8 FROM TEXT TO IMAGE

An Experimental Multimedia Information System based on Querying Texts in Natural Language

Anne-Marie Guimier-Sorbets
Professeur de Sciences de l'Information
Université de Paris X - Centre national de la Recherche scientifique (Paris)

In the field of Human Sciences researchers have for a long time been creating databases, the most operational of which are those of a documentary nature. These bases have in common the fact that they are produced by researchers and destined for researchers. If they have the advantage of meeting the researchers needs, they have the disadvantage of being long to build up : as we know, all the information must be extracted from the documents - at least that which we have decided to take into account, and analysed according to a pre-established descriptive system. Furthermore these data bases are not easily accessible for the public because to query them one must use, and therefore know, their formal query language. A way to make profitable the necessary investment of their building would be to let a wider public benefit from them, they could then serve not only to train students but also to inform a public who is, we know, more and more interested in the whole cultural scene.

We see appearing, moreover, multimedia products of a cultural nature which are destined for the public at large; in museums and more generally in the area of Fine Arts, videodisc, CD-ROM and CD-I collections are being built up, either as edited products to be commercialised or as works designed to be used in museums. Innovative in the information technology they use, and sometimes also in their aims, these products are time-consuming to build up, because the data that they contain is only brought together for this purpose and, if we want to guarantee the scientific quality of the work, it has to be done by specialists.

Would it be possible to build an information system capable of satisfying different types of public, from the same data, coming from different origins, and already built up for another objective? Can data collected by researchers be opened up and accessed by other methods of enquiry than that for which it was initially put together?

We saw a few years ago that the simple fact of adding pictures, in this instance images stored on a videodisc, opens up a documentary archaeological database for use by a new public. But is it possible to go further and bring together documents as diverse as those which are used traditionally texts in natural language coming from diverse sources and written to be looked up in a traditional manner, graphic and photographic images, description sheets from databases etc. and give access to a wide public by means of

electronic lookup? If the first part of the problem is essentially technical, the second is of a different nature: extending public use will necessarily entail different usages because, from the same specialised information, the system must not only meet the needs of the researchers and help educate students and pupils but must also meet the expectations of a wider public; now, if the researchers needs are for the most part identified, those of education are not; as for the expectations of the public "at large" in quest of cultural information, they are difficult to define since these products should, in reality, stimulate and even, to a great extent, arouse the need, or rather the desire, for information. We should then step from the world, a bit barren but relatively well-signposted, of scientific and technical information products, to the world now appearing and therefore more hazy, of products of a culture nature, what the Americans have already named *infotainment*.[1]

It was with these two directives in mind that we undertook our research at the *Centre de recherche Archéologie et Systèmes d'Information* (Université de Paris X- CNRS) in collaboration with the *Centre de Recherche sur l'Information specialisée* at the same university, and we built an experimental information system which would let us identify the different types of problems raised and try different types of solution. If this experimental base is specific to Antiquity, a scientific dicipline in which the recourse to texts, descriptive data and images is as necessary for research as for the diffusion of knowledge. The problems raised are not particular to this field of application, as we will try to demonstrate here. Indeed, although optical disks provide great possibilities for storing texts and high quality images, in lesser or greater number according to how they are recorded, and while softwares offer various possibilities for accessing the base by query and/or by navigation, the content of such systems remains to be defined in relation to the expectations of different types of public and the different ways of accessing different types of information, as well as the enquiry methods remain to be studied.

An experimental information system on the site of Delphi (Greece)

For this experiment, we chose the site of Delphi, liable, because of its diverse historical, archaeological, religious, philosophical, tourist and even athletic aspects, to interest archaeologists and experts in Antiquity, as well as students, schoolchildren and an even wider public.

The information system that we built therefore brings together a whole series of texts relating to the history and the topography of the site, to the architecture and decoration of the buildings, to the objects in the museum as well as to the sanctuary cults, the competitions in honour of Apollo, and even to the *Grande Fouille*, whose centenary was celebrated this year. These texts, in French, come from printed publications as diverse as tourist guides (the Guide Bleu), archaeological guides (the two volumes published in 1991 by the *Ecole Française* in Athens), volumes from the publication *Fouilles de Delphes*, monographs, articles, etc.

