OUVRIR A UN LARGE PUBLIC L'ACCES A UNE INFORMATION SPECIALISEE

Depuis de nombreuses années déjà les archéologues construisent des banques de données dont les plus opérationnelles sont les bases à visée documentaire. Après la période d'expérimentation, il en existe maintenant d'utiles et d'utilisées, qu'on met régulièrement à jour et on continue aussi à en entreprendre de nouvelles, tirant le meilleur parti des progrès spectaculaires de l'informatique, tant du point de vue du matériel que du logiciel. Ces banques ont en commun d'être produites par des chercheurs et destinées à des chercheurs. Si elles ont l'avantage de répondre à leurs besoins, elles ont l'inconvénient d'être lourdes à constituer : il faut, en effet, extraire toutes les informations des documents archéologiques, et les analyser selon un système descriptif pré-établi, le langage documentaire. Par ailleurs ces banques ne sont pas facilement accessibles au public puisque, pour les interroger, il faut utiliser — et donc connaître — leur langage documentaire.

Nos banques demandent des investissements de plus en plus lourds et — pour des raisons morales autant qu'économiques — il serait bon de pouvoir en faire profiter un public plus vaste, car elles pourraient servir ainsi à la *formation* des étudiants mais aussi à l'*information* d'un public de plus en plus intéressé, on le sait, par l'Antiquité.

On voit d'ailleurs apparaître des réalisations multimédia destinées à un public de non-spécialistes comme, par exemple, le vidéodisque *Parthénon* du Musée du Louvre ou l'encyclopédie *Perseus* des Yale University Press. Tout à fait innovants dans leurs objectifs comme dans les technologies d'information qu'ils mettent en oeuvre, ces deux produits aussi ont été lourds à construire, parce que les données qu'ils contiennent n'ont été rassemblées que pour ces seules réalisations; et ce travail est l'oeuvre de chercheurs, ce qui en garantit la qualité scientifique.

A partir d'un même ensemble de données d'origines diverses et, en partie au moins, déjà constituées pour d'autres objectifs, est-il possible de construire un système d'information capable de satisfaire divers types de public ? Peut-on leur permettre l'accès de nos données de chercheurs à travers d'autres types de consultation que ceux pour lesquels elles ont été initialement constituées ? C'est la recherche que nous menons actuellement dans le Centre fondé par le professeur R. Ginouvès, et dont je suis responsable.

Nous avons déjà pu montrer, il y a quelques années, que le simple fait d'ajouter des images à une base documentaire — en l'occurrence des images stockées

sur un vidéodisque pour illustrer trois bases dont celle sur la *Mosaïque dans le monde grec* — permet d'en ouvrir la consultation à deux types de publics nouveaux: — les spécialistes d'autres domaines qui choisissent dans la base les données — descriptions et images — pertinentes pour leur recherche et les complètent avec leurs propres informations,
- un public non formé à l'interrogation de la base, et même, parfois aussi, peu connaisseur de l'Antiquité, qui va pouvoir « feuilleter » les images du disque et lancer une interrogation automatique à partir d'une image qu'il aura choisie; ainsi la base de données lui fournit à la fois les autres images relatives au même document, les informations correspondantes et la forme de leur analyse, ce qui peut permettre alors de lancer une recherche à partir de ces données, ou de continuer la recherche par feuilletage et interrogation automatique.

Nous étendons maintenant notre recherche au traitement documentaire du langage naturel. A titre expérimental, nous constituons un système d'information sur le site de Delphes, en rassemblant toute une série de textes concernant l'histoire de Delphes, la topographie du site, l'architecture et le décor des bâtiments, les objets conservés au musée, aussi bien que les cultes dans le sanctuaire, les concours en l'honneur d'Apollon, ou même la Grande Fouille dont on fête cette année le centenaire. Ces textes en français proviennent de publications imprimées aussi diverses que des guides touristiques (le *Guide Bleu*), les guides archéologiques (les deux tomes publiés en 1991 par l'Ecole française d'Athènes), des volumes de la publication des *Fouilles de Delphes*, des monographies, des articles...

