LIBRARY HI TECH

A Special Issue on

Libraries and the Humanities in the 1990s

edited by

Fred Batt and
Charles Martell

Consecutive Issue 33
Volume 9 Number 1 1991

Library Hi Tech 1991 1

Retrieving Images Verbally:

No More Key Words and Other Heresies

Jocelyn Penny Small

The Lexicon Iconographicum Mythologiae Classicae is an international project of nearly forty countries to produce a pictorial dictionary of classical mythology. The author describes the system she has developed to address the problem of retrieving images, the way it differs from most cataloging systems, and how it can be applied to other types of information (e.g., text) to achieve the All-in-one, Nothing-left-out, Everything-at-your-fingertips, Computerized Humanities Emporium. Such an emporium can be developed today; the most important steps in achieving this leap into the future have already been taken.

Small is Director of the U.S. Center, Lexicon Iconographicum Mythologiae Classicae and a member of the art history department and library at Rutgers University, New Brunswick, New Jersey.

Consider a color slide or photograph of an Etruscan wall painting of Achilles hiding in the bushes behind a fountain to ambush the young Trojan prince, Troilos, riding in from the right.[1] In analog reproductions like these, the colors and detail can be well preserved. Either could be automated on a videodisk of some sort or digitized and then manipulated to change colors or even the figures themselves. A number of projects, in fact, are devoted to doing just those things. I, however, am not going to discuss any of them. It is not that I do not want an automated slide system with fifty thousand images on a single platter. It is not that I am not interested in distinguishing between a right arm raised to address the troops in an *adlocutio* and one waving "hi there." It is that pictures need words. Words for artists. Words for style. Words for dates. Words for subjects. Words for things. In other words, many, many words. I do not want even to contemplate the idea of looking through all 7,500 photographs in my comparatively small archive much less fifty thousand images on a disk to find the few I need.

I am going to describe the system I have developed to address the problem of retrieving images, the way it differs from most cataloging systems, and how it can be applied to other types of information to achieve the All-in-one, Nothing-left-out, Everything-at-your-fingertips, Computerized Humanities Emporium.

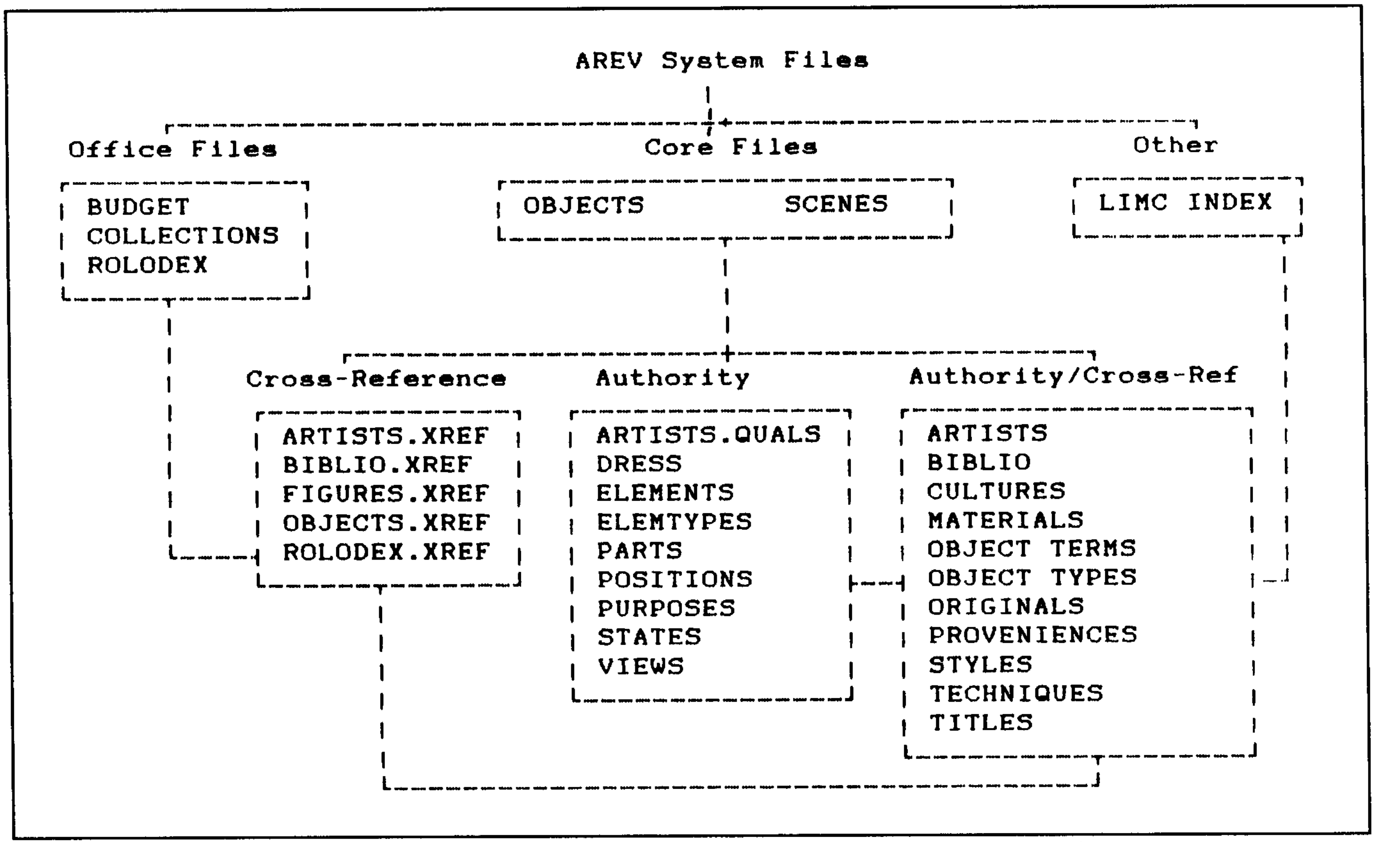


Figure 1: Flow-Chart for the Computer-Index of Classical Iconography.

