Full Text / Image DBMSs

Robert Rowland

When dealing with multimedial systems, of which image DBMSs are the clearest instance, discussion tends to focus on the implications of translating between one medium and another and on the relationship, for example, between images and the text used to describe them. Other chapters in this volume examine this relationship, the development of thesaurus-based systems, and the implications of allowing the end-user some control over the ways in which text is used to describe and retrieve the information contained in images.

Concentration on the differences between media and their implications may, however, lead to neglect of a number of more general issues arising out of the representation of information within database systems as such. When these are taken into account, multimedial systems can be approached as a special case among database systems rather than as a distinct kind of system, and the difficulties involved in controlling the process of translation between media can be regarded as a particular instance of the more general problem.

In such a context it is helpful to regard the representation of information within any database system as involving three distinct levels:

a) raw information
b) description (of relevant aspects of (a))
c) rules (governing (b)).

The relationship between these three levels form the substance of discussions about the modelling of historical data[1] and need not be expanded on here. Within such a framework, it is helpful to distinguish between closed and open systems. In the former, all three of these levels are pre-defined, and the user is constrained to work within the set of rules worked out by the author and embodied in the design of the system. In the latter, the end-user may be allowed to tailor the description of the raw information to his own requirements (while remaining within the pre-defined set of rules), or even to intervene directly at the level of the rules, changing the criteria according to which relevant aspects of information are selected for information and embodied in textual keys[2].

In a 'unimedial' system containing only textual information, the above schema would take the following form:

a)	raw information	full text of the documentary corpus
b)	description	structured description of (a)
c)	rules	rules for abstracting information
		record structure
		thesaurus of descriptive terms.

This is of course familiar ground for all those engaged in historical computing and I shall not rehearse old arguments, in particular those regarding the respective merits of full transcription / text retrieval versus more standard (i.e. more structured) database systems

[1] Cfr., in this same series, D. Greenstein (Ed.), *Modelling Historical Data*, Scripta Mercaturae, St. Katharinen, 1991

[2] Cfr. the chapter by Wendy Hall and Frank Colson, pp. 87-96.

and procedures. It is, however, worth remembering that for a system to be open the raw information must be contained within, or be accessible from within, the database. A system where each record consists of a structured description of a document is of necessity relatively closed, and the end-user's freedom of intervention is limited, for example, to the choice of the keys used to describe and retrieve a given structured (and usually very selective) summary description of any particular document.

It is perhaps for this reason that the question of 'openness' has been raised with particular reference to multimedial systems where the 'raw information' (a bitmapped image) is accessible from within the system. In terms of the above schema, a multimedial system of this kind would be formally equivalent to a hypothetical database system based on a verbatim transcription of an entire documentary corpus[3], and it is therefore understandable that there should be considerable discussion of the possibilities and implications of making it more 'open'. The same would have happened in relation to text-based systems if verbatim transcription of an entire documentary corpus had ever come to be regarded as an appropriate method for entering historical information into the computer and making it available to other scholars[4]. But since bitmapped images are much less costly than verbatim transcription[5] it is natural that the discussion should have arisen here and that it should have focussed on a particular case of structured description — the 'translation' of visual information into verbal descriptions — and on its inherently arbitrary nature.

Cost is one of the reasons why verbatim transcription has not been generally adopted; another is the fact that information stored in such a form is unwieldy, and that unless the system includes a structured description of each document it runs the risk of being about as useful as a map drawn to a scale of 1:1. If the question of cost is disregarded for the time being, everything hinges on the criteria underlying the structured description. If it were possible to design a system comprising a) verbatim transcription of the entire

[3] Perhaps the best-known example of such a system is the Earls Colne database created by Alan Macfarlane and his associates in Cambridge between 1971 and 1983. This database contained the entire (pre-edited) text of all the surviving documentation regarding the parish of Earls Colne, in Essex. The information considered relevant (by Macfarlane) was parsed into a relational database system (CODD), whence it could be accessed by means of a specially-designed query language (CHIPS). An end-user with different analytical preoccupations could, in theory, have re-edited the machine-readable text according to his own interests and criteria and then proceeded to form an entirely different database. For a description of the project cfr. Alan Macfarlane, 'The Origins and Organization of Research', (Appendix H of Sarah Harrison et al., Reconstructing Historical Communities with a Computer. Final Report to the Social Science Research Council, Department of Social Anthropology, Cambridge, 1983). I am grateful to Alan Macfarlane for allowing me to cite this unpublished report.

