriptions funéraires, par exemple) ou qu'il relève de certains thèmes généraux que les épigraphistes tiennent pour une manière pratique de classer les inscriptions (inscriptions funéraires, dédicaces, impériales, religieuses, etc.).

[5] C'est d'ailleurs trois et non deux versions qu'engendre finalement le texte initial de l'inscription, dans la mesure où les indices de vocabulaire recensent les termes selon une forme *canonique* (nominatif singulier pour les substantifs, par exemple).

— Ses propriétés de «contenu», c'est-à-dire le contenu propre du texte de l'inscription.
— Les propriétés orthographiques de l'expression de ce contenu (l'existence de cas nombreux d'abréviations, de réduction à des sigles, etc.).
— Les marques de «l'histoire interne» de l'objet, visible soit au niveau archéologique (dans divers remplois) soit au niveau linguistique dans des érasures, regravures, surcharges, etc.
— Les propriétés proprement «graphiques» relatives au mode de graphie du texte: types d'écriture, caractères des signes, de la disposition, unité ou multiplicité des scripteurs, etc.

Cet inventaire fait apparaitre qu'à la multiplicité des centres d'intérêt liés au matériel épigraphique, correspond une multiplicité de *niveaux*, d'*extension* et de *nature des informations* utiles qu'il comporte, et qui doivent être prises en compte.

d) Un dernier impératif peut être souligné. Cette multiplicité suggère que quelles que soient les décisions prises pour délimiter l'étendue et la nature des informations finalement retenues, la fonction des produits documentaires obtenus sera extrêmement variable suivant les thèmes de recherche. Dans le cas le plus simple, le produit documentaire —si le langage d'analyse est suffisamment fin et approprié—, peut constituer une part importante d'une interrogation particulière.

En revanche, d'autres thèmes de recherche, ou d'autres visées, peuvent exiger un travail ultérieur beaucoup plus approfondi. On peut être aussi amené à entreprendre divers classements, comptages, traitements statistiques, etc., *portant sur les réponses documentaires elles-mêmes,* et livrer ces nouveaux résultats à la réflexion historique. Or certaines de ces opérations, telles que comparaisons, comptages, classifications, exploitations statistiques, etc., sont des opérations que peut prendre en charge un ordinateur *à certaines conditions,* parmi lesquelles la définition des algorithmes représentant ces divers calculs et l'écriture des programmes correspondants sont naturellement les plus importantes. Dans cette hypothèse, il est donc nécessaire de prévoir que l'expression des données (tant du point de vue de leur contenu que de celui de leur représentation) puissent se prêter à une exploitation de cer ordre [6].

Cet examen permet de définir une première formulation des propriétés du système concernant les produits documentaires visés, la nature et l'étendue des informations à prendre en compte, et les propriétés structurelles de fonction-

[6] Cet aspect ne sera pas développé ici. Pour un exemple d'exploitation statistique de données épigraphiques, cf. M. Borillo *et al.* (1973). Pour un exposé relatif à la génération automatique de grilles descriptives à partir d'un système documentaire, cf. Igmraf. Enfin, on trouve dans Bourrelly *et al.* (1973) un exposé relatif à un système de traitement de l'information intégrant une bibliothèque de programmes scientifiques.

nement, en rapport avec les impératifs de la recherche épigraphique. Il convient d'examiner alors comment ces objectifs se traduisent au sein du système lui-même.

4.2. *Matérialisation des objectifs au sein du système documentaire*

a) *Nature des produits documentaires visés*

Les produits documentaires sont de deux ordres:

1 Des éditions *in extenso* des textes, en trois versions:

- Une version «brute» (B), correspondant à la version attestée.
- Une version «interprétée» (I), comportant un ensemble d'interprétations et de restitutions (mutilations, abréviations, etc.).
- Une version «canonique» (C) ou toutes les occurrences sont ramenées à des formes morphologiquement canoniques.

2. Des réponses documentaires correspondant à une généralisation des indices, et réalisant les possibilités suivantes:

- Interrogation du système sur la partie «textuelle» de l'inscription et/ou sur sa partie «objectuelle» (propriétés matérielles, archéologiques, etc.).
- Recherche de la présence ou de l'absence, et éventuellement extraction et édition de toute phraséologie particulière, dénotée soit par son énoncé, selon l'une des versions B, I ou C, soit au contraire par une notation conventionnelle (par exemple: EMPEREUR, OFFICIER, EDIFICE, etc.) pouvant être manifesté dans les textes selon diverses phraséologies.
- Possibilité d'éditer les produits de ces recherches selon des conventions de *classification* (ordre alphabétique des phraséologies, ordre conceptuel des éléments de contenu, ordre chronologique ou géographique des objets, etc.) de *hiérarchisation de ces classements,* et de *dispositions typographiques* (selon les axes lignes et colonnes) librement fixées par l'utilisateur.

b) *Contenu des produits documentaires*

Le contenu souhaité des produits documentaires permet de définir au sein des objets la *nature,* l'*étendue* et l'*organisation* des informations qu'il convient de repérer et de marquer lors de l'analyse documentaire préalable à l'enregistrement des données. Ces informations, comme on l'a vu sont de nature fort diverses, et mobilisables concurremment au sein d'une question.

1. Un premier groupe concerne le «contexte» du texte, c'est-à-dire les aspects matériels et archéologiques de l'objet qui porte ce texte (datation, lieu

de découverte, de conservation, situation de remploi, dimensions, matériau, support, décor, etc.) [7].

