AUTOMATISATION DE L'ETUDE DES SITES - UNE METHODE

Louis Bourrelly
Ingénieur au C.N.R.S. (URADCA)

L'introduction, dans l'étude des sites préhistoriques, d'un automate conçu pour mécaniser un certain nombre d'opérations de type calculatoire (au sens large du terme : calculs arithmétiques ou statistiques mais aussi calculs logiques, tris, tracés automatiques de plans, etc.) implique une objectivation aussi complète que possible du raisonnement archéologique dans son ensemble. Ce raisonnement est basé, ici, sur la description des objets de la fouille (outils, ossements, pierres, sédiments...) dans le contexte de leur découverte (coordonnées, stratigraphie, orientation...), il sera donc essentiel de formaliser et d'expliciter clairement le langage de représentation ainsi que les hypothèses qui sont à l'origine de ces descriptions. Il est évident que ces hypothèses, en privilégiant telles ou telles propriétés par rapport à toutes les autres (que décrire pour vérifier ces hypothèses?) vont conditionner fortement le raisonnement lui-même. Mais répondre à cette question n'est pas suffisant, il faudra aussi s'interroger sur la façon dont on va réaliser ces descriptions (comment décrire ?) et seule la cohérence de cette représentation confèrera au raisonnement une nature scientifique. Ce travail de rationalisation, s'il est nécessaire au "bon fonctionnement" de l'automate, n'en fait pas intégralement partie. Les codes descriptifs ainsi créés se situent en amont de la chaîne de traitement, ils sont le domaine réservé du spécialiste qui exprime là ses propres réflexions sur le domaine étudié. Bien qu'il ne soit pas de notre propos de parler ici des codes analytiques, signalons quand même qu'ils sont des ensembles de propriétés (descripteurs, attributs...) précisées par des définitions claires et munis de règles opératoires de description (orientation de l'objet, segmentation, seuils de différentiation...). La description d'un "objet" donné peut se faire de multiples façons et plusieurs codes peuvent (et doivent ?) être imaginés, chacun d'eux fournissant ses résultats (l'absence de résultat dans l'optique envisagée pouvant être considérée comme un résultat intéressant si la démarche utilisée est suffisamment formalisée). La visée de ces codes n'est donc pas une représentation unique et définitive du "continuum descriptif" que constituent les objets à décrire, mais plutôt la représentation, au travers d'un langage cohérent, de certaines hypothèses de base. Toute démarche ultérieure (rai-

sonnement et interprétation) ne pourra être validée qu'à l'intérieur de ces règles de représentation. Inversement il faut souligner que tout résultat obtenu à l'issue d'un "calcul" n'est pas forcément pertinent, parce que formalisé, par rapport au domaine d'application. Là encore l'intervention critique du spécialiste sera nécessaire et c'est lui et lui seul qui pourra juger de l'édéquation des conclusions obtenues.

Nous venons ainsi d'examiner le contexte dans lequel se situait l'automate (Fig. 1) et de montrer que cette "boite noire" n'était en fait, dans la chaîne du raisonnement qu'un outil, sans doute commode et générateur de savoir, mais dénué d'exprit d'initiative et dont l'efficacité ne peut qu'être liée à celui ou à ceux qui le manipulent.

