To appear in the proceedings of AHC'91
History & Computing
Odense, August 28-30 1991

From data structuring to data exchange: a simple path

Oreste Signore

CNUCE - CNR - via S. Maria 36 - 56126 Pisa (Italy)
Tel. +39 (50) 593201 FAX +39 (50) 576751
E.mail: EARN.bitnet ORESTE@ICNUCEVM.CNUCE.CNR.IT

Abstract

Data structuring, making use of the well established Database Design techniques, may lead to a clean model or art history objects. The solutions adopted by the Italian Institute for Cataloguing art objects (ICCD), are briefly outlined. The structuring of information has lead to the definition of a simple, but effective exchange format, overcoming the difficulties of establishing appropriate standards at physical, logical and conceptual level.
The experience made with 31 projects, who acted independently, on the basis of the frame of reference, and problems of standardization of language and descriptions, are discussed.

1 - Introduction

The structuring of art history data is not an easy task. However, it is necessary if we want to access and retrieve data avoiding "false drops". This implies that the information must be subdivided into small, semantically well defined chunks.
Unfortunately, this approach may lead to an indiscriminate proliferation of fields, and to excessively specialized data structures. It is therefore necessary to build up a frame of reference that will enable to "group", as far as possible, objects of the same type. Another key point is to identify with a specific tag each element of information.
This approach is necessary, and preliminary to the selection of the software and of the hardware. Never will exist any project, which will rely on the specific features of the target software, that will succeed in effectively storing, retrieving and exchanging data, if an accurate and clear data structure has not been defined.
This is also true when approaches like Information Retrieval or Hypertext, were the structuring of the information may appear as an useless or obsolete task, are selected.

2 - Italian Catalog Data Structuring

2.1 - Historical background

The first organisation of the Italian Catalog was based on a manual approach. Therefore, each object was described by a typewritten card, following some general rules. The basic ideas were very valuable, and all the subsequent work has been greatly influenced by the intellectual work that lead to the definition of the fundamental principles of the cataloguing rules.
The most relevant issues were:

- the identification of a reduced set of different cards, corresponding to different types of objects (art objects, archaeological objects, drawings, architecture, gardens, historical centre, etc.);
- the grouping of the information in several very general categories, like author, location, material, historical info, etc.;
- the topological arrangement of the catalogue cards.

On the other hand, it has to be pointed out that the cards were conceived for human usage, and therefore the various fields were to be filled in by the scholars, on the basis of their specific competence on the particular subject, following some general rules. As a matter of fact, the foreseen user of these cards was another scholar, able to understand the semantics of the content of the fields, and capable of identify possible inconsistencies, or interpret them in the correct way.
Around the mid '70, some attempts were made to directly transpose the cards into information retrieval systems. In spite of the initial enthusiasm, and the claims of the vendors, the results were very disappointing: as far as the quantity of data was increasing, the precision and recall factors were going down.
It soon became evident that the unsatisfactory results could not be ascribed to the peculiar systems (every product exhibits some strong and some weak points), but that a rethinking of the entire cataloguing schema was necessary, keeping in mind the constraints imposed by the automated treatment of the information.

2.2 - The very first step

The definition of the data structure started from the identification of the central role played by the object. Initially, a rich variety of object types was identified: single object, series, fragment, part of, etc.
After a while, it was realized that this kind of specialization was too complex, and would in fact constitute a barrier between the cataloguer and the user.
Finally, it was agreed to define a classification schema based on three different kinds of objects: simple objects, complex objects, aggregation of objects.
A simple object is an object such that all his attributes are pertinent to the whole object, and no components which may themselves be considered cataloguing objects may be identified.
A complex object may be either a simple object whose parts, physically or conceptually separable, exhibit some interesting peculiarities as cataloguing objects, either a set of objects which may be referred by a specific name.
The aggregation of objects arise when several objects are correlated on the basis of some conceptual criterion, but no name exists which identifies the aggregate.
It is obvious that the components of a complex object may be either simple either complex objects, and so is as far as the aggregate objects are concerned.

It is worthwhile to note that a specific object belongs to the different categories only on the basis of the quantity and the type of information: no list exists that specifies if a particular kind of object must be considered simple or complex or aggregate. The proposed model only establish a classification model, that is, the type of relationships that must be specified between the objects (a component of a complex object is an object itself), and the criteria of inheritance of the properties.

2.3 - The approach

The adopted approach was essentially based on the standard methodologies currently in use in the database design area, especially as far as conceptual design is concerned. As it is well known, the most popular approach is based on the Entity-Relationship model, and the conceptual model is independent on the software and obviously the hardware environment.