The chosen texts were entered, in digitised form, into a software which let them be queried in natural language, and which in order to do this relates the text of the question asked to the texts in the database, working from the principle that a text, or a part of a text has more chance of being relevant if it contains the same concepts as the question (since it deals with the subject which interests the user). To create this correspondence, we used

1 This word, made up from "information" and "entertainment", has the advantage of stressing the double nature of information as well as the idea of leisure activity and even of entertainment which must exist in this type of product.

the SPIRIT software (from SYSTEX) which, for each text in the base and for the text of each query, carried out a detailed automatic index (picking out the keywords). This was made possible by linguistic analysis (morphological, syntactical and locution analysis, and standardisation) followed by a statistical processing designed to calculate the "information weight" in relation to all the texts in the base, and to each concept identified and indexed. The user can also ask the system to widen the scope of his question by reformulating it automatically with words related semantically (e.g. synonyms). Moreover, factual information could be added to each document, they could be queried in the traditional way; if necessary, the same inquiry could combine criteria expressed in formal query language and the information retrieval in natural language, which is, obviously, the major advantage of the software.

Therefore the user could have a query such as "the sybil and the Delphic oracle", or "the sculpted decoration of the Treasure of Siphnos", or "sport competitions", or "chariot races in honour of Apollo", or even "What offerings in bronze were found in the sanctuary and what techniques were used to make them in ancient times?". In reply the system offers documents (texts, notices from other bases, related images) which seem to it to correspond to each of these queries, and categorises them in relation to the degree of relevance it has calculated. The user can then query the documents knowing what combination of criteria has led to their being selected.

This experiment allows us to raise questions related to querying in natural language, to the preparation of texts before they are put into the base, to the respective and complementary contributions of automatic indexing and manual indexing, to the accessing of images and documents coming from other databases, and to different ways of interactive querying. Finally, we will try to define the qualities necessary for a system of this type, and evaluate the limitations of this experience and the stakes of such research.

Querying in natural language

One of the primary concerns of our research was to access data in natural language by a query interface itself in natural language: this interface offers a flexibility and simplicity well-suited to our objective. Opening up to a wide public necessitates, first of all, an ergonomic system - there can be no question of the user having a long training period in order to query the system, and secondly a system flexible enough to be used without problem by an heterogeneous public.

Certainly, other types of solution have been set up to solve these user-interface problems: on one hand systems which work by tree structure menus and criteria lists, on the other hand "classic" information retrieval systems in which the user enters his query directly. The first have the advantage of simplicity since the user only has to indicate, at each step, the option or criteria chosen; these systems, efficient when they are well adapted to the public for whom they are prepared, have the disadvantage of being very rigid: first of all, because a too great choice of menus can not be given, they can only be designed for a relatively well defined user; also, the progression of these menus quickly becomes unbearable for the experienced user (even a person who starts out as a beginner makes progress quickly if he asks series of questions); finally, even if the retrieval can be done according to several criteria, it is impossible to combine them freely. For these reasons, such systems do not suit heterogeneous users, whether or not they are specialists in the field concerned, or trained in query techniques they are liable to ask a wide range of unforeseeable questions, it is therefore impossible to prepare criteria grids with enough precision to be relevant in all cases.

The "classic" information retrieval systems are, on the contrary, more flexible because the user formulates his request by choosing and combining his criteria as he wishes. However, in order to use it in his query, he must know the formal query language which was used to analyse the documents; and the correct use of Boolean operators requires a training period: if this type of interface remains the most efficient for specialists both in the chosen field and in document querying, it is not, as we know, designed for the public at large. So, neither of these two types of interface - each as efficient as the other for a defined public - is suitable for a heterogeneous public. That is why in our experiment we favoured query in natural language, even though studies done by other teams have shown its limitations[2].