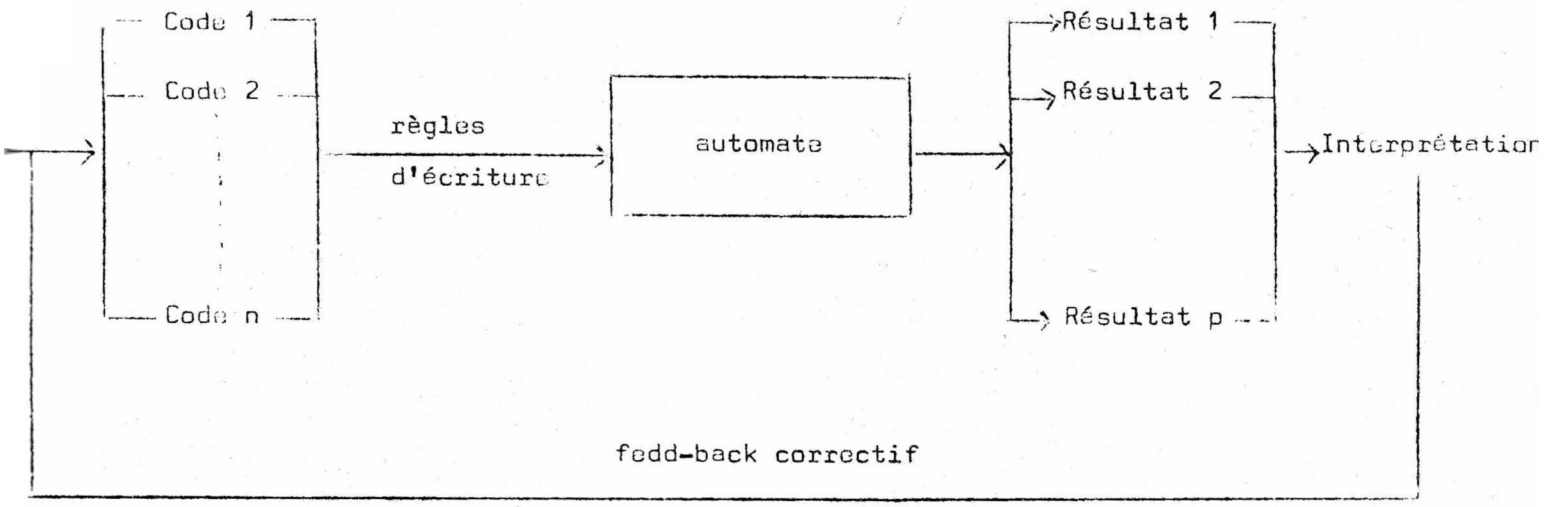
I - Description de l'automate (Fig. 2)

L'automate est donc un outil destiné à traiter d'importantes quantités d'objets (ou plus exactement les informations retenues pour leur description) au moyen d'un calculateur digital. Il se présente sous la forme d'un programme, c'est à dire d'une suite d'instructions données à la machine pour exécuter et résoudre les demandes des utilisateurs. Pour une approche très macroscopique nous pouvons distinguer, dans ce programme, trois parties essentielles : un langage de représentation des données, un lexique (ou thesaurus) et un langage d'interrogation.