2. Un second groupe concerne le «texte» proprement dit, et celui-ci se subdivise selon qu'il concerne la *lecture* du texte (il s'agit alors de ses propriétés graphiques et ortographiques, au sens le plus général) ou son *interprétation* (en terme de contenu, de faits de langue, etc.), ces deux opérations étant naturellement *solidaires* dans le processus d'analyse, mais pouvant constituer des niveaux isolables *quant à leurs resultats* (cf. ci-dessous).

c) *Structure du système*

Le système documentaire répondant à ces préoccupations suppose donc que chacun des deux temps de la représentation des données, puis de leur consultation reçoivent l'ensemble organisé des spécifications aptes à lui permettre de remplir les tâches visées.

La représentation, tout d'abord, n'est assurée qu'au terme d'une analyse des données comportant les quatre aspects principaux suivants (en ce qui concerne seulement la face textuelle de l'objet).

1. Une *lecture* du texte, effectuée par l'épigraphiste, et visant à créer les conditions intellectuelles et matérielles de la réussite des opérations suivantes. A ce niveau, c'est tout le *savoir* du lecteur qui est sollicité pour obtenir à partir de la version attestée sur l'objet l'ensemble des éléments qui seront ensuite mis en jeu. Il est évident que dans le cas considéré, les composants fondamentaux de ce savoir concernent d'une part la maîtrise de la langue latine, aussi bien selon un ordre général de connaissance que du point de vue des habitudes graphiques et discursives des rédacteurs d'inscriptions, et d'autre part, une culture historique centrée sur la civilisation romaine et allant jusqu'aux aspects plus proprement épigraphiques selon lesquels elle se manifeste.

2. Une *transcription* du texte visant à créer une version particulière dite «version d'enregistrement», sur laquelle s'appuiera le programme chargé de générer automatiquement les trois versions d'édition souhaitées. A ce niveau, il s'agit par conséquent de réaliser cette version d'enregistrement en manipulant simultanément.

— La version attestée et la «lecture» initiale qu'on lui associe.
— Un *langage de transcription* réalisant le passage de la version attestée à la version d'enregistrement.
— Un ensemble de *règles de transcription* qui définissent sur un mode opératoire, les instructions qui doivent être suivies pour la réalisation de ce passage. Ces règles, qui peuvent être vues comme un «mode d'emploi» du langage de transcription intègrent donc de manière unitaire les divers éléments provenant aussi bien d'une certaine mise

[7] Cet aspect ne sera plus considérée ci-dessous, où nous nous limiterons à la seule face «textuelle» de l'objet «inscription».

en oeuvre d'un savoir spécialisé que des configurations textuelles des objets analysés, et de l'algorithme de traitement sous-jacent au programme de génération des versions.

3. Une *indexation* [8] qui vise à associer aux éléments des textes des informations permettant lors de la consultation de générer des indices de tous ordres. Cette indexation elle-même comporte deux aspects, *liés* dans la pratique:

— Une *segmentation,* qui a pour but de découper le texte en unités particulières.
— Une *caractérisation* qui consiste à associer à chaque segment les informations d'ordre sémantique, syntaxique, stylistique, etc., nécessaires à la production des indices. L'indexation est assurée par un *langage d'indexation* qui lui même comporte deux éléments:
— Un *lexique,* qui recense les termes conventionnels, appelés *descripteurs,* permettant de caractériser les segments du texte transcrit. Ces descripteurs comportent de plus une organisation logique, visant à formaliser une partie du réseau notionel de l'analyste.
— Une *syntaxe,* constituée par un ensemble des relations logico-syntaxiques susceptibles de lier certains segments ou descripteurs entre eux.

La caractérisation du texte transcrit segmenté à l'aide du langage d'indexation est effectuée grâce à des *règles d'indexation* qui régissent les modalités de cette caractérisation de manière à la rendre *univoque* et *régulière,* conditions indispensables au fonctionnement satisfaisant ultérieur de la recherche documentaire (et qui sont aussi en jeu pour la transcription).

4. Une *représentation* proprement dite des résultats des opérations antérieures, qui traduit, selon des conventions formelles strictement définies ces résultats sous la forme d'un bordereau d'enregistrement.

Symétriquement, la consultation du système documentaire est assurée au terme d'un processus d'analyse recouvrant les aspects principaux suivants:

1) La *formulation* d'une demande d'information, telle que cette demande soit *pertinente* pour le système documentaire. Cette opération suppose non seulement une culture historique et épigraphique du même ordre que celle qui joue lors de l'analyse, mais aussi la connaissance des spécifications du système, pour ce qui concerne la nature du corpus traité, des informations retenues, la forme documentaire selon laquelle elles sont restituables.

2) A un second niveau, il s'agit de traduire le contenu de cette demande sous la forme d'una *question documentaire* proprement dite, bien-formée par rapport à des conventions que fixe le système de consultation. Celui-ci comporte:

[8] C'est tout à fait conventionnellement que l'on peut parler de «transcription» et d'«indexation», la distinction marquant que, au niveau des résultats livrés par l'analyse documentaire, on représente le document *à deux niveaux de représentation.*

— Des règles d'expression du contenu des questions selon des conventions qui rendent ces dernières *comparables* pour l'automate avec la représentation des documents enregistrés.
— Des règles de classification (critères de classification, hiérarchisation) des éléments indiqués dans les questions.
— Des règles d'édition, fixant le mode de tabulation de ces éléments.

Nous n'entrerons pas dans le détail du système de consultation ci-dessus, mais nous nous bornerons à donner quelques aperçus du système de représentation, en nous limitant à la transcription et à l'indexation.

4.3. *Le système de représentation (transcription et indexation)*

a) *Le langage de transcription*

Celui-ci comprend des *codes* et des *opérateurs* ayant les propriétés suivantes:

1. Chaque code a une *signification particulière*, bien précise, et correspondant à la désignation d'un phénomène reconnu dans la version attestée: abrègements, fautes d'orthographe, incomplétudes, mutilations diverses, etc.
2. Chaque code possède une *symbolisation* particulière (marque[s]).
3. Du point de vue de leur structure, on peut distinguer trois types de codes (appelés conventionnellement):

— Ponctuels, possédant une *marque*.
— Segmentaux simples, possédant deux marques, droite et gauche enfermant un texte.
— Segmentaux complexes, comportant trois marques, droite, centrale et gauche, et deux textes, droit et gauche, vis à vis de la marque centrale.