The first step has been the identification of the "basic" entities, like object, author, location, and so on. The identification of the relationships between these entities has been taken as the second step. This process lead to a simple, consistent model, were the object was playing a central role.

At first glance, the resulting model was not so different from a conventional "bill of materials" schema in a commercial enterprise. The very big difference was in the vagueness or uncertainty of some data (think for example to the problem of dates, multiple attribution, alternative names of the artists, etc.).

This lead to the identification of some "fundamental" entities and relationships, which are intrinsic to the representation of the real world, and of some "minor" relationships and entities, as those accounting for the name variants.

The analysis of this last and similar problems pushed us to the definition of some "authority files" as the only mean for normalizing the vocabulary, so insuring consistency of data.

Even if the Entity-Relationship model as been proved to be a very effective communication mean between end users and designers, it was found more productive to use an alternative way of representing information, that is, to switch to the conventional "cataloguing card" format.

From this point of view, the following choises were made:

- the information has been subdivided into small, semantically well defined, chunks;

- these chunks may be either a field, either a subfield of a structured field;
- each field may be defined as simple or structured;
- each field may be defined as repeating or non-repeating;
- each subfield may be a repeating or non repeating subfield;
- fields, either structured or unstructured, may be grouped into "paragraphs" in order to allow multiple occurrences of a set of fields.

Anyone familiar with the database design methodologies will easily recognize that, generally speaking, entities have been mapped onto paragraphs, (multivalued) attributes onto (repeating) fields, aggregate attributes onto structured fields.

It is also evident that the identification of a sequence of fields, with the characteristics of being repeatable and/or groupable, and references to "authority files", may be seen as the "linearization" of a non linear text.

Last, but not least, an effort has been made in order to maintain consistency between different cultural areas, so that semantically equivalent fields are identified by the same tag.

3 - The data exchange problem

3.1 - The different levels of standard

The exchange of data requires the definition of standards at three different levels: conceptual, logical, physical.

The physical level is the simplest to be agreed upon: almost everybody agrees to adopt a 9-track ASCII tape, or a 360Kb MS-DOS diskette. As a matter of fact, to agree at the physical level is exactly the same as two persons that agree that they will exchange information by means of paper sheets, or telephone.

The logical level is simply the definition of a key for decoding the information contained in the physical support. The exchange format, in this context, may vary from the very simple "card image format" to some sophisticated format, like the MARC format. But agreement at the logical level only means that everybody possesses a "decrypt key", exactly as two persons that exchange information on typewritten paper: will a chinese be able to read arabic or english?

Effective data exchange is possible only if a standard has been defined at the conceptual level: that is, when a model has been defined that allows everyone to share the knowledge of the world of interest.
Agreement at the conceptual level permits to understand the semantics of the fields, agreement at the logical level only permits to distinguish one field from another.
It is easily seen that, once the conceptual data model has been defined and agreed, the process of defining an exchange format is straightforward, and may be accomplished overnight.

3.2 - The normalization of the language

Even if the definition of a data model may be seen as the most relevant step toward the definition of an exchange format, peculiarities of art history data must be taken into account. As it has been already pointed out, a characteristic of this kind of data is their fuzziness, as the same concept (the name of an object, place, artist) may me designated in different ways, depending on the cultural background of the scholar. Once again, the problem is much more complicated that in a conventional business environment, as the normalization of data involves a great cultural effort, in order to reduce to a common frame different ways of thinking, each one based on valid and well established cultural traditions. Anyway, the experience has proved that retrieval of stored data may be effective only if a controlled vocabulary has been defined for data fields. Even better is to arrange the concepts, making explicit the synonymy, preference and hierarchical relationships between them, that is, build a thesaurus. As the building of a thesaurus is a very long and costly issue, it may be regarded as a long term target, while the normalization of the vocabulary is a task which is imperative and cannot be deferred.

4 - The italian experience

4.1 - The "giacimenti culturali"

In 1986, the italian government, with the intent of ecouraging the employement of young people, funded a Lit. 600.000.000.000 initiative, whose principal aim was the application of new technologies in the field of cultural heritage management. The

initiative took the name of "giacimenti culturali", as it was assimilating the cultural heritage to other types of resources, like oil or coal, that could be exploited.
After a call for proposal, some 39 projects were approved and financed. Of these projects, 31 were concerned in some way with the cataloguing of works of art. No guidelines were imposed as far as the technological environment was concerned, every project had the right to chose the hardware and software environment. The only constraint was that the results of the projects should be available to the central administration.
It is generally estimated that in this kind of projects a 15-20% of the budget is invested into hardware and software, and the remaining goes into the education of the personnel involved and gathering of data. If we agree on this estimate, it is easily seen that the value of the data was at large the most relevant issue of the projects.
Under this respect, a driving role was played by the classification model defined by the ICCD (Istituto Centrale per il Catalogo e la Documentazione) which has been illustrated above. It acted as a standard at the conceptual level, and was included as a constraint in the contracts signed by the firms which were executing the projects.

