SECOND INTERNATIONAL CONFERENCE

AUTOMATIC PROCESSING OF ART HISTORY DATA AND DOCUMENTS

PISA

SCUOLA NORMALE SUPERIORE

SEPTEMBER 24-27, 1984

PROCEEDINGS

EDITED BY LAURA CORTI AND MARILYN SCHMITT

SCUOLA NORMALE SUPERIORE
PISA

1984

THE J. PAUL GETTY TRUST
LOS ANGELES

PUBLISHED BY REGIONE TOSCANA

1. Integration in the last years

Over the last twenty years, with developments in the DP environment and the introduction of new ideas, the concept of "integration" has assumed several different meanings.

At the beginning (in the sixties), it was common practice to have spread data over different files, and programs were tightly related to the data organization and file structures. This environment is normally referred to as the "conventional DP environment". The concept of integration in such a context was essentially a set of cooperating procedures with data passing from one to another, and "integration" meant simply: "program integration".

In the seventies, the need for data integration grew, and the concept of "data base" became increasingly common. Initially, a number of "integrated file systems" were developed, giving origin to the first integrated Data Base Management Systems (DBMS); the theory was developed afterwards.

In the mid seventies the myth of the "integrated data base" was one of the major goals pursued by public and private enterprises (or perhaps one of the principal "paper makers" for the theorists). In fact, as shown by a European survey (/DBE 79/), /Signore 80/), integrated data bases were very rarely

implemented, integration often being limited to certain application areas. The data base concept, however, was a very important stimulus in attaining greater standardization and limited or controlled data redundancy, achieving, at least to a limited extent, "data integration". In the late seventies the first interest in data base design appeared.

From 1980 to 1984, in expectation of effective relational DBMSs, great attention was paid to data base design methodologies, and the conceptual schema design was identified as a major tool toward an integrated view of the real word. In the meantime, a new concept of integration was introduced: "media integration". This is a view of the data not simply as formatted strings of characters, but also texts, images and voice. In these last years, the viewpoints of other areas, such as Artificial Intelligence (semantic data models, expert systems, natural language understanding, etc.) have also been of relevance.

At present, we have a situation where:

- relational DBMSs are now available on the market, but their performances must be carefully tested;
- Information Retrieval Systems (developed outside the data base area) are still very important, even though it seems that no new special features have been added to them. As a matter of fact, in the IR field we are still concerned with traditional problems, such as thesauri, automatic indexing, feedback retrieval. Perhaps more important, the present IR systems cannot handle sophisticated data models.
- The DBMSs and IR systems are going to be integrated, especially within the Office Information Systems field. In this context, a lot of work is now being done on the integration of texts, voice, images. It is worthwhile noting that there are a number of ideas on how such an inte-

gration should be achieved.

2. Integration for Art History Data

The computerization of art history data needs integration in many senses, i.e., it needs data integration, media integration, and integration between different disciplines. It has been noted (/Arms 84/) that the level of integration needed is dictated by the level of the user.

In fact, in the computerization of art history data, several specialized data banks are often involved -- for authors, biographies, bibliographies, and so on. An appropriate framework to represent the integration of these different topics is the conceptual model. It seems that a first, necessary step towards integration is to represent the relationships of interest between objects, authors, locations, patrons, owners, original documents and so on. Obviously, such an integrated approach does not necessarily mean that the final goal is to build up a single, large integrated data base. The main aim is to understand the interrelations between different entities, and to set up an appropriate representation of the perceived reality. This may be especially useful for the exchange of data and experiences.

Moreover, once we enter into detail, some of these entities are very complex, and cannot be accommodated by simple flat data models, without a high degree of data redundancy. As an example, try to give a definition of "object" which may fit not only one simple object, but also a set of closely related although distinguishable objects, which from a certain viewpoint may be identified as a unique object (think of a dinner service, for example). Experience clearly shows that the kind of data which must be managed exhibit very complex relationships, and can only be normalized and formalized to a certain extent.

An important aspect which must be taken into account is media integration, as image management would add much "value" to an art history management system. However, as no image processing seems to be required, we can presume that at least for the moment the facilities provided by videodisk technology give satisfactory results. Voice integration does not seem to be a major commitment.

It is worthwhile mentioning a common query language to access different data bases managed by different software as another area of integration, which may be especially relevant for potential users.

As far as IR systems are concerned, there is a proposed ISO standard, based on the Common Command Language (CCL) developed for the EURONET-DIANE network (/Bartoli 81a/, /Bartoli 81b/, /Signore 83/). In the DBMS area, the SQL-like languages constitute a "de facto" standard.

In conclusion, leaving integration with different disciplines, e.g., Artificial Intelligence, as a future prospect, we may redefine the problem of integration as the integrated management of formatted and unformatted data, namely, to the integration of Data Base Management Systems and Information Retrieval Systems.

From my point of view, this is really the key item. In most cases, the user accessing the data base will have in mind a vague idea of the "objects" he is looking for, in the sense that he is not able to specify exactly the values of all the attributes that identify the object. An Information Retrieval System (IRS), offering facilities such as free text searching and structured thesauri, may help in formulating queries in a precise and exhaustive way, while a DBMS may accommodate for complex relationships between the documents.

3. Analysis of the projects reported in the CENSUS

Glancing quickly at the projects reported in the CENSUS (/CENSUS 84/), we can see that the kind of software adopted has been reported in 113 cases. In 42% of cases (48 out of 113) an IRS[1] was used, while in only 10% of cases, a DBMS [2] was adopted. The STAIRS system seems to be the most popular, as it accounts for 60% of IRS, or 26% of the total.