4. Les opérateurs, au nombre de trois, respectivement: «suppression et tassement» (—), «maintien» (+) et «remplacement par un blanc» (/) représentent les opérations de traitement effectuées par le programme de génération des versions B, I et C pour chaque élément (marque[s] et/ou texte[s]) de chaque code.

b) *Les règles de transcription*

La part la plus formalisée des règles de transcription est constituée par deux éléments:

1. Un *schèma de versions* qui indique les affectations des opérateurs aux éléments des codes pour chaque version B, I et C.
2. Des *instructions* relatives aux configurations que ces codes peuvent ou ne peuvent pas présenter lors de la transcription en vertu des opérations effectuées par le programme de génération (elles interdisent, par exemple, le chevauchement de marques de codes segmentaux, n'autorisant que leur succession

ou leur inclusion). L'algorithme prévoit le traitement en priorité des codes segmentaux (simples ou complexes), et parmi seux-ci, d'abord de ceux qui présentent des emboîtements; puis, celui des codes ponctuels, inclus ou non dans des segments isolés par les premiers. L'emboîtement de codes segmentaux est repérable par une suite de parenthèses fermées (seule la marque gauche des codes segmentaux est distinctes, la marque droite étant dans tous les cas une parenthèse fermée), ou par la présence, après une marque ouvrante (droite), d'une autre marque ouvrante (dans l'ordre de lecture) avant que la parenthèse fermée (marque fermante droite) du premier code ne soit rencontrée.

c) *Le langage d'indexation*

L'indexation est assurée par un langage d'indexation, comprenant un lexique organisé de descripteurs et une syntaxe, et par un ensemble de règles qui fixent les modalités de l'indexation.

1. *Le lexique*

Il comprend l'ensemble des descripteurs nécessaires pour l'indexation des textes, c'est-à-dire finalement, pour l'extraction, la tabulation organisée et l'édition de tout fragment de texte. Nous ne pouvons malheureusement pas nous étendre sur le contenu de ce lexique. Disons rapidement qu'il comprend une vingtaine de chapitres (pour ce qui concerne les textes seulement; le contexte donnant lieu à une autre section), couvrant l'ensemble des faits de contenu et de langue qui ont été *jugés pertinents pour la recherche*. Il s'agit, par exemple, des chapitres suivants [9]:

— PERSONNES (J): types sociaux de personnes attestées dans les textes: MAGISTRATS et HAUTS FONCTIONNAIRES, MILITAIRES, PRETRES, MAGISTRATS MUNICIPAUX, TITULAIRES et DIGNITAIRES, PERSONNES SANS QUALITE, etc.
— EMPEREURS (M), qui comprend l'ensemble des empereurs et des membres de leurs familles. Les EMPEREURS sont classés en «dynasties»: AUGUSTE (MA), JULIO-CLAUDIENS (MB), FLAVIENS (MC), etc., et à l'intérieur de chaque classe, par ordre chronologique: TIBERE (MBA), CALIGULA (MBB), CLAUDE (MBC), etc.; les membres de la famille d'un empereur sont représentés par des codes apparentés à celui de l'empereur concerné: DRUSUS IVLIUS CAESAR (MBAA), DRVSUS CAESAR (MBAB), TIBERE CAESAR (MBAC), etc.

[9] La signification conventionnelle des descripteurs est dénotée par des majuscules, leur symbolisation est notée entre parenthèses. On remarquera que l'organisation conceptuelle de ces descriptions est marquée par l'emboîtement alphabétique de leur symboles —ce parti permettant d'éviter l'enregistrement explicite des relations entre descripteurs ce qui est nécessaire lorsque leur symbolisation est arbitraire de ce point de vue.

— DENOMINATION DES PERSONNES (K) où sont indiqués les éléments de désignation: PRENOM, COGNOMEN, GENTILICE, TRIBV, ORIGO, SIGNUM, SEXE, etc.
— PARENTE: PATER, MATER, FILIVS, FILIA, etc.
— DESIGNATION DES EMPEREURS (N): IMPERATOR, CAESAR, FILIATION, NOM, AUGUSTUS, EPITHETES, etc.
— DIEUX ET DIVINITES (O) classés en DIEUX PRINCIPAUX (OA), ALLEGORIES ET PHENOMENES NATURELS (OB), ALLEGORIES MORALIA (OC), ALLEGORIES GEOGRAPHIQUES (OD), etc.

2. *Les relations* [10]

Les relations nécessaires sont de deux ordres et concernent:

— L'ordre linéaire du texte (qu'il est nécessaire de préserver en vue d'éditions et des questions portant justement sur des structures phraséologiques).
— Le contenu relationnel proprement dit du texte (relation de sujet à objet d'une action, de parenté, etc.).

d) *Les règles d'indexation*

Leur existence est rappelée ici pour mémoire, car il n'est naturellement pas possible d'entrer dans leur détail. Disons, en simplifiant, que les fonctions qu'elles doivent assurer consistent fondamentalement à fournir à l'analyste les moyens de réaliser *régulièrement* l'indexation du document en mettant en quelque sorte en regard son savoir, le document particulier auquel il est confronté, et le langage d'indexation dont il dispose. La segmentation et la caractérisation des segments du texte au moyen des termes du langage d'indexation, définissant les deux aspects cruciaux de ce travail, comme nous l'avons déjà dit. Les règles régissant ces deux opérations doivent comme il advient aussi au niveau de la transcription, définir un «mode d'emploi» opératoire du langage d'indexation vis-à-vis des phénomènes susceptibles d'être rencontrés lors de l'analyse, sous peine d'ambiguités, et par conséquent de dysfonctionnement du système documentaire lui-même.