4.2 - The exchange format

Keeping in mind the great variety of hardware and software the projects had selected, it was soon realized that the adoption of a sophisticated exchange format was unfeasible: in some cases, the projects were unable to produce files with variable length records!
Therefore, the decision was taken to define, as logical level standard, a very simple "card image" exchange format, that everybody would be able to adhere to.
Really, the exchange format was defined in a couple of hours!
A similar approach was followed for what thesauri and authority files were concerned. A very simple, card image, format, where the data were arranged into small, semantically defined fields, was defined.
The definition of the physical standard was even simpler: we were accepting sequential files, and the only limitation was that the magnetic supports could be readable by the machines we were using. So, we imposed that the support should be one of the following:

- 3.5" IBM compatible floppy (360Kb, or 720Kb, or 1.4 Mb)
- 5.25" IBM compatible floppy (preferably formatted at 360Kb)

- 9 tracks, EBCDIC or ASCII, tape, accordingly to the international standards (1600 bpi, 1/2")

Finally, the names of the files were standardized, in order to be able to identify the producer and the content of the files on the basis of their names.

At the beginning, only a few fields had a controlled vocabulary defined. Therefore, an exchange procedure was defined, in order to merge the dictionaries produced by the different projects. Some automated procedures were coded for this purpose.

All the projects were able to conform with these standards (even if in a few cases some slight differences may be detected), and the data produced by them are currently going to be massively processed in order to build a large data bank.
Processing of the data will enable to build "authority files" on several fields, and this will be another step towards the completion of the cataloguing schema.

5 - Conclusions

In passing from manual to automated treatment of the cataloguing data, a big effort is required to precisely define a cataloguing model, and the structure of the fields. A great help may come from the database design methodologies.
A clean data model leads, without any effort, to the definition of data exchange standards.
In a real case, where there was the potential risk of wasting a lot of resources, starting totally uncoordinated cataloguing projects, the availability of a cataloguing model resulted in an effective discipline of data.

6 - Acknowledgements

All the standardisation work, that lead to the definition of the classification model, would not have been possible without the encouragement by prof. O. Ferrari, at the time being the director of the Istituto Centrale per il Catalogo e la Documentazione

(Rome), who pursuived the cultural objective of building a "Catalog" and not simply making an inventory (or counting the objects)
A warm thank goes to S. Papaldo, who coordinated the activities of the ICCD in defining the standards and gave a great contribution to the definition of the reference model for art objects, and to M. Ruggeri, L. Cavagnaro, and others from ICCD who performed an in depth examination of the cataloguing data, testing the approach on several case studies.
Finally, I have to acknowledge my colleagues R. Bartoli, R. Gagliardi, R.D. Matteucci and G.A. Romano for their help in the definition of the general schema and support in several case studies.

7 - References

Aulisi90 Aulisi R., Ceccanti V., Signore O.: *Hermes: un ipertesto sugli stemmi nobiliari corredato da un thesaurus figurato*, Bollettino d' Informazioni, XI, n. 2 (1990) pp. 104-152, Scuola Normale Superiore - Pisa, ISSN: 0392-9957

Aulisi91 Aulisi R., Ceccanti V., Signore O.: *Hypertext for Hypertext: a Figured Thesaurus*, Database and Expert Systems Application, Proceedings of the International Conference in Berlin, Germany, 21-23 August 1991, (to be published)

Cavagnaro91 Cavagnaro L.,Matteucci D.R., Ramellini G.,Signore O.: *Strutturazione dei dati delle schede di catalogo: beni architettonici e ambientali*, Istituto Centrale per il Catalogo e la Documentazione (Roma) Istituto CNUCE (Pisa), (To be published)

Ceri83 Ceri, S. (ed.): *Methodology and tools for data base design*, North-Holland (1983)

Chen76 Chen P.P.: *The Entity-Relationship Model: Toward a Unified View of Data*, ACM TODS, Vol. 1, N. 1, (1976), pp. 9-36

D'Amadio89 D'Amadio M., Simeoni P.E.: *Strutturazione dei dati delle schede di Catalogo.Oggetti di interesse demo-antropologico*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Museo Nazionale delle Arti e Tradizioni Popolari (Roma), Tipografia Città Nuova della P.A.M.O.M., Roma (1985)