In our experiment the system could be queried in natural language, that is to say, with the subject's own vocabulary and a freely chosen syntax, as the examples show. The developed linguistic processing compensates for some of the inconveniences well known in using natural language for document retrieval; and if the quality of the replies obtained by this method do not match those obtained when using a regular and completely controlled formal query language, its easy use and flexibility make it a worthwhile query tool - it is up to us to improve its performance in developing the tools put into place by linguistic processing - even offering to those who wish it the possibility of combining the advantages of automatic and manual indexing, as we will see.

Manual indexing - Automatic indexing

As with all document systems, only the indexable information is taken into account in the querying, and it is therefore necessary to make explicit in the documents in the base all sorts of information implicitly contained in the text. Two systems, which can be combined, are available to do this: either the information is added to the text of the document - and will then be able to be located by the automatic indexing of the text in natural language - or the information is indexed manually, i.e. in the traditional way, with keywords in a actual field.

The possibility of adding these actual fields - flexible since they are optional for each document - allows us to complete the information carried by the text itself: in the experiment at Delphi, we chose to add signpost information such as a bibliographical reference for the text, its author, the date of publication, and analytical information such as which period the text was concerned with (ARCHAIC, CLASSICAL, MODERN...) and some keywords indicating the main subject of the text. Manual indexing, by adding keywords, is done traditionally: we had decided on these keywords beforehand - or at least decided on the nature of the information to index and the level of analysis. Manual indexing must remain light since it is only a complement to the automatic indexing of the text itself: we therefore limited it strictly to the main subject and a few basic ideas for querying which, perhaps, would not have been easily located by automatic indexing, as we have seen.

This double indexing system, which is particularly efficient, also allows us to evaluate the respective advantages and disadvantages of each of these two methods. Those of manual indexing are well known: a major advantage is its great reliability, the use of a formal access language assuring a strict regularity in the content description for a document; its inconvenience is the weight of its analysis processes. Automatic indexing, on condition

2 Specially for OPAC consultation, although here it concerns a different type of consultation. In any case, the first experiments led us to make the same conclusions on user text quality as Y.Polity and J.M. Francony at the RIAO 1991 Colloque (*Intelligent text and Image Handling*, RIAO 1991, *Conference Proceedings*, vol. 1, p.357-372).

that it is complete enough, has different advantages, which are appreciable in themselves: first of all the ease of the processes, since it is no longer necessary to analyse the contents of a document; also the indexing is not done a priori, but for each question, i.e. from the point of view of the user and not that of the archivist; a secondary idea in the text, which because of that would not be indexed manually, could be located by automatic indexing. So some silence factors are eliminated, at the price, perhaps, of a certain noise (if the located texts do not deal sufficiently with the idea looked for). And the irregularities well known in natural language obviously risk being silence factors, even if the depths of linguistic processing correct the effects of some of them. Under these circumstances we can understand the interest of combining the two methods of indexing, and of using them to their best advantage, which assumes on one hand series of experiments, and on the other hand the enriching of the software's dictionaries with the proper terms for the reference area of the texts, as well as their semantic relations.

Accessing images and documents from other databases

In addition, our information system gave access to 900 photographs of Delphi recorded on our videodisc *Images de l'archéologie* and which belong to the *Centre de Documentation Photographique et Photogrammétrique* (CDDP, CNRS-Université de Paris I); it is planned to add to these analog images with digital images, complementary photographs and, especially, maps, plans and drawings of the buildings on the site.

We can link these images to each of the documents in the base and this is what we have begun to do for the texts already recorded. During the lookup of the results of the querying, the text and the images illustrating it can be shown at the same time: the lookup of the texts gives access to that of the images.