Sous une forme numérisée, les textes sélectionnés sont introduits dans le logiciel SPIRIT (société SYSTEX), logiciel qui en permet la consultation en langage naturel. Pour ce faire, il met en correspondance le texte de la question posée et ceux des textes de la base, en partant du principe qu'un texte de la base, ou une partie de texte, a d'autant plus de chances d'être pertinent qu'il contient les mêmes concepts que ceux de la question (puisqu'il traite alors précisément du sujet intéressant l'utilisateur). Pour établir cette correspondance, le traitement consiste — et pour les textes de la base et pour celui de chaque question — en une indexation automatique très riche, qui est rendue possible par des analyses linguistiques (analyse morphologique, analyse syntaxique, analyse des locutions, normalisation) puis des traitements statistiques destinés à calculer le "poids informationnel", par rapport à l'ensemble des textes de la base, de chacun des concepts ainsi repérés et indexés. L'utilisateur peut aussi demander au système d'élargir sa question en la reformulant automatiquement avec des mots reliés sémantiquement (synonymes...) à ceux de la question initiale. De plus, des informations factuelles peuvent être ajoutées à chaque document, elles sont interrogeables de façon classique en utilisant des descripteurs; si c'est nécessaire,

on peut donc combiner, dans une même question, les critères exprimés dans un langage documentaire et la recherche en langage naturel, qui reste, bien évidemment, l'atout majeur du logiciel.

A ces textes, nous avons ajouté des documents issus de fichiers de chercheurs ou d'autres banques de données exploités avec d'autres logiciels; sous SPIRIT, ces documents, formés uniquement de descripteurs, deviennent interrogeables en langage naturel.

De plus, le système d'information donne accès aux 900 photographies de Delphes enregistrées dans notre vidéodisque *Images de l'archéologie* (elles appartiennent au Centre de Documentation Photographique et Photogrammétrique, CDPP, CNRS-Université de Paris I); à ces images analogiques, il est prévu d'ajouter des images numériques: photographies complémentaires et, surtout, des plans et dessins des édifices du site.

L'application de SPIRIT à des textes concernant l'Antiquité nécessite, pour profiter pleinement des fonctionnalités du logiciel, qu'on enrichisse le dictionnaire linguistique de tous les mots propres à notre domaine (comme, par exemple, "cnémides", "opus tessellatum", etc.), en indiquant leurs caractéristiques grammaticales afin que la "normalisation" de leurs différentes formes puisse être effectuée correctement avant l'indexation automatique; et qu'on enrichisse le thésaurus en précisant l'environnement sémantique de chacun de ces termes, de manière à permettre la reformulation automatique de la question à partir des synonymes ou autres mots reliés sémantiquement à ceux de la question. Cet enrichissement reste ensuite valable pour tous les nouveaux documents de la banque.

Ainsi, l'utilisateur pose une question du type « Comment se déroulaient les concours musicaux et sportifs en l'honneur d'Apollon ? », ou « La dispute d'Apollon et Héraklès pour la possession du trépied », ou bien « La sibylle et les rites oraculaires », ou encore « Quelles offrandes de bronze a-t-on retrouvé dans le sanctuaire et quelles étaient leurs techniques de fabrication à l'époque classique ? ». Le système propose en réponse les documents (textes, notices issues d'autres bases, images associées) qui lui semblent correspondre à chacune de ces questions, et il les classe en fonction du degré de pertinence qu'il a calculé. L'utilisateur peut ainsi consulter les documents en sachant quelles combinaisons de critères ont conduit à leur sélection. Les images elles-mêmes doivent pouvoir bientôt constituer des points d'entrée et permettre la consultation de textes et notices à partir d'un feuilletage. Enfin, l'utilisateur peut choisir, dans les documents qui lui sont proposés, une phrase, un paragraphe ou même un texte entier qui correspond particulièrement à ce qu'il cherche, et s'en servir pour lancer automatiquement une nouvelle question qui lui donnera accès à de nouveaux documents. Il peut ainsi explorer la base en affinant progressivement sa question, ou passer d'un concept à un autre — qu'il soit porté par une image ou par un

texte — suivant son propre cheminement. Si les liens entre documents et images sont évidemment fournis au système à chaque introduction de documents dans la base, les liens entre documents n'ont pas à être prédéclarés puisque c'est le système qui les calcule, par indexation automatique, au fur et à mesure des demandes de l'utilisateur et en ne prenant en compte que son propre point de vue: on atteint ici des fonctionnalités d'hypertexte dynamique particulièrement intéressantes.