Maffei90 Maffei S., Tarchi R., Signore O.: *Turms: una guida ipertestuale dei beni culturali dell' area grossetana*, Bollettino d' Informazioni, XI, n. 2 (1990), Scuola Normale Superiore - Pisa, ISSN: 0392-9957

Massari88 Massari S., Prosperi Valenti Rodinò S., Papaldo S., Signore O.: *Strutturazione dei dati delle schede di catalogo - Beni mobili storico-artistici: Stampe*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Istituto Nazionale per la grafica (Roma) - Istituto CNUCE (Pisa), 1988 Scuola Tipografica S. Pio X (Distributed by Istituto Centrale per il Catalogo e la Documentazione)

Montevecchi89 Montevecchi B., Vasco Rocca S.: *Metodologie di classificazione: Suppellettile ecclesiastica*, Centro Di (Firenze,1989), pp. 133 ISBN 88-7038-164-1

Papaldo85c Papaldo S., Ruggeri Giove M., Gagliardi R., Matteucci D.R., Romano G.A., Signore O.: *Strutturazione dei dati delle schede di Catalogo. Beni mobili archeologici e storico-artistici*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Istituto CNUCE (Pisa), Tipografia Città Nuova della P.A.M.O.M., Roma (1985)

Papaldo86 Papaldo S.,Ruggeri Giove M.,Gagliardi R.,Matteucci D.R., Romano G.A.,Signore O.: *Strutturazione dei dati delle schede di Catalogo: beni mobili*, Atti del Convegno sull' Automazione dei Dati del Catalogo dei Beni Culturali, Roma, 18-20 giugno 1985 (pag. 39-42) Ministero per i Beni Culturali e Ambientali - Istituto Centrale per il Catalogo e la Documentazione (Roma, 1986)

Papaldo88 Papaldo S., Ruggeri Giove M., Gagliardi R., Matteucci D.R., Romano G.A., Signore O.: *Proposta di strutturazione dei dati del catalogo: Beni mobili archeologici e storico-artistici. (Edizione riveduta e aggiornata)*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Istituto CNUCE (Pisa), Multigrafica Editrice, Roma (1988)

Papaldo89 Papaldo S., Signore O.: *Un approccio metodologico per la realizzazione di una banca dati storico-geografica (A methodological approach to producing a historical/geographical*

databank), Multigrafica Editrice, Roma (1989), pp. 573 ISBN 88-7597-105-6

Parise88 Parise Badoni F., Ruggeri M.: *Strutturazione dei dati delle schede di catalogo: Beni archeologici immobili e territoriali*, Istituto Centrale per il Catalogo e la Documentazione (Roma) - Istituto CNUCE (Pisa), 1988 Scuola Tipografica S. Pio X (Distributed by Istituto Centrale per il Catalogo e la Documentazione)

Signore84 Signore O.: *Data integration in Art History Information processing,* Conference Automatic Processing of art History Data and Documents, Pisa, Scuola Normale Superiore, September 24-27, 1984, Proceedings, pp. 312-319. Edited by L. Corti

Signore85 Signore O.: *Architettura di sistemi per la gestione di dati catalografici*, Atti del Convegno sull' Automazione dei Dati del Catalogo dei Beni Culturali, Roma, 18-20 giugno 1985 (pag. 51-58), Ministero per i Beni Culturali e Ambientali - Istituto Centrale per il Catalogo e la Documentazione (Roma, 1986)

Signore86 Signore O., Pardini S., Trasacco A.: *L' automazione del Catalogo dei Beni Culturali: realizzazione della base di dati per le schede OA e RA*, Internal report CNUCE C86-10

Signore89 Signore O., Bartoli R.: *Managing Art History Fuzzy dates: An Application in Historico-Geographical Authority*, Historical Social Research, Vol. 14, n. 3, (1989) pp.98-104 (Special Issue - Computer Applications in the Historical Sciences: Selected Contributions to the Cologne Computer Conference 1988)

Signore90a Signore O., Bartoli R.: *A case study for historico-geographical authority*, in Terminology for Museums, Proceedings of an International Conference held in Cambridge, England, 21-24 September 1988, Maney and Son Limited (1990), ISBN 0-905963-62-8,pp.150-158

Signore90b Signore O., Bartoli R.: *Implementation of a historical/geographical database with support of imprecise dates*, Database and Expert Systems Application, Proceedings of the International Conference in Vienna, Austria, 1990, (Tjoa A.M., Wagner R., Eds.) Springer Verlag Wien New York, ISBN 3-211-82234-8, pp. 271-274

Signore91 Signore O., Ceccanti V.: *An Hypertext on Heraldry*, Proceedings of Montpellier Computer Conference, 5e Congrès International de "Association for History & Computing", Montpellier, 4-7 Settembre 1990 (to be published)