The Lexicon Iconographicum Mythologiae Classicae is an international project of nearly forty countries to produce a pictorial dictionary of classical mythology.[2] As the Director of the U.S. Center, I was responsible somehow or other for coming up with a working computerization of our material.[3] I could hire outsiders—though not many, nor for long, my budget being what it was and is. As a result, over a period of time it became absolutely clear that I was the one who was going to have to do most of the work. My near total ignorance of computers probably was my greatest asset. I did not know that you could not do certain things, or even, if you could do them, you should not. So, over a period of five years my system has evolved and continues to evolve.

Consider the present incarnation (figure 1). I have given up trying to make a completely accurate flow chart with straight lines and right angles. A three-dimensional display might work, but the truth is that only a cobweb captures the tangle of connections. Logically and visually in the center are the two Core Files—Objects and Scenes. Objects contains all the information that pertains to the object as a whole, while Scenes records the data about the different representations that appear on a single object. In fact, the system actually allows for multiple interpretations of the same scene both to reflect the state of scholarship and to make the job of cataloging easier. My enormous staff—one other person and me—simply do not have the time to decide which scholar is right, and, even if we did, we would be no more believed than any other given scholar. Thus **Principle Number One** is Aristotelian: ***"Do not make your datum more accurate than it is."*** This principle may be rephrased as, ***"Preserve the Mess."***[4]

Preservation of ambiguity, however, does not mean a lack of either organization or controls. As to be expected in any structured database, information is divided among a number of boxes according to the **Second Principle**, which states that ***"Information should be reduced to its smallest unit or least common denominator."*** This method really works only in a relational database that permits broader and narrower definitions of terms in separate, but coordinated files. For example, the specific site where an object was found is noted in the Objects file (figure 2) and a second file, Proveniences (figure 3), classifies the site according to its modern location. Thus "Gela" is in Sicily, which is in Italy. The result is that not only can the system give all objects from Sicily or Italy, but also that information need be recorded only once, and, if Sicily ever secedes from Italy, only one record need be changed. Fields like Provenience lend themselves to controlled vocabulary.

FIGURE 2:

OBJECTS FILE: US.NO 5303

US NUMBER:	5303	
CITY:	New York	
COLLECTION:	Metropolitan Museum of Art	
INVENTORY NUMBER:	45.11.1	
EX COLLECTION:		
DISCOVERY DATE:		
PROVENIENCE:	Gela?	
FINDSPOT:		
OBJECT:	Vase	
TYPE:	Pelike	
SUB-TYPE:		
PURPOSE:		
CONDITION:	Repaired from fragments; incomplete.	
DIMENSIONS:		
MATERIAL:	Clay	TYPE:
CULTURE:	Greek	
REGION:	Attica	
SITE:		
TECHNIQUE:	Red-Figure	
STYLE:	Attic	
ARTIST:	Polygnotos	
BASIS:	Beazley	
ORIGINAL:		
DATE:	450 BC - 440 BC	
BASIS:		
INSCRIPTIONS:		
DECORATION:		
REMARKS:	Seen by Davies.	
BIBLIOGRAPHY:	ARV/2	1032 No. 55
	Para	442
	Addenda	155
	Henle, Myths	91 fig. 43
DATE ISSUED:	06-06-90	

FIGURE 3:

PROVENIENCES FILE: Gela

PROV	Gela	OBJECTS:	18
SYNONYMS			
GLOBAL			
NATION	Italy		
REGION	Sicily		
SITE	Gela		
USAGE			
REMARKS			
REFERENCES	[1]	PAGES	
PECS		346-347	
COMPLETE	YES		

But, and this can be an enormous "but," as the **Third Principle** states, ***"Never ever will any person or project, no matter how knowledgeable and experienced, be able to put together a list of words that will not need to be changed...continually."*** Two examples from my project illustrate the hazards. I spent a year experimenting with different designs for the database, and I might add I still experiment with different designs, but more on that later. During that first year I had developed several files to control vocabulary, among which that for the names of objects seemed comparatively simple, especially on the level of overall identification of the kinds of thing being cataloged. Was it sculpture, a vase, jewelry, and so on? At the same time I had made that field one that had to be filled in; if it was not, the data enterer could not proceed to any other part of the record. Imagine then my alarm when I learned that I, wonderful scholar that I am, had not included all possibilities and that staff with no choice and a true workaholic approach to complete everything now chose the second most common term, "Sculpture," as a default. Now all fields with a controlled vocabulary have "Problem" as a possible entry, and staff is instructed to describe the problem in the "Remarks" field. The fact that Advanced Revelation, the program I use, has all variable length fields with variable length storage up to 64 thousand bytes per record is a real boon in such situations.[5]