[4] I am not of course referring, here, to the computer-based edition of historical texts, where verbatim transcription is indispensable. The problems arise when considering very extensive corpora, such as the notarial records of an early modern town. The cost of verbatim transcription would obviously be out of all proportion to the benefits that the scientific community could ever hope to obtain.

[5] In terms of time, material resources and storage requirements.

corpus; b) a structured — and editable — description of each document; and c) a set of rules specifying the criteria adopted in drawing up (or in modifying) the structured descriptions, we would have an acceptable 'open' text-based system[6]. Since verbatim transcription is extravagant, however, it would make a great deal of sense to replace it (within the above schema) by scanned images of the relevant documents. Each document would be represented, within the database, by a structured description. The textual keys used to retrieve any such descriptive record could equally give access to the raw image of the document, and provided adequate guidelines had been laid down, the end-user could revise, refine and expand the description in accordance with his analytical preoccupations and his own assessment of the document's meaning and significance.

In most cases, retrieval would be effected by means of a set of textual keys, controlled by some kind of thesaurus, and these could, as has been mentioned, be redefined by the end-user. It would even be possible, in some cases, to remove such 'subjective' criteria from the database altogether, and to allow interactive information retrieval by means of 'textual keys' provided by the document itself. Such a solution would require that each document should already contain some kind of descriptive summary of itself (e.g. the 'sentence' in an Inquisition trial record), and that a verbatim transcription of this description be included as a field in the database record. With suitable information retrieval software the user could interactively build up and refine a free text query containing a set of terms which are effectively present and associated with one another in the documents themselves. Users with different interests and preoccupations could, with such a system, interrogate the same database in different ways; and the possibility of accessing an image of the original document from within the database can act as an aid to refining the query and as a check on the appropriateness of the set of terms finally chosen.

An illustration of the possibilities just outlined can be provided by a project on the records of the Portuguese Inquisition (currently at the planning stage) which I am developing in association with the Center for Medieval and Renaissance Studies of U.C.L.A. The 40 000 records of the Portuguese Inquisition (1536-1820), of which about 35 000 are full trial records, constitute a unique source for the social historian. At present they can only be accessed by means of an obsolete and inadequate catalogue drawn up in the 19th Century[7].

A typical trial record (whose length may vary between ten and several hundred pages) contains the following main elements: a) transcripts of testimony against the accused, drawn from depositions taken in other trials or from direct denunciations; b) an order for the accused to be arrested; c) the testimony of the accused about himself and his family, and about the accusations made against him; d) a formal statement of the accusation; e)

[6] Such a system would correspond very closely to the Earls Colne project described above. It is ironical that this project was brought to an end, largely because of the cost of continuing it on a mainframe, just as the personal computer was about to revolutionize the prospects for historical computing. It would not be difficult to resurrect the CODD/CHIPS system and run it on one of the more powerful microcomputers available today.

[7] An adequate catalogue now exists for the Évora tribunal, but even in this case only manual searches are possible.

further depositions of witnesses and the accused (possibly under torture); f) the conclusions of the court (sentence), including a final statement of the proven accusations; g) (usually) the statement of abjuration made by the accused on publication of the sentence; and h) follow-up documents (consignment of the accused to the appropriate civil authorities for execution, banishment or being sent to the galleys; release of the accused after a period of imprisonment or religious instruction, etc.). Except in the earliest trials, the prisoner was not informed about the identity of his accusers. But he was given the opportunity of naming all his enemies, and thus of disqualifying the testimony of at least some of these accusers. Since the accusations of enmity were investigated by the court, trials in which the prisoner tried to name all his potential accusers contain a great deal of circumstantial evidence of interest to the social historian.