To this group of texts linked to images, we have added documents from researchers files or from other databases using different softwares; with SPIRIT documents whose text is made up only of descriptors can be queried in natural language. The same question can concern all the types of document, or only one category. We could, for example, ask to query the bibliography of the Treasure of the Athenians, or ask the question "the mosaic at Delphi" and find texts dealing with this subject as well as analyses of the mosaics at Delphi taken from the factual databases on *Classical and Hellenistic mosaics throughout the Greek world* or on *Imperial and paleochristian mosaics in Greece*, or even items from the photo library of the *Centre de Recherche sur la Mosaïque* (CNRS) or from the *Centre de Documentation Photographique et Photogrammétrique*. The current situation of the experiment is that documents are extracted from databases run on UNIX with the SIGMINI software (Ecole des Mines de Paris) and from researchers files run on RapidFile with MSDOS; nothing prevents us, obviously, from using other softwares, provided that we can extract ASCII recordings which have information clear enough to allow automatic indexing and comprehensible when queried. If it is impossible to find all the richness of the information contained in different bases when it is brought together here (particularly information retrieved by the structural relations between descriptors in the same document, relations which can be important in some factual databases), a natural language interface allows us to simplify the questioning since, here, there is only one base to query and the system lets us substitute a single syntax - which is very simple - for the different softwares with which these bases had originally been made. This type of system allows us to integrate heterogeneous types of document and to query them in different ways, as we shall see.

Different ways of interactive querying

The querying of texts and documents can be coupled with the display of images which are linked to them. In a future version of the system, the images themselves should be able to be entry points and allow the querying of texts and notices when being leafed through: indeed, this function, which we developed for another software, is complementary to document-based retrieval; it allows other needs to be met and other types of users to be satisfied. Finally, at the present, the user can choose, from the texts which are offered, a sentence, a paragraph, or even a whole text which corresponds precisely to what he is looking for, and use it to automatically re-ask a question which will give him access to texts and notices which are themselves linked to images.

So, the functions of the programme, allowing the management of images and interrogation and navigation means which are both powerful and flexible, further enrich the interactive character of this multimedia querying. We also anticipate the case - which occurs frequently - of a query whose content takes shape during the interrogation: in fact, the user can explore the base by progressively refining his question, or go from one concept to another - whether this is done with a text or an image - following his own development. If the links between texts and documents are obviously put into the system each time a document is entered into the base, the links between documents need not be stated in advance since it is the system which calculates them, by automatic indexing, following the requests of the user and only taking into account his own point of view: here we reach particularly interesting functions of dynamic hypertext.

The necessary qualities of an information system; the limits of the experiment, the stakes of the research.

More generally, what qualities should we expect from an information system of this type so that it meets its aims?

The first is certainly the richness of the information contained, from a qualitative as well as quantative point of view. This requirement presupposes a judicious selection of documents, and, where long texts are concerned, their division into units appropriate to this type of querying. From a technical point of view, the quality and quantity of the information to take into account presupposes, almost necessarily, that the system can deal with heterogeneous data, coming from different environments. It also presupposes, particularly throughout the cultural domain and especially in those of Archaeology and History of Art, the management of multimedia data: images are essential, photographs and graphic documents, to which it is beneficial to be able to add moving images and even sound.

Moreover, if we speak here, by habit, of documents, it should really be a question of hyperdocuments, in the sense that J.P. Balpe (1990) gave to this term[3]. Indeed, even if the programme retains the idea of document for the part of the text divided up and characterised by descriptors, the querying allows us to navigate between bits of these documents - the parts of the text considered to be the most pertinent and which are not divided a priori but are determined by the automatic indexing in relation to each question.

3 "A hyperdocument is all computerised information content whose main characteristic is to not be subjected to a defined reading beforehand but to allow a more or less complex whole, which is more or less diverse, with more or less personalised reading (...) A hyperdocument is therefore all information content made up of nebulous fragments whose sense is built up, through computing tools, by each of the routes the reading determines". Cf.J.P. Valpe, *Hyperdocuments, hypertexts hypermédias*. Paris 1990, p.6.

Other information units - parts of the text chosen, as we have seen, by the user and which can be different from those which allowed the selection of the text - can be used to restart the querying without having been determined by the information system's author. These images themselves, taken in isolation or grouped in sequence, whatever their type, also constitute information units. So, the totality of the information obtained during querying comes from distinct units, linked, on the one hand, by the dynamic links which are created by the automatic indexing in relation to each question, and, on the other hand, thanks to the static links which have been pre-stated between the texts and the images by the authors of the system, links which add to the richness of the registered information.