On comprend le double intérêt d'un système exploitant directement le langage naturel : d'une part, il permet de donner accès à des informations portées par des textes rédigés dans une perspective traditionnelle sans qu'il y ait, pour construire la banque, à en extraire une sélection d'informations et à les analyser selon un système descriptif lourd et restrictif; d'autre part, le fait que l'interrogation soit elle-même en langage naturel permet d'ouvrir la consultation à tout type de public puisque l'utilisateur n'a plus besoin d'apprendre le langage documentaire. Et les fonctionnalités du logiciel, permettant la gestion des images et des modes de navigation à la fois puissants et souples, enrichissent encore le caractère interactif de cette consultation multimédia.

A travers ces recherches nous ne visons pas du tout à remplacer un type de base de données par un autre : la méthode traditionnelle d'analyse d'informations structurées par le biais d'un langage documentaire et leur interrogation par critères combinés à l'aide des opérateurs booléens demeure, malgré ses inconvénients, la plus efficace et elle seule permet une exploration systématique du matériel. Il s'agit seulement d'expérimenter un mode de consultation plus proche de la pratique traditionnelle avec ses avantages — mais aussi quelques uns de ses inconvénients bien connus. Il s'agit aussi de trouver un moyen d'accéder à des données très hétérogènes à travers un même outil, suffisamment puissant et convivial pour qu'il satisfasse des publics eux-mêmes très hétérogènes, interrogeant le système dans des perspectives très différentes.

Plus généralement, quelles qualités est-on en droit d'attendre d'un système d'information de ce type pour qu'il remplisse complètement ses objectifs ?

- La première est certainement *la richesse des informations* contenues, d'un point de vue tant qualitatif que quantitatif. Cette exigence suppose d'abord une sélection judicieuse des documents, et, lorsqu'il s'agit de textes longs, leur "découpage" en unités appropriées à ce type de consultation; elle suppose aussi la pertinence des liens établis entre textes ou notices et images. D'un point de vue technique, qualité et quantité des informations à prendre en compte supposent presque nécessairement, nous l'avons vu, de pouvoir gérer des données de type hétérogène, provenant d'environnements différents. Elle suppose enfin, surtout dans un domaine comme le nôtre, la gestion de données multimédia: les images sont indispensables, photographies et documents graphiques, auxquelles il pour-

rait être intéressant d'ajouter des images animées et même du son, que nous n'incluons pas aujourd'hui.

- La seconde qualité, complémentaire de la première, est la *pertinence des informations* fournies en réponse : il faut non seulement que les documents obtenus correspondent à la question posée, et qu'ils contiennent les informations recherchées, mais il faut aussi qu'ils correspondent au "niveau" de l'utilisateur: le *Guide bleu* intéressera aussi peu le spécialiste que certaines de nos publications de fouille intéresseront le "grand public".

- En troisième lieu, le système doit être *ergonomique* sous peine de perdre la majeure partie de son intérêt, auprès des divers publics.

- Quatrième qualité du système, une *réelle interactivité* et, pour cela, il doit offrir le choix entre *divers types d'accès*: à partir des textes seuls, ou à partir des critères factuels, ou encore à partir d'une combinaison de ces deux modes; à partir des images, à partir des documents précédemment sélectionnés...

- En cinquième lieu, le système doit être *ouvert* et permettre aussi bien la mise à jour de la base, et son enrichissement par l'*importation* de nouveaux documents, que l'*exportation* de certains documents vers d'autres environnements (traitement de textes, fichiers d'images...). Cette possibilité technique étant offerte, il restera évidemment aux responsables du système d'information à déterminer quelles opérations sont autorisées pour quels utilisateurs et dans quel cadre.