However, these data must be considered with care, as in many cases a single institution presented several projects, all using the same software. This is especially true as far as STAIRS, SELGEM and ISIS are concerned.

In addition, it may be supposed that the choice of the system has been often dictated by the computer facilities available. It often appears that some specific topics have been investigated to a very deep level. However, it may be noted that only rarely is there any indication that a complex data model has been supported, or that a conceptual schema has been developed.

In conclusion, it may be presumed that art historians encounter some difficulties in formalizing their data, and therefore their problems are more easily approached from the point of view of Information Retrieval Systems, especially when a single application area is considered. In this environment, a lot of attention is dedicated to the normalization of the vocabulary and setting up of structured thesauri, more than to the representation of the relationships between different application areas.

4. Experience with the Italian Catalog

The computerization of the Italian Catalog was begun

several years ago. After a first phase, when the approach was to store all the data in the computer, and query them via an IRS (namely STAIRS), it was found that the material input was unsuitable for effective retrieval. It was thus evident that a great effort at formalization and normalization was needed.

The first step was to stop thinking in terms of the target software, and start with a more abstract way of representing the reality: the conceptual schema. The methodology adopted was, to a large extent, inspired by the well known Entity Relationship Model (/Chen 76/), especially useful for its expressiveness as a graphic representation of relationships.

The conceptual schema constitutes the input for the subsequent work, in which an attempt is being made to resolve some of the major problems, at an abstract (i.e., software independent) level. During this phase, the conceptual schema is being refined, and different IRSs and DBMSs are being considered, in order to identify their merits and to decide on possible extensions and/or the need to integrate them.

The conceptual schema is also used in order to identify acceptable subsets of information to be provided by peripheral offices. It should be noted, however, that whereas standard design methodologies may be suitable for the billing of material or for general ledger applications, they generally do not fulfill the needs of a data base for our cultural wealth. In any case, it is clear that there are two major problems: the first is the normalization of the vocabulary, the second is the identification and representation of relationships. At present, single problems are examined individually, and the solutions proposed are verified against a significant sample. For simplicity's sake, in order to avoid the writing of ad hoc procedures, the test environment is an IRS. There are some indications that the final environment will be an integrated DBMS/IRS system.

5. Our experience shows that the conceptual schema design is a major tool towards data integration. It is of great help in identifying areas for normalization and the items to be used as links between different application areas. Moreover, the conceptual schema constitutes an excellent vehicle for communication between art historians and computer scientists.

The computerization of art history data seems to pose problems that cannot be entirely solved either by the DBMSs or by the IRSs available on the market: an integration of the facilities offered by both kinds of systems would provide a good tool to manage this kind of data. Integration with other media, especially as far as images are concerned, constitutes a valuable improvement.

As a future prospect, some issues from the Artificial Intelligence field may well lead to the implementation of much more sophisticated and powerful systems. However, even if it is possible to make reliable provisions, these enhancements are not foreseeable for the near future.

---

1. GRIPHOS, ISIS, MISTRAL-2, PARIS, PARIS-EASIS, SELGEM, STAIRS have been counted as IRSs.

2. DBASE2, FOCUS, IDMS, IMS/VS, MICROBASE, QBE, ADABAS, SYSTEM 2000 have been counted as DBMSs.

REFERENCES

/Arms 84/ W.Y. ARMS, "The museum prototype project: the search for integration", Automatic Processing of Art History Data and Documents, Pisa, Scuola Normale Superiore, September 24-27, 1984, L. CORTI (ed.).

/Bartoli 81a/ R. BARTOLI, G. A. ROMANO, O. SIGNORE, "Implementation of Common Command Language on STAIRS/VS - TLS. Computerized Legal Decisions. An interdisciplinary Approach", Proceedings of the International Study Congress on "Logica, Informatica, Diritto", C. CIAMPI and A.A. MARTINO (eds.), North Holland Publishing Company.

/Bartoli 81b/ R. BARTOLI, G. A. ROMANO, O. SIGNORE, "Realizzazione di un linguaggio comune di comandi per la rete EURONET-DIANE", Quaderni di Informatica e Beni Culturali, n. 4: Sistemi di trattamento di dati e immagini - Universita' degli Studi di Siena.

/CENSUS 84/ L. CORTI, (ed.), Computerization in the History of Art, Volume 1. Scuola Normale Superiore, Pisa, and The J. Paul Getty Trust, Los Angeles, 1984.

/DBE 79/ GMD (K. Supper, E. Gross, H. Muenzenberger, J. Walther, U. Baumgarten, M. Speake). INRIA (D. Potier, D. Goldwasser, B. Euzenat) NCC (D.R.A. Coan, J. M. Dransfield, K. B. Hannah, J. D. Lomax, D. P. Nicholls) CNR (O. Signore, C. Thanos, G. Gasparotto): "Experience of DBMS usage in Europe", Report to European Commission (1979).

/Chen 76/ P.P. CHEN, The entity relationship model: toward a unified view of data, ACM TODS, 1 (1) (1976) 9-36.

/Signore 80/ O. SIGNORE, C. THANOS, "EEC Database Project: Italian National Survey Report", Rivista di Informatica, X, 1, (Gennaio-Marzo 1980), pp. 69-85.

/Signore 83/ O.SIGNORE, R.BARTOLI, G.A.ROMANO, "La realizzazione di uno standard per i sistemi di information retrieval, l'esperienza STAIRS/VS-TLS". Congresso L'informatica giuridica e le comunita' nazionali e internazionali, Roma 9-14 Maggio 1983.