My second example is worse. I had had discussions with a university press about doing a dictionary of ancient things. While table and chair might be easy to identify, an iunx—a love charm of Aphrodite—might not be. Moreover, the degeneration of the grape bunch on fourth century South Italian vases makes the last in the series difficult to identify without knowledge of its predecessors. One of the first questions the publisher raised was how many terms were involved. From a variety of sources I compiled a list of nearly 2,000 terms. As it worked out, the project was blessedly dropped by both of us simultaneously, but I still had my file of terms, which is used currently to classify things and figures depicted in our scenes. For instance, a chariot is cataloged as a vehicle in order that all representations of transportation can be found. Last year, three years after I had made the original list, I decided to compare our actual usage in the Scenes File with that in the Things or Elements file, as I call it. The results were curious and slightly appalling. Only thirty-five percent of our actual usage overlapped with my independent compilation. The remaining sixty-five percent discrepancy cannot all be plurals. Separately and unbeknownst to me, Donald Walker and Robert Amsler at Bellcore had done a vocabulary study that compared words defined in *Webster's Seventh Collegiate Dictionary* with those actually used in the *New York Times*. Not terribly surprising was the result that two-thirds of the words in the dictionary did not appear in the *Times*, but rather startling was the fact that two-thirds of the words in the *Times* did not appear in the dictionary and clearly they were not all proper names.[6] The Bellcore figures and mine match all too well and lead to the same conclusion, embodied in **Principle Four**: ***"No controlled vocabulary should be produced in the absence of actual usage."***

I do not mean that one has to wait until the results from usage are in. One does have to start sometime somewhere. Nonetheless, one has to realize that any list of terms compiled ahead of time will be subject to change in the form of substitution of words and additions. Three further issues are associated with controlled vocabularies: alternative terms, usefulness, and accuracy of retrieval. Even if a database such as mine uses a controlled vocabulary, the same access must be given for synonyms. I can give a simple example from my own field. All the vases I have to deal with are made of fired clay, which is how terracotta is defined in the dictionary. Yet the scholars who study vases say that they are made of clay, while those who work on statuettes, which are made of exactly the same fired clay, generally say they are made of terracotta and even may refer to those objects as "terracottas." Which term do you use? Do you follow tradition and use clay for vases and terracotta for statuettes? You will still have the lone scholar in the field and certainly those outside of the discipline using the so-called wrong term. I decided not to follow custom, and opted for the shorter term that was easier to type. I still cannot, however, ignore the other usage. In fact, as part of an international project my language problems are more than quadrupled. The publication uses four languages: English, French, German, and Italian. I therefore treat foreign languages as synonyms. For the moment I shall pass over the problems of lemmatization in German, classical Greek, and Latin.

Usefulness is a characteristic rarely considered by scholars when designing their own databases, but always when searching someone else's. In March 1990, I participated in a conference on a joint project of Rutgers and Princeton Universities to establish a national center for machine-readable text. The attendees were evenly balanced among computer types, scholars, and librarians with a couple of publishers. It took less than two hours for the question "What is a text?" to arise. Despite Marshall McLuhan's contention that "the medium is the message" the scholars established to their total satisfaction that ontologically a machine-readable text exists totally divorced from its medium, be it tape, disk, paper, what-have-you, because it could be and probably would be transferred from medium to medium. Moreover, the details of a particular encoding are so specific to the individual incarnation that the compiler needs to be contacted in any case. Clearly a field for medium was worthless and should not be included in any catalog of machine-readable texts.

The librarians from their experiences in the trenches, however, disagreed. Imagine a scholar who wants some obscure text like one of the lesser known novels of Bulwer-Lytton or Frances Trollope. He or she goes to the reference librarian and finds out *mirabile dictu* that some other equally benighted soul had indeed encoded the text. The scholar is ecstatic, the librarian who has to do the scut work much less so. He or she has to get in touch with the compiler with the hope that the person has not moved, much less died in the meantime. With good luck in a true Panglossian world a response would come back within a reasonable amount of time, a couple of weeks say, that indeed the text is in machine-readable form and available to any scholar with no strings attached. That the text consists of keypunched cards for a machine that exists at one site serviced only by Greyhound, which is currently on strike, should not be considered a deterrent. When this example is multiplied by a large number of queries, and they will inevitably increase, the librarians face an administrative horror. While medium may not matter philosophically, it sure does practically. Thus the **Fifth Principle** of design concerns reasonableness: ***"The amount of effort to record a particular piece of information must be weighed against the amount of usefulness returned from that piece of information."*** This principle has two corollaries, of which the **First Corollary** is: ***"Utility always takes precedence over philosophy."***

FIGURE 4:

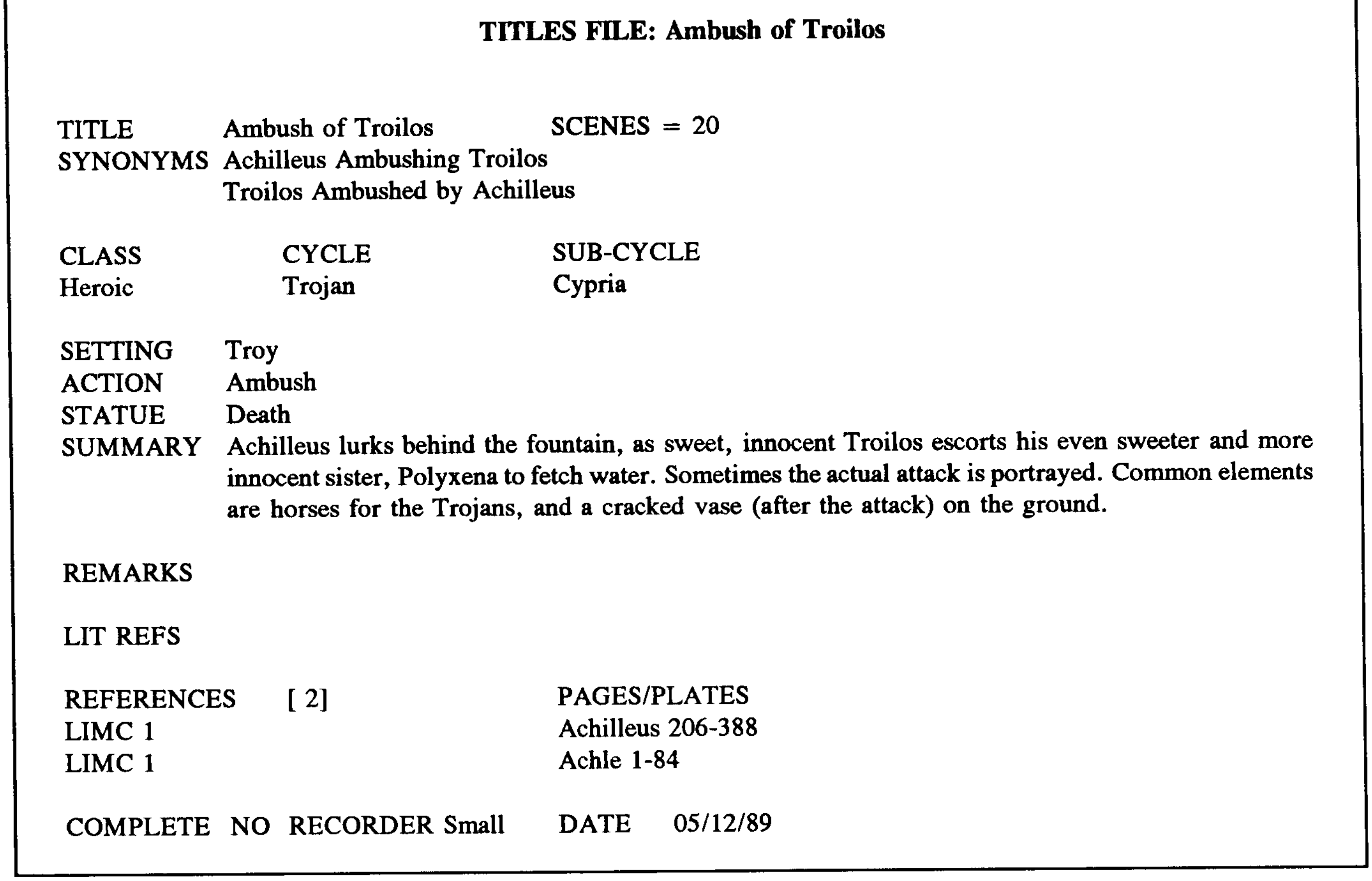
SCENES FILE: SCENE.NO 4442-A1

US NUMBER:	4442
SCENE NUMBER:	4442-A1
CITY:	New York
COLLECTION:	Metropolitan Museum of Art
INVENTORY NO.:	01.8.6 (GR 521)
PHOTOGRAPH(S):	MusPh 140548
VIEW: Side A	
PART:	
POSITION:	
TITLE:	Ambush of Troilos
DESCRIPTION:	Achilleus (greaves, baldric with scabbard, spear in right, Boeotian shield, high-crested helmet) pursues Troilos (white chitoniskos) who rides one horse, with another to his side and Polyxene (draped) who flees on foot from a fountain house (Doric portico) shown at left. Polyxene's oinochoe lies abandoned on the ground between Achilleus' legs; a hare bolts beneath the feet of the horses and a bird flies above the horses' backs toward Achilleus.
OTHER TITLES:	
REMARKS:	
DATE ISSUED:	06-06-90

The Fifth Principle works in both directions. We have just considered adding useful information. Its **Second Corollary** states that ***"Not all information useful to a scholar is worth recording,"*** because the costs of recording it, especially in human time, can be too great when weighed against its benefits. Without any doubt, pose is an important element in any pictorial rendering, but equally without any doubt it is a nightmare to record in a meaningful manner and a quagmire of unending detail. Take a simple example of the ambush of Troilos on an Attic cup from ca. 575 B.C. in the Metropolitan Museum of Art, New York.[7] (See figure 4.) The moment chosen is later than that in the Tomb of Bulls, because Achilles has left his hiding place to pursue Troilos and, in this representation, his sister, Polyxena. Now let me describe Achilles. His head is tilted forward, as is his upper torso. He holds a shield in his left hand, slightly lowered and to the side, and a spear, somewhat behind him by the middle of its shaft in his right hand. Merely saying he is pursuing his quarry is insufficient. Is he moving swiftly, slowly, deliberately, intently? Do I say he is striding or is he running? Maybe a detailed description can get around the problem. He moves to the right with a lengthy stride that puts his left foot forward and flat on the ground, while only the ball and toes of his rear right foot rest on the ground. Where do you stop? Even this simple scene with three human figures and two animals will take a fair amount of time to catalog. Imagine doing a Roman sarcophagus with forty figures. While pose is obviously useful to the scholar, it is far too subject to individual interpretation which, in turn, implies that a standardized vocabulary will be difficult to develop and that without a consensus on terms not much will be retrieved anyway. Over and above that, optimist that I am, I think that pose will be best handled by visual pattern matching.

