James M.Bower
Getty Art History Information Program, Santa Monica, California

Vocabulary Control and the Virtual Database

Bower, J.M.: **Vocabulary control and the virtual databases.** Knowl.Org. 20(1993)No.1, p.4-7, 21 refs.
Efforts to build "virtual museums" have focussed predominantly on solving problems of rapidity changing interface technology. Insufficient effort has been spent on planning for the "virtual databases" on which these multimedia environments depend, particularly from the viewpoint of scholarly research. The Getty Art History Information Program has developed vocabularies that encourage consistency among scholarly documentation projects — regardless of their technical implementations — through terminology standards. Two vocabularies intended for control of terminology at the point of data capture are described, and scenarios are proposed for their further utility in navigating the complex databases that underlie the virtual museum.
(Author)

1.Introduction

In the 1989 film *Field of Dreams*, a Midwestern farmer facing foreclosure is driven by a disembodied voice to build a baseball diamond in the middle of an Iowa cornfield. The voice urgers, "If you build it, they will come". The farmer's persistence in realizing this apparent folly is rewarded when a ghost team of champion baseball players materializes to play on the new field. In the film's closing shot, as dusk settles, the camera pans back to reveal an endless stream of cars, full of eager baseball fans willing to pay the price of admission to see this fantasy team play, wending its way to the cornfield.

Similarly many single-minded museum administrators have invested scarce resources to build dazzling interactive multimedia systems called "virtual museums", imaginary environments in which simulated objects (analogous to the farmer's ghosts), displaced from the constraints of real time and space, can be selected, observed, and manipulated by the museum "visitor". Like the farmer in the movie, the administrators hope that, once built, these systems will attract a large and eager public to their institutions to be stimulated, educated, and entertained.

A virtual museum relies for its effects on an underlying "virtual database" in which is stored the information necessary to simulate the museum's objects, and often their display environment. The "visitor" accesses the database through computer programs in order to select and manipulate the objects, which are simulated through a digitally encoded mix of text, recorded sound, and images (still photographs, computer animation, film, video)[1].

The theme of the virtual museum has been well represented in the recent literature (7), (12),(14), 18). Most authors deal primarily with the technical issues involved in integrating the various media used to simulate the objects, and with solving the problem of how to present these multimedia packages to the general public without completely overwhelming them. By contrast, few writers have tackled the issues of content and structure of the virtual databases on which these systems depend, and fewer still have addressed these issues from the perspective of the scholarly researcher (16). The present article attempts to partially redress this imbalance.

For the purposes of this paper, the concept of the virtual database extends beyond any single multimedia project with a small domain of objects to encompass the full repertoire of polyglot digitized descriptions and images of objects being compiled in museums, libraries, and archives scattered temporally and geographically throughout the world, capable of being interchanged freely among institutions and individuals, using computer networks. Such a virtual database would be the art-historical equivalent of the virtual libraries created by the Research Libraries Information Network (RLIN) and the On-Line Computer Library Center (OCLC)[2].

A comparable database of objects cumulated from the holdings of many separate museums would be of intense interest to scholars (5, p.52), and in fact, such databases are already being compiled through national inventory programs in Canada, France, and Italy, among others. More recent initiatives such as the European Museums Network are bringing us even closer to the virtual database in this expanded sense (19). The prospect of adding objects from the thousands of museums worldwide that are now automating their collections is exhilarating, but when envisioned on such an international scale, the problems of the virtual database are magnified equally with the opportunities[3].

2. Weaknesses of the Virtual Database

In 1987 Gary Schwartz recognized four types of weakness common in databases of art objects (16, p.58). "Conceptual weakness" results from inconsistency between the purposes for which a database is created and the major elements that comprise the database (e.g., a database of photographs of artworks which omits any description of the photographs themselves). "Contentual wea-

kness" occurs when the quality of information in the database is compromised, as when commercial interests determine the data recorded in auction databases. Schwartz describes "resolutional weakness" as shifting focus in the degree to which data are analyzed. Finally, "political weakness" is seen as the result of conflicting agendas made manifest in the merging of data developed separately by collaborating institutions.

It is easy to imagine that the problems afflicting art databases individually increase by a level of magnitude when databases are conceptually linked and physically merged into something approaching the virtual database. Political weakness, ironically, may be the most easily overcome; only the institutions and scholars most willing to open their data for examination in the public sphere will contribute initially to the virtual database. Over time, as museums become more comfortable with distinctions between data suitable for collections management and data for scholarly research (in effect, private versus public consumption), they may be motivated to contribute to the virtual database in a *quid pro quo* for accessing its growing resources.