[10] Nous ne considérons ici que les relations «syntagmatiques», c'est-à-dire explicitement attestables entre des éléments de représentation (descripteurs, ou segments de textes transcrits); celles-ci se distinguent des relations «paradigmatiques», tout aussi explicites, mais prédéfinies au sein du lexique. Elles s'opposent globalement, à toute la partie relationnelle du réseau conceptuel du spécialiste *non explicité au sein du système*. D'autre part nous n'évoquerons que le *contenu* de ces relations, et non leur expression, laquelle peut être fort variable (lien explicite sous forme d'un syntagme du type aRb, convention d'enregistrement, par exemple pour l'ordre mutuel des termes d'un texte, co-affectation de descripteurs à un même segment, etc.).

5. Relations entretenues entre les propriétés et les fonctions du système de représentation

Nous nous proposons de conclure cette étude par quelques remarques plus propres à établir un pont entre les deux développements précédents, et à éclairer les points les plus fondamentaux justifiant mutuellement, pensons-nous, la démarche suivie dans l'entreprisse documentaire relatée et le cadre méthodologique auquel elle se réfère. Nous organiserons ces remarques autour de deux pôles de réflexion, concernant le système de représentation (contenu, organisation, rôle, § 5.1.) et les conditions de validation de la démarche suivie (§ 5.2.).

5.1. *Le contenu, l'organisation et le rôle du système de représentation*

Le contenu et l'organisation propres du système de représentation correspondent en dernière analyse aux conventions fixées pour représenter les informations issues de la classe particulière de textes, les inscriptions latines, informations sélectionnées et caractérisées de manière à ce qu'une procédure mécanique puisse ensuite créer, selon des opérations définies, divers documents (produits documentaires).

5.1.1. *La définition du corpus*

En ce qui concerne la *classe des textes concernés,* on peut signaler la situation suivante.

La plus grande difficulté existe à définir de manière stable les objets mêmes auxquels s'intéresse l'épigraphie latine. Les «inscriptions», en effet relèvent beaucoup plus d'une définition empirique liée à la pratique épigraphique (en ce sens est une inscription tout texte qui est traditionnellement incorporé dans le Corpus) que d'une définition explicite et stable. En particulier, ni les supports, ni le contenu, ni les types de personnages attestés, ni les auteurs, ni la finalité ne peuvent semble-t-il fournir les éléments d'une telle définition. De sorte qu'initialement, on est conduit à mettre relativement entre parenthèses la nécessité de préciser quels sont les objets qui relèvent du corpus envisagé. En revanche, après l'examen des différents aspects informatifs pertinents des objets tenus pour des inscriptions, il est possible *en retour* de fixer (au moins provisoirement) une définition opératoire du corpus. Par exemple, la liste des matériaux recensés, ou celle des supports ou certains éléments de contenu offrent à l'aide d'une combinaison de traits de ce type, une sorte de «seuil» en deçà ou au delà duquel un objet est normalement tenu ou non pour une «inscription».

5.1.2. *Délimitation de l'information pertinente*

En ce qui concerne les informations retenues, c'est évidemment la projection des objectifs documentaires propres de la discipline épigraphique sur les

textes qui préside fondammentalement aux choix effectués [11], aussi bien à propos de la réduction opérée sur le matériel que sur les modalités de représentation.

1) Cette situation nous semble particulièrement claire en ce qui concerne les dispositions prises à propos de la transcription des textes.

Comme nous l'avons vu, les conventions de rédaction suivies par les épigraphistes se réfèrent à plusieurs points de vue intervenant dans la détermination et le sens de ces conventions:

— La difficulté de lecture, que le texte soit tronqué, ou que les signes aient subi diverses injures. Une distinction est opérée entre ce qui reste identifiable, malgré des dommages, et ce qui n'est restituable qu'avec incertitude ou ne l'est pas. L'état de conservation interfère par conséquent avec la lisibilité, et cette dernière avec le niveau de conjecture exigé.
— Le traitement de diverses abréviations et de symboles particuliers représentant des graphies plus développées.
— La marque de ces symboles (monogrammes, symboles abréviatifs non alphabétiques).
— Les séparateurs (points, dessins variables, passages à la ligne, etc.).
— Les irrégularités orthographiques (ommision, répétitions, confusions de lettres).

Comme on le voit, le premier point ci-dessus est lié à l'*histoire de l'objet*, et à sa conservation actuelle, tandis que les suivants sont liés aux habitudes et aux *pratiques mêmes des producteurs d'inscriptions*.

Ces conventions de présentation constituent fondamentalement un traitement de l'*incomplétude* et de la *symbolisation*.

Sur un plan général, il apparaît assez bien que le traitement de l'incomplétude peut consister soit à *signaler* un phénomène, sans le décrire (par exemple, une lettre est abimée, mais reste identifiable; on signale qu'elle est abimée, sans décrire cependant comment), soit au contraire à *signaler et à décrire*, selon des modalités fixées, un phénomène particulier (par exemple, l'existence d'une abréviation est signalée, et décrite par le développement du texte abrégé dans la transcription). Or rien ne prédispose «naturellement» les phénomènes de mutilation et d'abrégement à connaître tel ou tel type d'analyse; et ce n'est

[11] La *pertinence* de l'information, comme critère majeur de choix peut d'ailleurs être diversement contrariée dans les faits par d'autres considérations, telles que, per exemple, *les difficultés d'accés* à l'information, si les représentations visées exigent une *observation directe* des objets, ou d'une reproduction (photographique par exemple), effectués selon des conventions à préciser (et que la construction du système documentaire permet de préciser). Ou la *crédibilité* de certaines données (ce qui peut concerner certaines données contextuelles dans le CIL). Ou encore les *difficultés rencontrées à formaliser* certains aspects de l'analyse, ce qui en épigraphie latine concerne tout particulièrement les données paléographiques, du moins actuellement.

au contraire qu'un ensemble de décisions, justifiables vis-à-vis des impératifs de la recherche épigraphique, qui permet de fixer les termes de cette analyse.