Verbatim transcription of approximately 2 million pages of difficult manuscript text would obviously be prohibitively expensive, and quite pointless; but, given the range of information contained in these trial records, no structured description could possibly correspond to the full range of interests displayed by those wishing to consult them. It is therefore proposed to construct a database each of whose records would contain an extended structured description of an individual trial[8], including a verbatim transcription of the sentence. Each such record will be linked to a bitmapped image (stored on CD-ROM) of the entire original document[9]. The user will be able to inspect on-screen, or print, the bitmapped reproduction of any document retrieved. The structured description could be edited at any time by the user so as to include aspects of particular relevance to him. Records can be retrieved by an interactive procedure making use of any combination of terms present either in the structured description or in the transcribed text of the sentence[10], and the actual contents of the document (in this case, as summarized in the sentence) can serve as a check on the textual keys used for information retrieval.

[8] The project will build on the experience gained with an earlier project started between 1982 and 1987 at the Gulbenkian Institute of Science, Lisbon. Each trial record was there represented by an extended structured description of up to 60 fields. This number will be considerably reduced in the proposed project.

[9] Although no final decision has been taken on the software to be used, the project has been designed to make use of the capabilities of the database / information retrieval system MUSCAT, which in its current version allows both interactive information retrieval and cross-referencing, from within the database, to bitmapped or analog images (on CD-ROM or videodisc).

[10] For a description of the information retrieval system see M. F. Porter, 'Implementing a Probabilistic Information Retrieval System', *Information Technology: Research and Development*, 1, 1982, 131-156.

Halbgraue Reihe
zur Historischen Fachinformatik

Herausgegeben von
Manfred Thaller
Max-Planck-Institut für Geschichte

Serie A: Historische Quellenkunden

Band 14

Erscheint gleichzeitig als:

MEDIUM AEVUM QUOTIDIANUM

HERAUSGEGEBEN VON GERHARD JARITZ

26

Manfred Thaller (Ed.)

Images and Manuscripts in Historical Computing

Max-Planck-Institut für Geschichte
In Kommission bei
SCRIPTA MERCATURAE VERLAG

St. Katharinen, 1992

Printed in Germany
Druck: Konrad Pachnicke, Göttingen
Umschlaggestaltung: Basta Werbeagentur, Göttingen
ISBN: 3-928134-53-1

Table of Contents

Introduction

Manfred Thaller

This book is the product of a workshop held at the International University Institute in Firenze on November 15th, 1991. The intention of that workshop has been to bring together people from as many different approaches to "image processing" as possible. The reason for this "collecting" approach to the subject was a feeling, that while image processing in many ways has been the "hottest" topic in Humanities computing in recent years, it may be the least well defined. It seems also much harder to say in this area, what is specifically important to historians, than to other people. In that situation it was felt, that a forum would be helpful, which could sort out what of the various approaches can be useful in historical research.

To solve this task, the present volume has been produced: in many ways, it reflects the discussions which actually have been going on less, than the two companion volumes on the workshops at Glasgow and Tromsø do. This is intentional. On the one hand, the participants at the workshop in Firenze did strongly feel the need to have projects represented in the volume, which were not actually present at the workshop. On the other, the discussions for quite some time were engaged in clarifying what the *methodological* issues were. That is: what actually are the topics for scholarly discussion beyond the description of individual projects, when it comes to the processing of images in historical research?

The situation in the area is made difficult, because some of the underlying assumptions are connected with vigorous research groups, who use fora of scholarly debate, which are only slightly overlapping; so, what is tacitly assumed to hold true in one group of research projects may be considered so obviously wrong in another one, that it scarcely deserves explicit refutation.

We hope, that we have been successful in bringing some of these hidden differences in opinion out into the open. We consider this extremely important, because only that clarification allows for a fair evaluation of projects which may have started from different sets of assumption. So important, indeed, that we would like to catalogue here some of the basic differences of opinion which exist between image processing projects. The reader will rediscover them in many of the contributions; as editor I think however, that sumarizing them at the beginning may make the contributions — which, of course, have been striving for impartiality — more easily recognizable as parts of one coherent debate.