Conceptual weakness may be exacerbated in the virtual database by the juxtaposition of datasets created for similar purposes, but with different methods. A researcher may have difficulty using data merged from the catalogs of two photo archives if one treats photogoraphs strictly as surrogates for the artworks depicted, while the other catalogs them as primary documents in the historiography of art documentation. Similar difficulty emerges from resolutional weakness across the contents of the virtual database; a researcher may be frustrated when a query for information on Italian cassoni reveals a scholar's richly detailed catalog of cassoni in Tuscan museums, but only the most summary data on cassoni from other collections.

Contentual weakness is to the scholar the most dangerous and, in the contenxt of the virtual database, the most insidious weakness[4]. Judgments of quality and veracity must be made up front by the database sponsors, before datasets are subsumed into the virtual collection. Criteria for evaluating potential additions to the databank should be established and disseminated, to promote realistic expectations of what the database will be. One such criterion might be the extent to which a candidate dataset embodies standards — whether for data structure, data values, or syntax — that have gained consensus among the sponsors and their intended audiences.

3. Ambiguity, Vocabulary Control, and Vocabulary Coordination

Another weakness in the virtual database is the ambiguity that occurs naturally by the accretion of vast amounts of object information from different sources. Data that are unambiguous within the context of their initial capture (e.g., a scholar's personal research database) may become ambiguous when juxtaposed with data in different languages, data from other disciplines where overlapping terms have not been rendered referentially unique, or data from the same discipline that use different but equivalent terms to express names and concepts (17, p.2).

An effective method of dealing with ambiguity in the virtual database is vocabulary control, which structures terminology so that the language of the object's cataloger is brought into coincidence with that of its researcher (13, p.8), (15). Svenonius distinguishes between different kinds of vocabulary control (defined according to the data elements being controlled — e.g., personal names, iconographic themes), and different degrees of vocabulary control (defined according to the range of terminological relationships built into the structure of a given vocabulary) and cites synonym control, hierarchical-term control, and related-term control as three mechanisms suited to art information (17, p.6-7).

Conceptually, vocabulary control is valid only within a database in which all material is consistently subject to the same set of controls. By systematizing terms at the point of data capture, the controlled vocabulary serves as a linguistic filter through which the cataloging and subsequent retrieval functions are reconciled. The virtual database, however, is composed of many datasets defined and controlled according to varying local criteria, recontextualized into a heterogeneous mix. In this situation, where the constituent datasets predate the virtual whole, it makes little sense to speak of vocabulary control. The task becomes one of integrating, after the fact, the multiple vocabularies used to control the data subsets — what the author calls vocabulary coordination.

To the optimist, the future is always longer than the past. While the number of art databases completed or already under construction is substantial, it still represents a relatively small percentage of the art objects extant in the world. As technological and financial barriers to the automation of museum collections are overcome, scholars and museums have an opportunity to coordinate their documentation projects through the application of shared vocabularies. While satisfying local needs, coordinated use of controlled vocabularies offers the added benefit that data merged from different sources will be inherently consistent at the terminological level.

It is to this goal that the Getty Art History Information Program (AHIP) has applied itself increasingly in recent years. Because inconsistent use of terminology has been seen to hamper both the retrieval and sharing of electronic information, AHIP has actively promoted consistency and compatibility in the creation ,of art-historical databases. AHIP has given priority to vocabulary projects — creation of standards, development of common resources, and technical assistance in vocabulary use — because these have shown the greatest potential to benefit the field.

4. AHIP Vocabulary Projects

Svenonius argues that vocabulary control in art databases will be most effective when applied to the data

elements most frequently used by art scholars in their research (17, p.7). A recent AHIP inquiry into the information-seeking behavior of humanities scholars has shown a strong preference for personal names, geographic names, chronological terms, and common subject terms as access points into automated databases (6, p.14-15). AHIP has addressed three of these areas in its vocabulary projects.

Subject terms for art and architecture (as distinct from iconographic terms) have been developed by the Art and Architecture Thesaurus, a project of Getty AHIP located in Williamstown, Massachusetts, that has been reported widely in the literature (and elsewhere in this journal)[5]. So pressing was the need for in-depth controlled vocabulary in this area that the Art and Architecture Thesaurus (AAT) was published in 1990 in partial form (2), with a supplement and electronic edition appearing in 1992 (3). The complete thesaurus containing nearly 70,000 terms will be released in print and electronic forms in 1994.