Cet exemple est illustratif de la structure interne du système de représentation, et de son rôle par rapport au savoir épigraphique. Toujours en ce qui concerne le cas d'une «lettre abîmée», il est clair qu'il est nécessaire qu'au moins empiriquement les épigraphistes soient d'accord pour définir une sorte de «seuil d'identification», ce qui peut poser des problèmes, par exemple, dans le cas de lettres graphiquement voisines, telles que B, P, R, ou E, F, etc.

De la même manière, on doit définir sur l'axe continu des écarts par rapport à la norme orthographique un seuil séparant les cas de fautes et ceux d'abréviations [12]. Des *règles* d'analyse doivent par conséquent expliciter les usages habituellement suivis par les épigraphistes, usages qui sans doute circonscrivent un seuil empirique de reconnaissance des signes, et où l'entourage est appelé à jouer un rôle plus ou moins important. De la même manière on doit se mettre d'accord sur ce qu'est un «registre» dans une inscription, c'est-à-dire finalement indiquer explicitement selon quels critères on décide que l'on passe d'un texte à un texte différent sur le même support (différences de contenu, d'écriture, changement de face du support, marques formelles de séparations, etc.) [13].

2) Le même ordre de question intervient à propos du langage d'indexation. Les objectifs poursuivis imposent, par exemple, de reconnaître au sein des textes les mentions des empereurs romains. Or il apparaît à l'examen qu' *épigraphiquement,* la catégorie d'empereur est beaucoup trop macroscopique telle quelle, et que si l'on tient compte des types de questions documentaires qui peuvent la concerner, il est nécessaire de la décomposer en divers éléments constitutifs. On peut tout d'abord en reconnaître trois principaux: le nom de l'empereur, ses fonctions et titres, et les mentions résiduelles (résiduelles de ce point de vue). Chacune de ces classes peut encore faire l'objet de subdivisions; pour la première, par exemple, elle peut comporter, la filiation, et le nom, qui lui-même se décompose en prénom, gentilice, cognomen, etc.

De sorte que l'on est conduit, au moment de la constitution du système de représentation à définir une sorte de schéma théorique de la désignation des empereurs romains, à partir duquel on peut ensuite observer et décrire les désignations effectivement attestées, puis rédiger divers produits documentaires sur la base d'une sélection et d'un classement des éléments du schéma.

Dans le cas considéré, ce schéma consiste ainsi en une organisation des

[12] Par exemple, les épigraphistes considèrent que l'absence de la lettre N dans COS, pour CO(N)S(UL) ne constitue pas une faute de graphie (lettre manquante), mais fait au contraire partie du mécanisme d'abrègement du terme CONSUL. Plusieurs raisons peuvent naturellement justifier cette interprétation (grande fréquence de cette réalisation; tendance au monogramme; etc.).

[13] Il convient d'ailleurs de préciser que ces règles n'ont pas pour but de fixer définitivement une pratique épigraphique, ni de trancher parmi des usages divergeants. Leur rôle consiste seulement à fixer des conventions, qui seront respectées lors de l'analyse, et sur lesquelles on peut par conséquent s'appuyer lors de la consultation documentaire.

éléments pouvant intervenir dans la désignation d'un empereur, une question donnée pouvant alors se traduire dans les termes de certains ou de tous les éléments de cette désignation. Cette organisation reflète des groupements d'éléments informatifs tenus pour solidaires, et pouvant éventuellement être dénotés conjointement lors de la consultation, la désignation de ces groupements constituant lui-même un élément d'information, connu du système.

Par ailleurs, à côté de cette description des éléments lexicologiques intervenant dans la désignation des empereurs, on doit aussi posséder, au sein du système de représentation, une liste des empereurs eux-mêmes, désignés conventionnellement par un descripteur propre. Et ce pour deux raisons. La première réside dans le fait que des questions documentaires peuvent porter sur un empereur donné (par exemple, Trajan), indépendemment de toutes les variantes phraséologiques selon lesquelles cet empereur est désigné dans les textes (on peut dire que le rapport entre le concept désignant un empereur donné et l'ensemble des phraséologies qui constituent ses désignations effectives est voisin de celui qui existe entre un radical et ses différentes variantes morphologiques). La seconde tient au parti même de représenter les éléments de désignation indépendemment de leur sens vis-à-vis d'un empereur donné, parti auquel nous avons dit que l'on était conduit afin de répondre à certains objectifs documentaires. En effet, rien dans le schéma ne signale que les éléments attestés dans une désignation particulière constituent solidairement la désignation d'un même personnage. Et lorsque plusieurs désignations d'empereurs interviennent dans une même inscription (ce qui est, par exemple, le cas lorsqu'un empereur est le fils d'un autre empereur et que sa filiation est indiquée) il est absolument nécessaire de marquer cette relation. On est donc conduit à indiquer d'une part une relation de «solidarité» entre les descripteurs représentants les éléments de désignation d'un empereur, et d'autre part une autre relation (disons d'«identité») entre cet ensemble solidaire et le descripteur représentant l'empereur particulier attesté.