Three basic differences in opinion seem to exist today:

(1) Is image processing a genuine and independent field of computer based research in the Humanities, or is it an auxiliary tool? Many projects assume tacitly — and some do so quite outspokenly — that images on the computer act as illustrations to more conventional applications. To retrieval systems, as illustrations in catalogues and the like. Projects of this type tend to point out, that with currently easily available equipment and currently clearly understood data processing technologies, the analysis of images, which can quite easily be handled as illustrations today, is still costly and of uncertain promise. Which is the reason why they assume, that such analytical approaches, if at all, should be undertaken

as side effects of projects only, which focus upon the relatively simple administration of images. Their opponents think, in a nutshell, that while experiments may be needed, their overall outcome is so promising, that even the more simple techniques of today should be implemented only, if they can later be made useful for the advanced techniques now only partially feasible.

(2) Connected to this is another conflict, which might be the most constant one in Humanities data processing during the last decades, is particularly decisive, however, when it comes to image processing. Shall we concentrate on levels of sopbistication, which are available for many on today's equipment or shall we try to make use of the most sophisticated tools today, trusting that they will become available to an increasingly large number of projects in the future? This specific battle has been fought since the earliest years of Humanities computing, and this editor has found himself on both sides at different stages. A "right" answer does not exist: the debate in image processing is probably one of the best occassions to understand mutually, that both positions are full of merit. It is pointless to take permanently restrictions into consideration, which obviously will cease to exist a few years from now. It discredits all of us, if computing in history always promises results only on next years equipment and does not deliver here and now. Maybe, that is indeed one of the more important tasks of the *Association for History and Computing*: to provide a link between both worlds, lending vision to those of us burdened down by the next funding deadline and disciplining the loftier projects by the question of when something will be affordable for all of us.

(3) The third major underlying difference is inherently connected to the previous ones. An image as such is beautiful, but not very useful, before it is connected to a description. Shall such descriptions be arbitrary, formulated in the traditionally clouded language of a historian, perfectly unsuitable for any sophisticated technique of retrieval, maybe not even unambigously understandable to a fellow historian? Or shall they follow a predefined catalogue of narrow criteria, using a carefully controlled vocabulary, for both of which it is somewhat unclear how they will remain relevant for future research questions which have not been asked so far? — All the contributors to this volume have been much to polite to phrase their opinions in this way: scarcely any of them does not have a strong one with regard to this problem.

More questions than answers. "Image processing", whether applied to images proper or to digitalized manuscripts, seems indeed to be an area, where many methodological questions remain open. Besides that, interestingly, it seems to be one of the most consequential ones: a project like the digitalization of the *Archivo General de Indias* will continue to influence the conditions of historical work for decades in the next century. There are not only many open questions, it is worthwhile and neccessary to discuss them.

While everybody seems to have encountered image processing in one form or the other already, precise knowledge about it seems to be relatively scarce. The volume starts, therefore, with a general introduction into the field by J. v.d. Berg, H. Brandhorst and P. v. Huisstede. While most of the following contributions have been written to be as self supporting as possible, this introduction attempts to give all readers, particularly those

with only a vague notion of the techniques concerned, a common ground upon which the more specialized discussions may build.

The contributions that follow have been written to introduce specific areas, where handling of images is useful and can be integrated into a larger context. All authors have been asked in this part to clearly state their own opinion, to produce clearcut statements about their methodological position in the discussions described above. Originally, four contributions were planned: the first one, discussing whether the more advanced techniques of image processing can change the way in which images are analysed and handled by art historians, could unfortunately not be included in this volume due to printing deadlines: we hope to present it as part of follow up volumes or in one of the next issues of *History and Computing*.

The paper of M. Thaller argues that scanning and presenting corpora of manuscripts on a work station can (a) save the originals, (b) introduce new methods for palaeographic training into university teaching, (c) provide tools for reading damaged manuscripts, the comparison of hand writing and general palaeographic studies. He further proposes to build upon that a new understanding of editorial work. A fairly long technical discussion of the mechanisms needed to link images and transcriptions of manuscripts in a wider context follows.