I - 1. Le langage de représentation des données

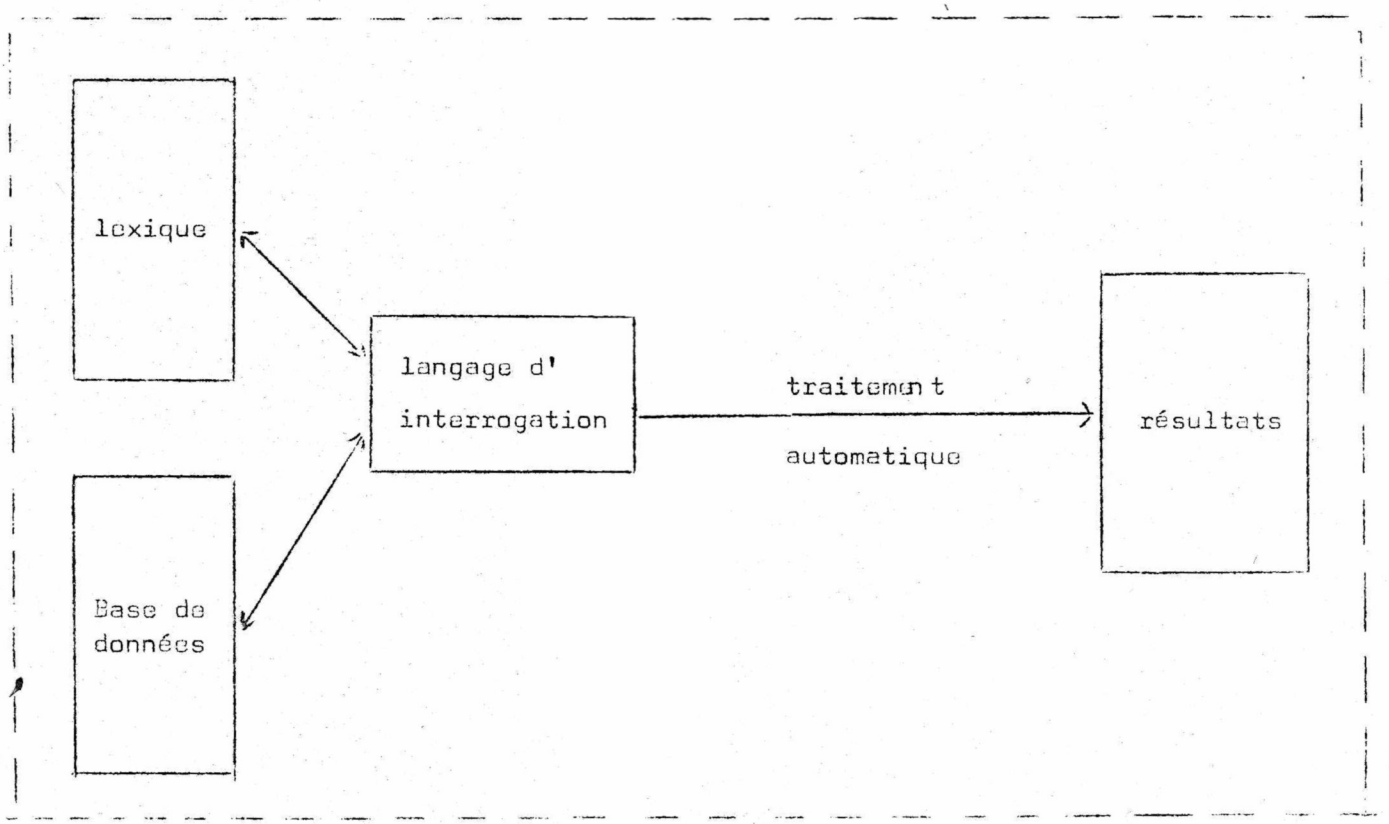
C'est un langage formel qui constitue une structure d'accueil autorisant la représentation des codes analytiques. Il est suffisamment général pour ne pas être remis en cause à chacune des modifications apportée aux descriptions des documents, ni à chaque test d'une hypothèse différente, ni même à chaque champ d'application. Cette généralité trouve sa justification dans des raisons économiques (l'investissement financier et intellectuel étant le plus souvent assez considérable) mais aussi dans des raisons d'ordre méthodologique (puissance du langage pour autoriser au maximum les finesses d'une analyse). Enfin, en respectant au mieux les contraintes imposées par les calculateurs digitaux (représentation interne de l'information, organisation des fichiers pour minimiser les temps d'acces,...) la structure du langage doit assurer une complète mobilité de la base de données : rajout ou retrait d'informations (apport de nouveaux objets, intégration de nouvelles propriétés...).

Ce langage est articulé sur un découpage en catégories d'informations qui sont autant d'aspects différents de l'objet retenus pour l'analyse (coordonnées, stratigraphie, technique de taille, retouche, dimensions,..). Chacune de ces catégories introduit un nombre quelconque de propriétés qui caractérisent l'objet étudié (longueur, épaisseur, abscisse, retouche directe, troncature,...). Ces descripteurs eux-même peuvent être mis en relation (opposition, adjacence, localisation,...) ou précisés (tendances, poli, couleur,...). Tous ces divers éléments de l'analyse sont représentés



- Fig. 1 -

Chaîne de traitement des informations



- Fig. 2 -

L'automate

suivant les règles strictes de la syntaxe du langage qui assure la cohérence du contenu informationnel par rapport aux algorithmes de traitement.

Il va de soi qu'un tel langage va réagir sur la structure de l'analyse (c'est à dire les codes analytiques qu'il va représenter) et imposer ses solutions formelles aux problèmes posés par les descriptions. Le détail des solutions retenues ne peut, en effet, qu'être différent suivant que le code est utilisé au moyen d'un calculateur, d'un appareil de tri à aiguilles, d'un ensemble à sélection visuelle ou bien encore par simple lecture sur un support imprimé.