On voit donc qu'en définitive, la catégorie initiale d'«empereur romain», comprise *dans sa réalité épigraphique, et en fonction des objectifs documentaires visés* correspond à un ensemble de dispositions au sein du système de représentation comprenant:

— Une liste des éléments de désignation (du type IMPERATOR, CAESAR, NOM, CONSUL, etc.).
— Une organisation de ces éléments.
— Une relation marquant la co-référence des éléments de désignation attestés (dite ci-dessus de «solidarité»).
— Une liste des empereurs romains.
— Une relation marquant l'«identité» entre les éléments de désignation et un empereur particulier.
— De plus, pour des raisons comparables à celles qui suggèrent de nantir les éléments de désignation d'une organisation, les listes des

empereurs doivent elles-mêmes être organisée (en fonction de leur chronologie, et de groupements de type dynastique utilisés habituellement par les historiens: Julio-Claudiens, Flaviens, Antonins, etc.) afin là aussi de simplifier la consultation.

Ainsi le niveau de finesse de décomposition et les dispositions prises pour la représentation des «empereurs romains» sont fixées en dernier ressort par les besoins mêmes de la recherche documentaire, exprimant les impératifs qui président à la conception du système [14].

5.1.3. *A propos des aspects linguistiques*

Les modalités selon lesquelles ont été prises en compte les propriétés linguistiques des textes épigraphiques nous semblent aussi particulièrement illustrative de cette situation. Elles méritent un commentaire particulier, car elles éclairent de plus un point méthodologique important (cf. § 3.3). Quatre remarques nous semblent devoir être faites à ce propos:

1) Tout d'abord, les choix qui ont été faits résultent d'une pondération initiale apportée aux différents points de vue, provenant de disciplines différentes pouvant intervenir dans l'analyse des données épigraphiques. Diverses raisons ont en effet conduit à centrer cette analyse par rapport aux disciplines historiques (archéologie, histoire romaine), et à envisager par conséquent les différents aspects des objets par rapport à cette focalisation. De sorte que comme on le comprendra, non seulement le système en question ne vise pas la totalité du domaine latin (LE latin), mais au contraire, ce n'est que *par rapport aux impératifs de la recherche archéologique et historique* projetés sur le matériel épigraphique que les propriétés linguistiques ont été éxaminées.

2) Ceci posé, on a donc opéré un choix parmi les segmentations (en radicaux, désinences, mot, propositions, etc.), et catégorisations (catégories grammaticales, cas et fonctions, genre, nombre, temps, personnes, etc.) qui pourraient être effectuées sur les textes d'un point de vue linguistique.

Par exemple, la *morphologie* latine est traitée formellement au sein du système, mais elle n'est pas *interprétée* (en termes de cas, genre, nombre, etc.),

[14] Ce qui vient d'être dit des empereurs, pour l'analyse et la représentation desquels nous avons reconnu la nécessité d'un schéma théorique de désignation suggère que le même type de schéma, en tant que référence organisée, pourrait être nécessaire, ou du moins fort utile pour d'autres éléments lexico-conceptuels des inscriptions, voire pour des classes restreintes d'inscriptions *dans leur ensemble*. Ce problème ne peut être résolu qu'empiriquement, la solution dépendant visiblement d'un certain rapport entre les *objectifs* de la recherche épigraphique et un *état donné de ses connaissances* à propos du matériel épigraphique. Il est évident, par exemple, que si l'on possédait un modèle des inscriptions funéraires (ou d'une classe particulière de celles-ci), l'ensemble des problèmes de représentation se poseraient dans un contexte différent. Il convient de remarquer néanmoins que certains des produits documentaires peuvent constituer la matière première d'une étude de ce type (cf.: M. Borillo *et al.*, 1973, pour un exemple particulier; et ci-dessous § 3.2, à propos du statut des objectifs scientifiques par rapport au système documentaire).

dans la mesure où l'on a besoin de connaître les différentes variantes d'un même lexème, mais où la signification morpho-syntaxique de ces variantes *ne correspond pas*, dans le domaine épigraphique et *pour la visée choisie, à des demandes d'informations autonomes*.

Pour les mêmes raisons, seules certaines segmentations (en radicaux et désinences, mots, et phrases) ont été prises en considération. On peut observer cependant que d'autres segmentations, ne provenant pas directement d'une procédure linguistique, sont en revanche nécessaires —comme, par exemple, celle qui est opérée par les passages à la ligne ou celle que projette le lexique du langage d'indexation sur le contenu des textes.

On peut constater de plus que ces caractérisations interviennent de manière diversifiées dans le fonctionnement du système documentaire. Ainsi la segmentation en mot est nécessaire pour la composition des versions I et C, mais elle est ignorée pour la version B; à l'inverse, la segmentation en ligne est attestée pour les versions B et I, et supprimée pour la version C.

3) De plus, non seulement les spécificités reconnues du domaine et de la visée propres délimitent l'étendue de la prise en compte des propriétés linguistiques, et fixent leur pertinence, relative aux objectifs visés, mais encore elles définissent les modalités de leur expression au sein du système de représentation. Par exemple, les erreurs orthographiques sont *signalées* et *décrites* lors de la transcription, en termes de lettres répétées en trop, manquantes restituées, ou substituées, et *caractérisées*, lors de l'indexation, dans les termes du chapitre FAITS de LANGUE; selon le cas: CONSONNES SIMPLE POUR DOUBLE, ABSENCE DE CONSONNE FINALE, ERREUR DE DECLINAISON, ANOMALIE DE RECTION DE CAS DE PREPOSITIONS, etc.

Là encore, ce n'est que l'examen des éléments entrant en jeu dans le système de représentation qui définit un choix de cet ordre, ou de tout autre jugé plus pertinent pour le propos recherché.