F. Colson and W. Hall discuss the role of images in teaching systems in university education. They do so by a detailed description of the mechanism by which images are integrated into Microcosm / HiDES teaching packages. Their considerations include the treatment of moving images; furthermore they enquire about relationships between image and text in typical stages in the dialogue between a teaching package and a user.

W. Hall and F. Colson argue in the final contribution to this part the general case of open systems, exemplifying their argument with a discussion of the various degrees in which control about the choices a user has is ascertained in the ways in which navigation is supported in a hyper-text oriented system containing images. In a nutshell the difference between "open" and closed systems can be understood as the following: in an "open system" the user can dynamically develop further the behaviour of an image-based or image-related system. On the contrary in static "editions" the editor has absolute control, the user none.

Following these general description of approaches, in the third part, several international projects are presented, which describe in detail the decisions taken in implementing "real" image processing based applications, some of them of almost frigthening magnitude. The contributors of this part were asked to provide a different kind of introduction to the subject than those to the previous two: all of them should discuss a relatively small topic, which, however, should be discussed with much greater detail than the relatively broad overviews of the first two parts.

All the contributions growing out of the workshop came from projects, which had among their aims the immediate applicability of the tools developed within the next 12 - 24 months. As a result they are focusing on corpora not much beyond 20.000 (color) and 100.000 (b/w) images, which are supposed to be stored in resolutions manageable within $\leq$ 5 MB / image (color) and $\leq$ 0.5 MB / image (b/w). The participants of the workshop felt strongly, that this view should be augmented by a description of the rationale behind

the creation of a large scale project for the systematic conversion of a complete archive. The resulting paper, by P. González, describes the considerations which lead to the design of the *Archivo General de Indias* project and the experiences gained during the completed stages. That description is enhanced by a discussion of the strategies selected to make the raw bitmaps accessible via suitable descriptions / transcriptions / keywords. A critical appraisal, which decisions would be made differently after the developments in hardware technology in recent years, augments the value of the description.

The participants of the workshop felt furthermore strongly, that their view described above should be augmented by a description of the techniques used for the handling of images in extremely high resolution. A. Hamber's contribution, dealing with the *Vasari project*, gives a very thorough introduction into the technical problems encountered in handling images of extremely high quality and also explains the economic rationale behind an approach to start on purpose with the highest quality of images available today on prototypical hardware.

As these huge projects both were related to institutions which traditionally collect source material for historical studies, it seemed wise to include also a view on the role images would play in the data archives which traditionally have been of much importance in the considerations of the AHC. E.S. Ore discusses what implications this type of machine readable material should have for the infrastructure of institutions specifically dedicated to Humanities computing.

Image systems which deal with the archiving of pictorial material and manuscript systems have so far generally fairly "shallow" descriptions. At least in art history, moreover, the rely quite frequently on pre-defined terminologies. G. Jaritz and B. Schuh describe how far and why historical research needs a different approach to grasp as much of the internal structure and the content of an image as possible.

Last not least R. Rowland, who acted as host of the workshop at Firenze, describes the considerations which currently prepare the creation of another largescale archival database, to contain large amounts of material from the archives of the inquisition in Portugal. His contribution tries to explore the way in which the more recent developments of image processing can be embedded in the general services required for an archival system.

This series of workshop reports shall attempt to provide a broader basis for thorough discussions of current methodological questions. Their main virtue shall be, that it is produced sufficiently quick to become available, before developments in this field of extremely quick development make them obsolete. We hope we have reached that goal: the editor has to apologize, however, that due to the necessity to bring this volume out in time, proofreading has by neccessity be not as intensive as it should have been. To which another shortcoming is added: none of the persons engaged in the final production of this volume is a native speaker of English; so while we hope to have kept to the standards of what might be described as "International" or "Continental" English, the native speakers among the readers can only be asked for their tolerance.

Göttingen, August 1992