- En dernier lieu, le système d'information ainsi réalisé sera d'autant plus facilement *diffusable* qu'il sera implanté sur des configurations matérielles et logicielles répandues, et/ou portable sur plusieurs types de configurations.

On comprend que toutes ces exigences ne sont pas faciles à satisfaire. La dernière notamment pose de réels problèmes techniques: jusqu'à quel point, en effet, un tel système peut-il être indépendant de l'environnement informatique sur lequel il est exploité ? Pour notre part, nous avons utilisé un logiciel tout à fait spécifique, même s'il tourne sous plusieurs systèmes d'exploitation (actuellement Unix et OS/2). Et, pour les données elles-mêmes, on se heurte notamment aux problèmes de format d'enregistrement, qu'il s'agisse du texte, de l'image — fixe ou, plus encore, animée — ou du son, de leur standard de compression, etc.

Si ces problèmes techniques ne sont pas simples, ils sont, à plus ou moins long terme, en voie d'être résolus. Toutefois, il reste des problèmes plus fondamentaux comme « quelles informations pour quel public ? ». Si, on le sait, constituer les données destinées à des spécialistes bien identifiés n'est déjà pas chose simple, que dire lorsqu'on travaille pour des publics hétérogènes, dont les besoins sont difficilement définissables avec précision ? La recherche est loin d'être terminée, d'autant qu'après une phase de conception et de mise au point des outils elle doit ensuite s'appuyer sur une longue période d'expérimentation auprès de publics divers qui, seuls, pourront apporter une nécessaire validation à

ces travaux. Cette recherche a commencé dans notre Centre en 1991, elle est donc encore loin de son terme mais ses premiers résultats sont prometteurs.

ANNE-MARIE GUIMIER-SORBETS
Centre de Recherche
"Archéologie et Systèmes d'information"
Université de Paris X - CNRS

BIBLIOGRAPHIE

Pour des informations complémentaires, on pourra consulter:

— *sur les travaux du Centre de Recherche "Archéologie et Systèmes d'information":*

GUIMIER-SORBETS A.M. 1990, *Les Bases de données en Archéologie. Conception et mise en oeuvre.* Paris, CNRS.

GINOUVÈS R., GUIMIER-SORBETS A.M. 1991, *Un Centre de Recherche sur les systèmes d'information en Archéologie*, « Archeologia e Calcolatori », 2, 7-12.

— *sur le vidéodisque « Images de l'Archéologie » que nous avons réalisé en collaboration avec le Centre de Documentation photographique et photogrammétrique (CDPP, CNRS-Université de Paris I), et grâce à un financement du Ministère de l'Education nationale:*

Images de l'Archéologie. Vidéodisque, Paris 1986.

— *sur l'apport des images aux banques de données et aux systèmes d'information:*

GUIMIER-SORBETS A.M. 1988, *Apports du vidéodisque à la recherche scientifique*, dans S. CACALY (ed.), *Image et vidéodisque*, Paris, 121-133.

GUIMIER-SORBETS A.M. 1990, *Nouveaux axes de recherche dans la constitution de systèmes documentaires intégrant analyses et images*, dans G. LOSFELD (ed.), *Colloque Sciences historiques, sciences du passé et nouvelles technologies d'information*, Lille, 329-335.

GUIMIER-SORBETS A.M. 1991, *Recherches sur l'apport des images aux banques de données et systèmes d'information archéologique*, dans *Informatique et Statistique dans les Sciences humaines*, 27, Liège, 129-135.

— *sur le système d'information sur la sculpture hellénistique de Délos:*

GUIMIER-SORBETS A.M., JOCKEY Ph., *Système d'information sur les sculptures de Délos*, communication au *Colloque européen Archéologie et Informatique* (Saint-Germain-en-Laye, Musée des Antiquités nationales, 21-25 Novembre 1991). Texte à paraître dans les *Actes* en 1994.