4) Il nous semble nécessaire de signaler ici pour mémoire seulement un autre aspect de la démarche décrite, aspect que l'on pourra si l'on veut appeler linguistique, dans un sens particulier du terme. Certains exemples antérieurs nous ont permis d'apprécier que si la construction du système documentaire ne va pas sans tout un ensemble de décisions concernant ses proprietés de tous ordres, un moment particulièrement important du travail sous-jacent à ces décisions concerne les mises au point visant la *terminologie propre* du domaine considéré, et à travers celle-ci, les concepts et les opérations épigraphiques (et les relations qu'ils entretiennent avec les objets concernés). Rappelons simplement ce que nous avons dit à propos de la «mutilation» d'une lettre, de la définition d'une «inscription» ou d'un «registre» ou de celle d'un «empereur romain». Cette élucidation systèmatique du réseau conceptuel sous-jacent au travail épigraphique, qui peut être abordé par l'examen du champ lexico-sémantique des spécialistes, est absolument nécessaire pour obtenir la stabilité voulue des représentations.

5.2. *Problèmes liés à la validation de la démarche et des résultats*

Dans la cadre d'un système documentaire, où l'opération fondamentale consiste finalement à retrouver, en fonction de divers tris, des informations préalablement enregistrées, les procédures de validation de la démarche suivie et des résultats obtenus sous semblent pouvoir être définies à deux niveaux.

a) A un premier niveau, il s'agit de vérifier la cohérence interne du système de traitement de l'information. C'est la *stabilité des représentations obtenues* au terme de l'analyse documentaire, et la *pertinence des produits documentaires obtenus* qui fournissent ici un critère, expérimental, de jugement. Nous n'insisterons pas sur cet aspect, car il set assez évident que la réussite à ces deux types de tests est une condition indispensable.

b) A un second niveau, il s'agit de vérifier l'adéquation entre l'ensemble des dispositions prises on vue de réaliser le système documentaire et les objectifs scientifiques de la discipline considérée. En principe, comme nous l'avons précisé, l'examen des objectifs initialement visés lors de la construction d'une chaîne de traitement de l'information constitue une phase cruciale de cette construction. Et nous avons décrit l'ensemble de la construction du système comme une sorte de traduction de ces objectifs sous la forme de propriétés données du système. Cependant, dans cette «traduction», les objectifs préalablement reconnus sont loin de rester intacts et immuables. Non seulement l'épigraphie possède son rythme propre de développement, et peut sans doute connaître comme beaucoup de sciences et disciplines historiques des évolutions ou des mutations quant à ses objectifs, son objet, ses modes de raisonnement, etc. On n'a naturellement aucune garantie particulière que les objectifs fixés, à une époque donnée, pour un système de traitement de l'information, ne soient pas tenus ultérieurement pour restrictifs, voire caducs partiellement. Mais surtout, le recours à l'informatique pour le traitement de l'information peut lui-même introduire des dimensions nouvelles pour les objectifs d'abord envisagés. D'une part les capacités de calcul (au sens le plus général du terme, qu'il s'agisse de tris et de comparaisons, ou de calculs d'ordre numérique ou logique), des ordinateurs peuvent suggérer certaines recherches en épigraphie comme ailleurs, peu envisageables, sinon concevables, autrement. Ce contexte peut se manifester tout particulièrement dans un domaine comme l'épigraphie latine qui s'appuie sur une longue tradition documentaire où les indices du CIL constituent en définitive un cas très particulier d'extraction et de tabulation de textes, et où le travail d'analyse nécessaire à leur fabrication mécanique permet d'envisager de constituer bien d'autres tabulations. Il semble que la conception qui a présidé à la constitution du CIL, dans la seconde moitié du XIX^{ème} siècle, a joué implicitement un certain rôle dans la définition même du travail documentaire, et plus largement proprement scientifique de l'épigraphie. A partir du moment où les objectifs initialement fixés induisent la possibilité d'engendrer des indices beaucoup plus diversifiés, on conçoit que ces objectifs

eux-mêmes peuvent subir en retour des modifications, et imposent à terme de nouvelles tâches à un système de traitement de l'information.

D'autre part, la formalisation de l'analyse provoque certainement un effet propre, dont les retentissements ne sont pas nécessairement perceptibles dans l'immédiat. Nous avons indiqué à propos de plusieurs problèmes comment la nécessité de formaliser l'analyse, qui est imposée initialement par le projet de mécaniser certaines phases de la documentation, réfléchit en quelque sorte un ensemble de questions vers le matériel en cause, et à travers lui, vers l'épigraphiste. Dans cette sorte de confrontation entre l'objet dont on doit formaliser l'analyse, le savoir de l'épigraphiste sur cet objet, et les contraintes apportées par le formalisation, la réflexion apportée dans la prise des décisions arrêtées à un moment donné, et le contenu même de ces décisions peuvent constituer l'occasion d'un examen approfondi des conditions mêmes de la recherche épigraphique. Et cet examen peut naturellement lui-même livrer des résultats ou des hypothèses qui retentissent sur les fonctions initialement assignées au traitement de l'information.

Il est évident que par ces considérations, on sort du cadre de procédures de validation, *stricto sensu*. Mais nous pensons en revanche que celles-ci tirent en dernier ressort leur signification profonde d'une confrontation entre les *moyens* et les *objectifs* d'une discipline donnée, et qu'il y a là aussi un niveau méthodologique qu'il faut reconnaître [15]. Cette perspective en tout cas éclaire d'une manière que nous pensons assez satisfaisante le rôle fondamental que nous accordons à l'existence explicite —c'est-à-dire communicable, et par conséquent éventuellement criticable et perfectible— d'un système de représentation des données analysées.

Il conviendrait, et ce sera notre dernière remarque, de situer le rôle effectif des moyens informatiques dans cette perspective.