Most art historical databases stop at the level of the title of a scene, which presents two problems for the ancient art historian. First, classical representations do not automatically come with agreed upon titles like "Rembrandt Contemplating the Bust of Aristotle." Instead we are faced with a number of scenes, which from the specificity of figures and action must be telling a story, but unfortunately we do not know what it is. In some cases, we can identify a single figure, like Herakles, and even the action, a fight, but not which particular one. Other times a number of copies of the

FIGURE 5:



TITLES FILE: Ambush of Troilos

TITLE	Ambush of Troilos	SCENES = 20	
SYNONYMS	Achilleus Ambushing Troilos Troilos Ambushed by Achilleus		

CLASS	CYCLE	SUB-CYCLE
Heroic	Trojan	Cypria

SETTING Troy
ACTION Ambush
STATUE Death
SUMMARY Achilleus lurks behind the fountain, as sweet, innocent Troilos escorts his even sweeter and more innocent sister, Polyxena to fetch water. Sometimes the actual attack is portrayed. Common elements are horses for the Trojans, and a cracked vase (after the attack) on the ground.

REMARKS

LIT REFS

REFERENCES [2]	PAGES/PLATES
LIMC 1	Achilleus 206-388
LIMC 1	Achle 1-84

COMPLETE NO RECORDER Small DATE 05/12/89

same type, often Roman copies of Greek statues, were made that are distinguishable primarily by their location and inventory number. I pass over the problem of fragmentarily preserved objects. Still, a sufficient number of straightforward cases exists to make the category of title useful. While we catalog every single figure and element in a scene, restricting the information to just them and not a made-up title will not provide sufficient information for those figures, like the gods, who have engaged in a number of events, especially across cycles. For example, Athena appears as an on-looker in scenes with her pet heroes like Herakles, and as an active participant in others like the fight with the giants. At the same time scholars are interested in certain groups of information, such as all representations from the Trojan Cycle or the *Iliad.* Thus individual titles provide points of entry to the Titles file where, for instance, the "Ambush of Troilos" is placed not only in the Trojan Cycle but in its sub-cycle, the Cypria (figure 5).

As I have already mentioned, each scene is broken into its individual elements, which can be figures, flora, fauna, architecture—in short, anything within the scene that is not part of the ornamental decoration, such as the bands that divide scenes on Attic vases (figure 6).

Principle Six states that ***"It is easier to catalog whole groups of entities than to remember which ones are the right ones."*** In other words, do not do some animals. Either do all of them or none of them. In this respect it should be noted that the only way you can tell one specific rendering of a particular story from another is by the figures and elements present. Thus in the Ambush of Troilos what counts are not just the different moments, but whether, for instance, Polyxena appears.

CLASSIFICATION SYSTEMS FOR ART HISTORY

It is perhaps appropriate at this time to briefly discuss the two major classification systems for art history: ICONCLASS and the Art and Architecture Thesaurus.

ICONCLASS

ICONCLASS is based on what I call the Roget Thesaurus Approach to the World. Take everything, absolutely everything, you can think of. Then divide

FIGURE 6:

SCENES FILE: SCENE.NO 4442-A1 - Grid

SCENES............	ELEMENTS...............	TYPE...........................	DRESS/ATTRIBUTE....................
4442-A1	Fountain		
	Water		
	Column	Doric	
	Achilleus	Warrior	Helmet, Corinthian; Baldric; Greaves; Sheath; Spear; Shield, Boeotian
	Oinochoe		
	Bird		
	Troilos	Clad	Chiton, Short; Reins
	Horse		
	Horse		
	Hare		
	Polyxene	Clad	Peplos

it into one hierarchical and interlocking set of broad categories with examples. Don't stop there. Add codes so that additions are difficult to fit in. Codes, while short, are also, when enough are accumulated—and there are nine volumes to ICONCLASS—impossible to remember. They are also unbreakable. My "Ambush of Troilos" may not be the best of titles. I can imagine a compelling case being made for "Achilleus Ambushing Troilos" or "Troilos Ambushed by Achilleus," depending on your predilections. Nonetheless, the recording of the actual words makes parts of the title searchable. The codes do enable ICONCLASS to give more explicit explanations, which I find simultaneously too detailed and too ambiguous—a neat feat. In our example, code "94 E 21" is defined as "Achilles watches Troilus, Priam's son, at the horse-trough; sometimes Polyxena present."[8] I pass over the fact that the setting for the action is a fountain, not a horse trough, where Polyxena has gone to fetch water with her brother along as protector. The maddening problem with the overly specific ICONCLASS description is that it inconsistently tells who Troilos is but not Polyxena, and then omits the information the iconographer really wants to know. Is Polyxena present or not? Furthermore, their arcane way of creating new codes inevitably leads to different projects creating duplicate numbers with different subjects. **Principle Seven** is ***"No codes...ever."***