The second quality, complementary to the first, is the relevance of the information given as an answer. For this, it is first of all necessary that the texts obtained contain the information looked for, and we have seen that this objective presupposes a great deal of text preparation before they are entered into the base and also depends on the richness of the dictionaries installed for the automatic indexing, this is especially the case in cultural domains; we have also seen that turning to manual indexing means that we must compensate for some insufficiencies. We see that in spite of its richness this type of natural language processing which extracts keywords does not solve all types of problem: we know that several teams are working on the extraction of text knowledge in natural language - a more advanced stage of what we used to call content computing - in the more or less near future this research will see important developments whose results will considerably enrich these systems. Relevance also concerns the links which the author pre-states, as we have seen, between the texts - or parts of text - and the images; this problem is more complex than it seems: because we reach the image by different methods, directly by leafing - and the image will then direct us to one or several texts - or indirectly through one or several texts, each being able to direct us to several images, it becomes necessary, at least, to be able to qualify these links in order to inform the user of their nature (information that can be introduced in some hypertext programmes by playing with the different "buttons"). Moreover, as the user comes to a fragment of text by a "point of view" which only depends on dynamic relations, how can we be certain that the text-image relations, which themselves must be pre-stated, will be relevant in all cases?

And, more generally, how can we evaluate the degree of relevance of the answers given by such a system? Already, in querying by request, the rates of silence and noise can not be measured in the same way in a probabilist system, such as SPIRIT, and in a Boolean system; but this takes another dimension when we query a base by hypertext type navigations, and goes even further with hypermedia, where the user is completely free to choose his own path at each stage. Of course the "right " information is that which meets the users needs, but in relation to what "requests" and what type of public must we evaluate the relevance of the "answer" and user satisfaction? The heterogeneity that we took as an objective further complicates the problem, but, in a specialised text information perspective, this criteria must keep its full value. Moreover, at the present time, evaluation tools and methods for comparing different means of accessing information do not exist: we can see that there are still problems to be solved that this experiment can shed some light on.

The heterogeneity of the target publics raises another question about relevance: the documents provided as an answer must correspond to the "level" of the user; indeed the Guide bleu will be of as little interest to a specialist as certain excavation publications will be to the public at large. The correspondence of terms used in the question and in the texts can provide a certain filtering (for example, specialists will speak of "cnemids", others of "greaves") at least so that we do not bring into play the reformulation with synonyms, but this solution is often not enough. No document characterisation has yet been attempted, because of the problems it presents.

In third place the system must be ergonomic less it lose the greatest reason for its interest to different publics; notably in the presentation of texts and images, work remains to be done to have a more pleasant display than in the screen windows; the present state of the technology already allows for a simple querying, but it is clear that the really multimodal interfaces of the future will bring considerable advantages to these information systems, at least when they are standardised enough to be accessible on a large range of desktop computers.

The fourth quality for such a system is a real interactivity; for that different types of access must be offered: access by texts only, or by factual criteria, or even by a combination of these two means; access by images, or by documents selected beforehand. This system should not be designed according to its internal logic - essentially of a technical nature - but rather according to the logic of its use: if we can say a priori, on the one hand, that interrogation by asking a question is better suited to the user who knows what information he is looking for and who can formulate the question during a session where he can have some help if needed, and, on the other hand, that a hypertext (or hypermedia) navigation is better suited to a less precise querying, we understand that the researcher prefers interrogation while navigation will be better suited to a public who is trying to become cultured. Nonetheless, it is important to allow changing from one mode to another during the same session because the same user can, according to the subject, need a navigation stage before asking his question, or inversely, the lookup of information retrieved by a question could lead to navigation. In addition, we can envisage allowing the user to become the author of his own system, but it would be necessary to find a way of protecting the database (information units as well as their static links).

In fifth place, the system must be open and allow the updating of the base, and its enrichment by the importation of new documents, and the exportation of some documents to other environments (word-processing, image files). If this technical possibility is offered, it is obviously up to those in charge of the information system to decide what operations are authorised for what users and in what context.