La construction d'une chaîne de traitement de l'information, et plus généralement, le recours à l'informatique à propos d'un matériel scientifique particulier, suppose qu'existe un certain accord entre les spécialistes de ce domaine, accord concernant le plus visiblement les *objets* concernés et les *propriétés* recélées par ces objets. Moins visiblement, mais tout aussi impérieusement, cet accord doit porter sur les modes d'exploitations scientifiques de ce matérieu. Dans le travail relatif à la reconnaissance et à la fixation de cet accord entre spécialistes —travail qui à tous égards nous paraît scientifique—, un certain nombre de problèmes de méthode doivent nécessairement être abordés, et surtout recevoir une solution explicite. Le travail d'interpretation à quoi revient l'analyse des documents et l'établissement des données d'analyse recouvre, comme il est naturel, des opérations extrêmement variées; ces opérations doivent recevoir une définition explicite au moins jusqu'au point où leur application atteint une stabilité suffisante.

De sorte que l'*utilisation réfléchie* des moyens que l'informatique rend dis-

[15] Cf. M. Borillo, 1974.

ponibles débouche sur des préoccupations qui par nature sont d'ordre méthodologique, et concernent fondamentalement la nature du raisonnement et de ses opérations, dans le contexte d'une évolution vers un statut moins empirique. Ces préoccupations sont en définitive, nous semble-t-il, constitutives de l'attitude scientifique elle-même. Si le recours aux moyens offerts par l'informatique, tout particulièrement en ce qui concerne les données textuelles, suppose que certaines conditions soient remplies, ce recours peut créer en retour des conditions nouvelles de réflexion sur les méthodes et les objectifs scientifiques proprement dits, si du moins on a reconnu tout ce que peut avoir de néfaste la dissociation entre l'outillage informatique et le contexte intellectuel de son utilisation.

RÉFÉRENCES

Borillo, M., 1970, *La vérification des hypothèses en archéologie: deux pas vers une méthode,* «Archéologie et Calculateurs», CNRS, Paris.

Borillo, M., 1971, *Construction d'un raisonnement déductif au moyen de la simulation d'un travail archéologique traditionnel.* A paraître dans «Archéologie et Calcul», UGE, Collection 10 × 18, Paris.

Borillo, M.; Fernández de la Vega, W.; Guenoche, A.; Janon, M., et Virbel, J., 1973, *Analyse des textes et raisonnement en histoire. Une expérience de recherche historique à partir de l'analyse d'un corpus d'inscriptions funéraires romaines,* in «Archéologie et Calcul», UGE, Collection 10 × 18, Paris.

Borillo, M., 1974, *Techniques de traitement de l'information et procédures formelles en archéologie,* in «Archéologie et Calcul», UGE.

Borillo, M., et Virbel, J., 1974, *Analyse des données textuelles et constructions scientifiques dans les disciplines historiques. A propos des recherches sur le discours illuministe au XVIII^e siècle, de G. Gayot et M. Pecheux,* in «Informatique et Sciences humaines», Paris, à paraître.

Borillo, A.; Borillo, M.; Bourrelly, L.; Chouraqui, E.; Fernández de la Vega, W.; Guenoche, A.; Hesnard, A.; Tognotti, J., et Virbel, J., 1973, *Description des outils (mathématiques, linguistiques et informatiques) impliqués par la construction d'une chaîne automatique intégrée de traitement de l'information textuelle et graphique,* «Information Storage and Retrieval», *9,* 10, 527-560.

Bourrelly, L., et Chouraqui, E., 1973, *Le langage d'interrogation du système SATIN 2 conçu comme un outil de traitement scientifique des données relatives à un corpus quelconque de documents,* Rapport URADCA, CNRS, Marseille.

Chouraqui, E.; Janon, M., et Virbel, J., 1974, *Un système d'exploitation automatique du Corpus des Inscriptions Latines: le SYCIL. Table ronde sur l'application à l'Epigraphie latine des méthodes de l'informatique,* Marseille, CNRS, 812 72, in «Antiquités Africaines», VIII, 1974.

Cros, R. C.; Levy, F., et Gardin, J. C., 1968, *L'automatisation des recherches documentaires. Un modèle général, le SYNTOL,* Gauthier-Villars.

Farinas del Cerro, L.; Fernández de la Vega, W., et Hesnard, A., 1974, *Contribution à l'établissement d'une typologie des amphores dites Dressel 2-4.* Communication au Colloque sur les «Méthodes classiques et les méthodes formelles dans l'étude typologique des amphores», Rome.

Gardin, J. C., 1967, *Methods for the descriptive analysis of archaeological material,* «American Antiquity», *32,* 1, 13-30.

GARDIN, J. C., 1969, *Semantic analysis procedures in the sciences of man,* «Information sur les Sciences Sociales», VII, 1, pp. 17-42.

GARDIN, J. C., 1974, *Les banques de données en archéologie: problèmes méthodologiques, technologiques et institutionnels,* in les «Banques de données en archéologie», CNRS, ed. Paris.

GUENOCHE, A., et TCHERNIA, A., 1974, *Essai de construction d'un modèle descriptif des amphores Dr. 20.* Communication au Colloque sur les «Méthodes classiques et les méthodes formelles dans l'étude typologique des amphores», Rome.

HEMPEL et OPPENHEIM, 1948, *Studies in the logic of explanation,* «Philosophy of Science», XV, Baltimore.

JANON, M., et VIRBEL, J., 1974, *Travaux pour l'exploitation automatique du Corpus des Inscriptions Latines,* in les «Banques de données en archéologie», CNRS, ed. Paris.

VIRBEL, J., 1973, *Methodological aspects of the segmentation and the characterization of textual data in Archaeology,* in the «Explanation of Culture Change: models in prehistory», Duckworth (C. Renfrew, ed.).

Le texte de cette communication est publié dans *Sémiotica* (Mouton, ed.).