Art and Architecture Thesaurus

The Art and Architecture Thesaurus (or AAT) is also a hierarchical listing of words with definitions and notes on usage, both of which, unfortunately, are not searchable in the current system, a contradiction of **Principle Eight**: ***"Everything, and especially free text, should be fully and Boolean searchable."*** It is far more flexible than *ICONCLASS*, not the least because it uses words rather than codes. It also allows for synonyms, and broader and narrower terms. At the same time usage at actual projects is recorded. Even more important it is available electronically. It too has skimped on classical art, but more significantly, like all the other projects I know of, it only records hierarchical information. Some kinds of information are simply informational. For example, the Ambush of Troilos takes place at Troy, the Birth of Erichthonios at Athens. Myron is Greek and is a sculptor. None of these bits of information necessarily follow from the other, but they are all of interest. So I call my authority files "classification files" because they perform the duties not just of thesauri, but they also contain important non-hierarchical information.

As you have probably gathered by now, almost no guides exist for my discipline. I am slowly, sometimes it seems extremely slowly, choosing terms and retroactively imposing consistency. Let us return to the New York cup with Troilos. Ten elements have been recorded in three related fields (figure 6). The database program I use blessedly does not require the designer to calculate how many elements will exist in the most complex scene or to set up a separate file with each figure having a separate record. Instead each field can contain many values which, in turn, can be associated with other values. **Principle Nine** states that ***"You***

should always buy a program more powerful than you anticipate you will need, since inevitably you will not have anticipated everything you will need, for which compare Principle Three. " Thus we have "Elements" for the various items that occur in scenes. These elements are further defined as to their "Type." For this Ambush of Troilos the Fountain House is Doric, Achilleus is a warrior, and two figures are clad. Two comments: we follow the orthography of the publication of the *LIMC* for the names of figures (hence Achilleus and not Achilles); and, as every art historian knows, many ancient art historians are obsessed with clothes or drapery, as they call it. The Fountain House is further defined in the Elements file as Architecture. Similarly and in the same Elements file all figures, like Troilos here, are given with their parents, mates, and children—all repeating or multi-valued fields to allow for multiple liaisons and offspring (figure 7). In addition Troilos is again classified as Heroic, Trojan, and in the Cypria, as was the title of this scene in the Titles file. **Principle Ten** states that ***"The official rule for relational databases that no piece of information should be repeated exists only to be broken. If using the database is easier with information repeated, then do so. Storage and elegance are the least of the problems. "*** The **Corollary to Principle Ten** says ***"Multiple and overlapping points of access allow for better retrieval. "***

The third of the three related fields in the Grid, as I call it, contains both dress and attributes. The combination of the two fields evolved. In the beginning there were five fields in the Grid, but experience quickly got rid of the fifth and less quickly led to the demise of the fourth. It is all well and good to separate dress from attributes, but what do you do with the lion skin of Herakles which is both. What do you do with that same lion skin when he is not wearing it, but has it draped over his arm? or hanging from a tree? The last brings up the knottiest problem. In most cases the location of an item reflects its ownership. Herakles wears, holds, or sits on the lion skin. Sometimes, however, the attribute is either some place separate from its owner or worse yet in the case of Herakles and Apollo fighting over Apollo's tripod either in Herakles' possession or jointly held by both figures. Since determining intellectual ownership was often impossible for less well-known figures, but the physical location was obvious, I decided that we would record that information only. We proceed arbitrarily from left to right and top to bottom both to ease the cataloging and to provide a minimum idea of the collocation of figures and things.

In the dress/attribute field not only are multiple items possible for any element, but they in turn can be further modified. Troilos wears a short chiton and Achilles holds a Boeotian shield. Again, however, there is a limit to the amount of detail recorded. We note the devices decorating the outside of shields, which are generally single, simple ornaments like a bull or a triskeles. When devices are complex, like an Amazonomachy, we merely record the overall descriptive term in accordance with the second corollary to Principle Five. Every item of information is bound to be interesting to somebody, and the more picayune that item the more vociferous that somebody will be. Nonetheless, it is necessary to draw the line somewhere.

FIGURE 7:

ELEMENTS FILE: Record for a Figure

TERM:	Troilos
SYNONYMS:	Troilus
GREEK:	Trwi/los
ETRUSCAN:	Truile
LATIN:	Troilus
FRENCH:	Troilos
GERMAN:	Troilos
ITALIAN:	Troilo
MAJOR CAT:	Figure
GENDER:	Male
PROFESSION:	Child
CLASS:	Heroic
CYCLE:	Trojan
SUB-CYCLE:	Cypria
FATHER:	Priamos
MOTHER:	Hekabe
MATE(S):	
CHILDREN:	
USAGE:	The handsome, young son of Priamos and Hekabe, who bravely escorts his sister, Polyxene, to the fountain to fetch water.
REMARKS:	
US REPS:	16
REFERENCES:	Simon 1973a, passim