Names of artists and architects, and of geographic places, have been developed by the Vocabulary Coordination Group (VCG), an AHIP project established in 1987 to identify and coordinate controlled vocabulary resources for use in automated documentation systems within the J.Paul Getty Trust, and elsewhere in the art-historical and museum communities.

4.1 The Getty Union List of Artists Names

To foster consistent usage of artist names among Getty Trust databases, the VCG developed the Getty Union List of Artist Names (ULAN), a database of artists' and architects' names, biographical data, and bibliography produced from the merged authority files of nine Getty projects[6]. These projects span a broad spectrum of art documentation types, including abstracting and indexing services, research photograph archives, scholarly databases that document primary source materials, and object collections from archives and museums. As a result, the ULAN contains over 200.000 names representing almost 100,000 individuals. Nearly 60% of these artists are represented by one name and one project only; the remaining 40% overlap in two or more project databases, some with over 100 different forms of name[7]. Overlapping records for the same artist are linked to form "clusters", but the original structure of each contributed record is retained, showing the projects' choices of preferred and variant name forms. This pluralistic approach reveals patterns of usage among the names and biographic data that are helpful to projects in building their individual authority files.

To encourage use of the forms most appropriate for scholarly documentation, names in a cluster are algorithmically "ranked" by contributor, according to criteria such as preference for names in the vernacular, and the depth of bibliographic sources used in the verification process. The name that emerges will serve as the focus of the cluster in online displays, and will determine the alphabetic entry point for the cluster in the published ULAN. All other names linked to the cluster serve as access points to it, either through keyword retrieval online or as "see references" in the printed alphabet.

Although the database only links synonyms, the dense web of name variants makes this an extremely effective tool for control of artist names. ULAN names can be flagged to indicate specific standards by which names have been formulated (e.g., the RILA Verification Manual, or Anglo-American Cataloging Rules, 2nd Edition). Ambiguity among names is rare, skeletal biographic strings (typically including nationality, role, and lifedates) are included to help discern among artists[8]. The ULAN will be distributed in print and electronic forms early in 1994.

4.2 Thesaurus of Geographic Names

While the ULAN was developed solely from Getty-contributed data, VCG based its Thesaurus of Geographic Names (TGN) on machine-readable geographic data available commercially from cartographic publishers. The TGN contains hierarchies of geographic names and related data representing ca. 300,000 places, produced by merging records for United States and international populated places, features, and political units from Rand McNally Corporation's RANDATA database with records from the geographic authority files of three Getty projects[9]. Although 9,000 historic place names from Times Books Ltd's Atlas of World History have been incorporated into the file, the primary emphasis of the database is on contemporary geography.

A geographic place in the TGN database can be a physical or political entity, either current or historic. Places include topographical feature (e.g., lakes and mountains), and entities described by political boundaries (e.g., cities, counties, regions). Records for each place include the current name for the place, any historic names from our contributors' datasets, and at least one "place type" (a word or phrase used to characterize the place according to its current physical aspect or political autonomy). Records may also contain physical coordinates, dates, notes describing the place and its changing boundaries and characteristics over time, and bibliographic citations.

Data are organized differently in TGN than in ULAN. Place records are constructed by collapsing together all contributed data for a place. An authoritative form of the current name in the vernacular is identified, with additional forms (including historical equivalents) designated as variants. Place names and place types are additionally flagged according to their language (vernacular, English, or other).

In addition to controlling synonyms, TGN provides hierarchical-term control, including multi-parent relationships. Ambiguity among place names is common, but is mitigated through inclusion of place types and geographic coordinates in "thesaural" displays, and through display of hierarchies. The Thesaurus of Geographic Names will be released on compact disk in 1994.

5. Vocabulary Control and Vocabulary Coordination, again

AHIP's three vocabulary resources - the *Art and Architecture Thesaurus, Union List of Artist Names,* and *Thesaurus of Geographic Names* — have been developed in part to help institutions and individuals on the threshold of automating their collections, or preparing to build scholarly databases of art information, to control the data they capture. This mode of use will be supported by the release of each resource in electronic form with the Authority Reference Tool (ART), an AHIP-designed software program that allows rapid navigation through complex vocabulary data, allowing the user to "cut and paste" terms or names from the resource file into the user's local application. AHIP anticipatess that widespread use of its vocabulary tools in this form will result in *de facto* terminological consistency within and among the many scholarly databases now being planned and built.