I - 2. Le lexique

C'est d'abord l'ensemble des termes descriptifs utilisés pour la représentation des objets. Cette exhaustivité a pour but l'homogénéité des descriptions dans le temps (d'une série d'analyses à une autre) aussi bien que dans l'espace (travail simultané de plusieurs personnes). Cet ensemble est généralement structuré de la même façon que le découpage analytique et reflète donc les mêmes catégories d'informations.

Pour certaines applications, quelques unes de ces catégories introduisent des caractéristiques qui présentent entre elles des relations statiques ou "a priori" (relations paradigmatiques) qu'il peut être commode d'interpréter et qu'il est donc nécessaire de porter à la connaissance de l'automate. Ces relations sont le plus souvent de nature hiérarchique (inclusion, appartenance, filiation, etc...), le lexique va prendre alors une forme arborescente et l'automate interprétera cette structure lors de son fonctionnement. Pour illustrer ce type de relations citons la classification animale (mammifères - carnivores - canidés...), la structure du corps humain (bras - avant bras - coude - poignet...), ou la structure administrative d'un pays (région - département - commune). Ce sont ces informations qui permettront à l'automate de savoir, par exemple, lors d'une interrogation sur les os de carnivores que ceux appartenant à des chiens sont pertinents.

Signalons enfin parmi les autres fonctions du lexique la réduction des synonymies, la vérification de la cohérence du vocabulaire (interdiction des doubles définitions de descripteurs) et la codification automatique des termes issus du langage naturel en vue de leur représentation en machine.

I - 3. Le langage d'interrogation

Maintenant que l'automate est muni, d'une part d'une base de données constituées des analyses d'objets et d'autre part d'un lexique, il faut pouvoir l'utiliser pour isoler à l'intérieur de cette masse d'informations des sous ensembles d'objets répondant à certaines combinaisons de caractèristiques, c'est le but du langage d'interrogation. Cette "recherche rétrospective d'informations" (information retrieval en américain) ne pourra porter que sur le contenu des analyses mais la puissance et la rapidité de

l'automate autorisent une combinatoire dont la richesse ne pourrait pas être obtenue par des seuls tris manuels.

La combinaison des termes descriptifs s'exprime grâce à des opérateurs logiques (booléens) dont les plus courants sont le ET (présence simultanée de deux termes), le OU (présence de l'un ou de l'autre ou des deux à la fois) et le NON (absence obligatoire du terme cité). Les expressions logiques ainsi écrites peuvent avoir plusieurs niveaux de parenthèses (cf. les expressions algébriques) exprimant ainsi les moindres détails d'une question élaborée. Enfin, il est possible d'utiliser les opérateurs arithmétiques (= , < , $\leqslant$, > , $\geqslant$) pour manipuler, dans ces expressions, les variables numériques telles que "longueur", "épaisseur", ou "coordonnées".
Exemple fictif d'écriture d'une telle expression :
"outil à retouche marginale (M) directe (D) ou inverse (I) sur le bord latéral droit (LD) dont la longueur (L) est supérieure à 40 mm".
En supposant que les termes soulignés constituent des entité au niveau de la description, on obtiendrait la représentation suivante :

$$(((M.ET.(D.OU.I)).ET.LD).ET.L>40)$$

ce qui est une expression très simple à un seul niveau de parenthèses (pour l'opérateur OU, les autres ne servent ici qu'à bien marquer les différentes phases séquentielles des opérations). Il est certain que cette technique de représentation autorise une richesse telle que ces expressions sont le plus souvent difficilement formulables dans le cadre d'une langue parlée (au moins de façon simple et directe).

Ce langage d'interrogation permet en outre de définir les sorties du système : type de produit attendu en réponse support de sortie (imprimante, disque, bande...), format d'impression ou d'édition des références. Sa forme et la simplicité de son utilisation doivent permettre l'établissement d'un véritable dialogue entre l'utilisateur et l'automate.