Finally, an information system done in such a way will be even more easy to diffuse if it is set up with widespread hardware and software, and/or it is able to be used with several types of configuration.

We understand that these requirements are not easy to satisfy. The last, especially, raises real technical problems: up to what point, really, can such a system be independent from the computing environment where it is used? For this experiment, we used a particular software, even if it works with several operating systems (for the moment Dos/Windows, Unix and OS/2). And, for the data itself, we principally came up against record format problems, whether this concerned text, image - still or moving - or sound, and their compression standard.

If these technical problems are not simple, they are at least on their way to being solved, in the longer or shorter term. However, more fundamental problems remain, such as "what information for what public?". If, as we know, building up data destined for well identified specialists is no easy thing, what can we say about working with the public "at large", which is, above all, a diversified public and even more difficult to "model"? Targeting a public which is heterogeneous to this point can seem like wanting to do the impossible: the research is far from finished, after a design period and the perfecting of the tools, it must go through a long experimental period with different types of public, who are the only ones who can give the necessary validation to this work.

It can seem doubly paradoxical to suggest creating "general" and "multi-function" systems at a time where, generally, one prefers to multiply the products in targeting them for specific types of use - but this type of evolution can only take place once the product has

reached a certain maturity, which is not the case today - and at a time where we are realising that to be worthwhile and therefore profitable, information products must be specialised and give a great deal of added value to the basic information they deal with. However, on the one hand, this last statement is only valid for information products destined for specialists, on the other hand, the idea of profitability cannot be applied in the same way to Human Sciences - particularly in the cultural domain - as it is to Economy or Finance, domains where information is given its true worth. Moreover, the products that we have studied here are essentially destined for a wider public: without being able to substitute for them, they will find themselves in competition with products already on the market - whether this concerns books, audiovisual programmes, or even the first creations on optical disks; consequently, their necessary added value[4] must be in the quantity and the quality of the information they offer, combined with the richness of interactivity when querying.

If it is a success, the integration of texts from different origins and published in a traditional manner will greatly contribute to the richness of information of these new products; and inversely, the new technology will allow the valorisation of the existing cultural heritage. And the researchers will have a double interest in collaborating in the preparation of these multimedia creations: on the one hand these systems can provide the researchers with an appreciable way of distributing the results of their work, displayed in textual form or even in databases in some cases; on the other hand, these systems will give researchers a basic documentary resource, which they can then use for their own studies, even in the development of more specialised information products. At the present time, the researchers are not on "speaking terms" with the few products that they have, they think that they lead to popularisation; however, they do not hesitate to seek out popular works or even those destined for tourists in order to find the high quality colour photos that they need for their research and teaching, and that, for reasons of cost, they do not find in scientific publications: multimedia products which are well done provide them with reservoirs which are even more extensive. They will also discover the new practices which allow the querying and manipulation of these vast groups of data, henceforth largely freed from the constraints of time (in accessing) and place (where they are stored) and enriched by the power of the processing offered.

In order that these products are not just distributed in small number in the universities and libraries, they must also interest a larger public who will ensure their profitability; for this type of public, the richness of the information will not be enough to make these products competitive: as well as having interesting subject matter, they must also offer greater interactivity than the present products, which, often, only reproduce electronically the traditional querying means. In this respect, too, we are only at the beginning of the research which is being carried out in different countries, and a study of the different uses can only be based on a very small number of experiments.

On one hand, the progress in multimedia microcomputing opens a production and consumer market which is as much document information as cultural information; on the other hand, in order to widen the marketing of their products to the public at large, the computer manufacturers need to package them with attractive multimedia interactive products: several factors can therefore help the development of these new types of cultural information products, and we emphasise the interest that they could rouse if they met the needs of the public, as well as satisfied a part of the need for information on behalf of the specialists for teaching and research (whether this be a single product or a series giving the same basic information in different versions appropriate to each type of use

4 Now, we speak more and more of "value added goods".

and user). It is therefore worthwhile to study their design for they constitute a great stake, as much for the world of electronic editing as for that of the communication of cultural information.