At the sametime, AHIP recognizes that its vocabularies have tremendous potential as retrieval tools in the virtual database, where constituent datasets often use different — sometimes conflicting — vocabulary standards. Optimally, if the linking structures inherent in the AHIP vocabulary have also been captured into the dataset being queried (e.g., the ULAN Cluster ID, which uniquely identifies an artist), the AHIP terms can be used transparently (to the searcher) to retrieve synonyms and related terms, or to navigate across hierarchies to retrieve data at higher or lower levels of specificity. Even if the linking structures have not been captured in the query file, AHIP vocabularies acting as search "filters" could recognize potentially ambiguous data and prompt the scholar to refine his search accordingly.

The retrieval potential of AHIP'S vocabularies expands as one examines the technological scenarios being forecast beyond the year 2000. Weissman, for example, foresees fundamental changes in the way data are structured and stored in computers, and a radical shift away from tightly bound software and at applications toward generic software tools that can be mobilized in any combination around multimedia documents in an operating system database environment (20). In the virtual database, this would break down structural barriers among constituent datasets and increase the likelihood of successful retrieval using linked vocabularies. Search and retrieval paradigms once appropriate only to specific types of information will blend as applications become highly integrated (10), (21). One can imagine using AHIP tools to search in the virtual database and retrieve fixed-field data from a 50-year old national heritage file, relational data from a 19th-century inventory automated in the 1980's, and biographic information from an artist authority file created in 1996 using SGML (Standard Generalized Markup Language) tags. Retrieval will not be limited to text-based searches, either; iconic criteria will also be used (4, p.37).

6. Conclusion

The use of structured terminology for vocabulary control and coordination in art-historical databases at a local level will, over time, enable the successful integration of those datasets into a virtual database greater than the sum of its parts. Weaknesses to which art databases are now subject will be mitigated, as vocabulary structures ensure that commonality can be exploited through uniformity, while necessary differences (e.g., in language, or levels of specificity) are accommodated through appropriate control mechanisms. Vocabulary control and coordination, when linked with increasingly sophisticated computer capabilities, will allow the researcher to navigate more and more effectively in the database of the virtual museum.

Notes

1 Buckland (9) aptly points out that, although the term "multimedia" is used as an umbrella term for the various types of original documents from which the virtual object is composed, the data thus rendered are actually stored in a monomedium.

2 Strictly speaking, the RLIN and OCLC databases are not "virtual libraries" because they do not simulate the materials represented by the bibliographic records. The analogy extends only to the text-based descriptions of objects and books maintained in the respective systems.

3 Glushko (11) offers a useful analysis of the technical reasons why such projects fall under the burden of "scaling up".

4 Arnold (12, p.200) contends that the more "satisfying" multimedia applications seem to viewers, the more discouraged those viewers will be from cross-checking the data they receive.

5 A full bibliography of AHIP publications and articles is available; contact the author at the address listed.

6 The nine participating projects are the *Avery Index of Architectural Periodicals;* the *Bibliography of the History of Art (BHA);* the *Census of Antique Art and Architecture Known to the Renaissance;* the *Foundation for Documents of Architecture;* the *Photo Study Collection of the Getty Center for the History of Art and the Humanities;* the *J.Paul Getty Museum; the Provenance Index; the Vocabulary Coordination Group;* and the *Witt Computer Index.*

7 The ULAN captures equally the significant variations in name (e.g., "Giovanni Antonio Bazzi" and "Il Sodoma"), and minute variations based on verbatim transcriptions of artist names from art-historical documents such as collection inventories and auction records.

8 Data from the ULAN were used to test a set of computerized name-matching algorithms developed by AHIP as an outgrowth of its Museum Prototype Project, itself an early attempt to create a virtual database: cf. Borgman/Siegfried (8).

9 The projects are the *Bibliography of the History of Art (BHA);* the *Foundation for Documents of Architecture;* and the *Photo Study Collection of the Getty Center for the History of Art and the Humanities.*

References

(1) Arnold,l S.E.: The large data construct: a new frontier in database design. Microcomputers for Inform.Management 7(1990)No.3, p.185-204

(2) Art and Architecture Thesaurus. New York: Oxford Univ.Press 1990. (Supplement 1.New York: Oxford Univ.Press 1992)

Continuation on page 34