Using the Computer-Index of Classical Iconography

How has the system worked in practice? When I was just beginning the process of setting up the database, I asked as many scholars as I could what kinds of questions they expected me to answer. My two favorite replies both occurred at a 1985 international conference on Etruscans. A Danish archaeologist had a brief request: Everything. A Canadian scholar wanted to know the significance of Dionysos riding on an elephant. Fortunately neither has queried my database. Real questions answered recently have been: all subjects i.e., titles) on Greek objects made between 600 - 400 BC with women; all Greek scenes with spinning; all Etruscan mirrors in the Metropolitan Museum of Art; and all objects made or from Umbria, a region in Italy. So far we have been able to answer all questions. The system certainly could be expanded to include genre scenes like banqueting, although we have been able to give partial responses, as in the case of spinning. A related question concerned figures running and the fact that our free text description of each scene is fully searchable enabled us to satisfy that scholar. As a parenthetical to this discussion, you might be interested to know that most everyone wants the data in print-outs and not in machine-readable form.

Related Observations

Having gotten this far in developing a computerized system, I think it is foolish to stop. I want to have it all or rather I do not see why it took me only a couple of minutes to find all the Greek vases with multiple representations of Theseus in the United States and more than three weeks to finish not finding photographs of the remaining fifteen. I don't just want information about objects. I want bibliography. I want literary sources. I want articles. I want books. I want pictures. All on-line. All with one searching engine. Some of these tools are available or in the works, at least for classical studies. All ancient Greek literary sources are available on a CD-ROM from the Thesaurus Linguae Graecae (which is discussed in detail elsewhere in this issue of *Library Hi Tech*). A project has just begun to put into machine-readable form all of *l'Année philologique*, the major bibliographical work for classics. Not only is one search engine needed for disparate types of materials, but the current models need much improvement.

Structured information, as in databases like mine, tends to exist separately from text bases, when the two should be combined to enhance markedly the retrieval from text bases. The problems of inadequate retrieval in text systems exist on two levels: the focus on key words or abstracts; and limited searching engines for full text. While I am more than willing to grant that an excellent abstracter may summarize a given work better than the author, the abstract remains an abstract. It cannot cover all the text, much less the information in the footnotes. Yes, I want the information in the footnotes, especially in the footnotes. Key words are even worse. They cannot give a full picture of the content because a handful of words or even two handfuls and a fistful simply cannot cover the range of topics within a work. Systems that record the frequency of usage of words in free texts present the fewest problems, but even here it should not be forgotten that they too can mislead. If I write an article about Etruria, it is entirely possible that once I have set the stage, so to speak, I may never use the word "Etruria" again, even though it pervades the article. Over and above that, as long as text retrieval systems work only on the level of the individual word or term, they will produce only partial responses.

For a very small investment in effort the current retrieval systems could be significantly improved by combining them with a structured database that sits on top of their regular retrieval system. In this manner if I want information about the Trojan cycle, I won't have to rack my brains to think of all the names of all the participants. My system will not only produce those names, but also it will give them broken down into the various sub-cycles. It will give me non-hierarchical kinds of information like all Greek sculptors. Moreover, my interchangeable modules can be swapped with people in other fields. For example, if someone wanted all references to stone, he or she could take my Materials file, because marble is stone no matter what period is involved. For modern materials, of course, entries for fiberglass, steel, and plastic among others would have to be added. Here Principle Three applies. New words and classifications will be added, others changed, and perhaps even some deleted. Nevertheless, the modular approach means that specific classification sets can be developed independently and individually over a period of time. They can be added gradually or not so gradually to my ideal retrieval system. At this point we come to the real beauty of this all-in-one system. Not one bit—or should I say byte?—of previously encoded text needs to be altered. I know this is true, because I did a test several years ago with an independent bibliographical database. Because the system works on top of existing information, with comparatively little effort an enormous leap in accuracy of recall will occur.

Now I do not mean by any stretch to contend that all retrieval problems are solved. Natural language processing is not yet here, although when it comes it is likely to be dependent on vast dictionaries, or should

I say, classification schemes. The person who wants Paris may get the city, the Trojan hero, or plaster of Paris. Most scholars, however, would rather have too much rather than too little with one caveat. If the wrong responses significantly outnumber the right ones, the scholar will be disgusted, as I was on a recent search of an outside database. As a parenthetical to librarians and information scientists, I find the approaches that emphasize either precision or recall limited, when they should allow the inquirer to specify the degree of accuracy. Sometimes I want only general information; other times I want everything including the kitchen sink; sometimes those needs involve the same information, but at different stages in my research. Choice and flexibility are the operative words.

I have not mentioned the prime use of text bases today for stylistic analysis. I believe, and the recent Rutgers/Princeton University conference on machine-readable texts confirms me in my belief, that the majority of scholars involved with computers in the humanities is thinking in rather limited terms about the possible uses for machine-readable text. The range is not just broader, it also is not fully predictable at this time; as Niels Bohr said, "It is very difficult to make predictions, especially about the future."

In conclusion, you need words. Words not just for pictures, but words for the texts. Many, many words. But what is really very nice is that one set of words will do for both. We are clearly on the brink of having the most marvelous scholarly tools. The Humanities Emporium I have just described can be developed today, and I know how, although its first incarnation on a two shoe-string budget might be "hypotext" rather than hypertext. I am not saying that the path will be one smooth, easy course, but that the most important leap has already been taken. According to Phyllis Rose in *Parallel Lives* (41), "The big step is between not conceiving something at all and conceiving it to be impossible. Once you have conceived it as impossible, it is but a short step to finding it possible—probable—certain."