II - Les produits de sortie (quelques résultats)

L'automate qui vient d'être décrit ici est un système de recherche documentaire dont la fonction directe est de trier au sens où nous venons de le définir, ses produits directs seront donc le reflet de ces tris :

- comptage des objets répondant à une question donnée
- extraction et édition des références de ces objets
- fabrication d'index
- extraction de valeurs numériques (coordonnées, dimensions) ou descriptives (propriété)
- fabrication de grilles à usage de calculs (classification)

Sur ces produits immédiats il sera possible d'articuler une bibliothèque d'autres programmes, lesquels auront pour fonction de fournir, à la demande d'autres formes de résultats plus élaborés dans leur forme ou ayant fait l'objet de divers calculs :

- tracés de plans montrant la répartition des objets sur un sol d'habitat (à partir de l'extraction des coordonnées)
- tracés de graphiques cumulatifs
- impressions de tables de pourcentages
- calculs d'indices
- classification automatique
- étude d'une répartition statistique

Cette liste d'exemples est donnée à titre indicatif et ne prétend pas épuiser le sujet, il sera toujours possible d'imaginer d'autres sorties en fonction des besoins de telle ou telle étude (voir la liste des références illustrant quelques uns de ces types de travaux).

III - Conclusion

L'utilisation d'un tel automate conduit à une redéfinition claire des objectifs visés et des moyens à mettre en oeuvre pour mener à bien l'étude envisagée, en ce sens nous pensons que son influence sur les méthodes même de la recherche n'est pas négligeable. Le développement récent des banques de données en archéologie montre, si besoin était, l'intérêt que suscite désormais ce type d'approche dont l'ambition première est de libérer le spécialiste des travaux longs et fastidieux que sont les compilations sans cesse répétées. Enfin la masse croissante des informations disponibles et l'impossibilité devant laquelle on se trouve de les exploiter correctement sont autant de facteurs qui conduisent à des solutions du type de celle qui vient d'être décrite ici.

Bibliographie sommaire

BINFORD Sally R. et BINFORD Lewis R., 1966 - A Preliminary Analysis of functional Variability in the Mousterian of Levallois Facies. In "Recent Studies in Paleoanthropology; eds J.D.CLARK et F.C.HOWELL, "American Anthropologist" 68, 2, Part 2, 238-295.

BORILLO M., 1971 - Formal procedures and the use of computers in Archaeology. "Norwegian Archaeological Review", 4, 1, 2-27.

BOURRELLY L., 1973 - The automatic processing of the results of an excavation at a prehistoric site - examination areas. In "The explanation of culture change - Models in Prehistory". Ed. COLIN R. - 109, 114. Editions Gerald Duckworth and Co - Gloucester.

COWGILL G., 1968 - Computer Analysis of archaeological data from Teotihuacan, Mexico. In "New perspectives in archaeology" eds S.R. et L. R. BINFORD (q.v.), 143-150.

FERNANDEZ DE LA VEGA W., 1968 - Analyse quantitative du mobilier funéraire de la fouille de Sala Consilina. In "Calcul et formalisation dans les sciences de l'homme". Ed. B.JAULIN, 121-129. Editions du C.N.R.S., Paris.

GARDIN J.C. et BORILLO M., eds., 1970 - Archéologie et calculateurs - Problèmes sémiologiques et mathématiques. Editions du CNRS, Paris.

HODSON F.R., 1970 - Cluster Analysis and Archaeology : Some new developments and applications. World Archaeology, 1, 3, 299-320. Londres.

LAGRANGE M.S., 1973 - Analyse sémiologique et Histoire de l'art. Edit. KLINCKSIECK, Paris.

LANDAU J., 1971 - Les représentations anthropomorphes mégalithiques de la région méditerranéenne (IV° au Ier millénaire). Paris.

WHALLON Jr R., 1973 - Spatial analysis of palaeolothic occupation areas. The present problem and the "functional argument". In "The explanation of culture change - Models in Prehistory", ed. COLIN R., 115-130. Editions Gerald Duckworth and Co. Gloucester.