APPENDIX:

SOME DESIGN PRINCIPLES FOR DATABASES

1. Do not make your datum more accurate than it is. This principle may be rephrased as, "Preserve the Mess."

2. Information should be reduced to its smallest unit or least common denominator.

3. Never ever will any person or project, no matter how knowledgeable and experienced, be able to put together a list of words that will not need to be changed...continually.

4. No controlled vocabulary should be produced in the absence of actual usage.

5. The amount of effort to record a particular piece of information must be weighed against the amount of usefulness returned from that piece of information.

   Corollary 1: Utility always takes precedence over philosophy.

   Corollary 2: Not all information useful to a scholar is worth recording.

6. It is easier to catalog whole groups of entities than to remember which ones are the right ones.

7. No codes...ever.

8. Everything, and especially free text, should be fully and Boolean searchable.

9. You should always buy a program more powerful than you anticipate you will need, since inevitably you will not have anticipated everything you will need, for which compare Principle Three.

10. The official rule for relational databases that no piece of information should be repeated exists only to be broken. If using the database is easier with information repeated, then do so. Storage is the least of the problems.

    Corollary: Multiple and overlapping points of access allow for better retrieval.

NOTES

1. This scene is in the Tomb of the Bulls in Tarquinia, and is dated to ca. 530 BC. Illustrations may be found in a number of sources, among which are M. Pallottino, *Etruscan Painting*. (Geneva: 1952), 31; and S. Steingräber, *Catalogo ragionato della pittura etrusca* (Milan: 1985), 157-158.

(Notes continued on page 67.)

is a meaningless 5-1/4 inch disk. As noted above, some 500 TLG CD-ROMs are currently in circulation worldwide. Of this total, twenty have been acquired by university libraries. In virtually all twenty instances, the TLG has received complaints from the members of the respective institutions that—though willing (or even eager) to acquire TLG CD-ROMs—the librarians lack either the ability or the inclination, or both, to procure and maintain the highly specialized hardware and software necessary to make meaningful use of the TLG's disk. Many librarians, these complainants tell us, still tend to make a distinction between information on one hand and the means to manipulate information on the other. Today's librarians, they feel, have the responsibility to acquire and maintain not only collections such as those represented by the TLG but also all of the technological resources necessary to gain access to these electronic collections.

Many other scholars, however, oppose centralization of electronic resources within a university library: while granting that a TLG CD-ROM is fundamentally a collection of texts, they argue that the very nature of the compact disk calls for individual (or at most departmental) CD-ROM ownership and control, and that efforts on the part of librarians to acquire and control certain types of electronic data constitute an unwarranted intrusion on the part of librarians into basic research processes.

If the last twenty years have produced vast amounts of electronic resources such as those represented by the TLG data bank, the coming decade will most likely witness a struggle between efforts to centralize these resources and attempts to keep them in the hands of the individual scholarly user. It is far too early to predict the outcome of this struggle. It is clear, however, that, thanks to the Thesaurus Linguae Graecae, classics is today a discipline far different from what it was twenty years ago. In a way, classics and the Thesaurus Linguae Graecae represent a paradigm: the oldest of the humanistic disciplines, for centuries wedded to traditional research methodologies, was suddenly transformed through the acceptance of modern technology. Other humanistic fields can be expected to follow suit as more and more humanities-oriented data banks come into being. In the final analysis, the humanities and technology need no longer simply coexist as friendly antagonists, but can truly benefit one another. It is perhaps ironic that, in the case of classics, it will be the computer—the most advanced product of modern technology, indeed the symbol of modern technology—that will help man to achieve a full understanding of those ancient writings that constitute the very roots of Western thought and civilization.

(Notes continued from page 60.)

2. Every two years a volume of text with a separate volume of plates is published by Artemis Verlag of Zurich.

3. I am extremely grateful to the National Endowment of the Humanities, Research Tools and Rutgers University for their continuing support. For the current grant the *US LIMC* has also received funding from the Samuel H. Kress Foundation.

4. For a summary of the various principles (surely not all that there are) propounded herein, see the Appendix—"Some Design Principles for Databases."

5. Advanced Revelation is produced by Revelation Technologies, Inc. in Seattle, Washington. It runs on IBM PCs and its clones.

6. This information was reported at the meeting of the Text Encoding Initiative (TEI) in Chicago on 18 February 1989, and is recorded in TEI document number, TEI AB M 1, p. 8.

7. New York, Metropolitan Museum of Art 01.8.6 by the C Painter. J. D. Beazley, *Attic Black-Figure Vase-Painters* (Oxford: 1956) 51, no. 4. T. H. Carpenter, *Beazley Addenda—Additional References to ABV, ARV ^{2} and Paralipomena.* 2nd edition (Oxford: 1989), 13. *The Metropolitan Museum of Art Bulletin* 31:1 (Fall 1972): no. 5.

8. H. van de Waal, *ICONCLASS. An Iconographic Classification System*, vol. 8-9. (Amsterdam, Oxford, and New York: 1980), 131.