MAARAV 5-6 (Spring, 1990): 23-32

CYBERNETICA MESOPOTAMICA[1]

GIORGIO BUCCELLATI
UNIVERSITY OF CALIFORNIA, LOS ANGELES

Under the watchful eye of its patron deity, Nisaba, the art of writing has progressed along paths that only the gods could have foreseen some 5000 years ago. By most standards, we have come to associate "history" with the evidence provided primarily by the written record, and "prehistory" with the evidence provided exclusively by the archaeological record. By the same standards, we may be tempted to envisage our age as the beginning of a third major phase in human history. The computer holds a status, vis-à-vis writing, quite analogous to what writing held vis-à-vis preliterate or pre-scribal information techniques, and as such it may be heralding a major transition in human development into some sort of "post-history." But just as we are being weaned from Nisaba's maternal cares, and as we are trading clay for silicone, we seem to be growing in our commitment to the past, to the domain over which she has ruled for so long. It seems ironical to say that no Mesopotamian ever knew as much about his cultural past as we know today, but there is much that is true in this assertion: while our accent in speaking their language may be at best abominable, which Mesopotamian scribe could ever have had access to as full a gamut of text types, and from as many different periods, as we do today?

The reasons why the computer is of such dramatic consequence as to lead us slowly beyond the realm of Nisaba are much more

1. The text of this paper follows that of an oral presentation given at the national meeting of the American Oriental Society in Los Angeles in March, 1987. I dedicate it with friendship to Stanislav Segert, with a special personal recollection of our first acquaintance in Chicago, at a time when he was pioneering the use of the computer on a research project he had undertaken with I. J. Gelb.

than just greater ease or increased efficiency. Not that ease and efficiency are unimportant: most of us have become apt at using the computer for both word-processing and, perhaps to a more limited extent, for data base management—two functions which have certainly made a convert of many a skeptic in recent years. But the major impact to be felt in the very immediate future is in the way in which we *conceptualize* both the data and the utilization we can make of it. I would like to address here some aspects of these changes. I will do this in a rather minor key, because I will deal with the practical dimension rather than with theoretical issues. Specifically, I would like first to speak about the general scope of our project; second, to explain the notion of distribution disks; and third, to illustrate the specific disks which are currently available.

I started applying data processing techniques to Mesopotamian materials back in the age of dinosaurs—1968, to be precise. The dinosaurs were the large mainframe computers: no thought, then, of microcomputers and of general widespread use of magnetic media. By 1972 we had progressed, with the support of funds from the Research Committee of the Academic Senate at UCLA, to a point where we felt ready for a major grant application from NEH, which we in fact received. Under the terms of the grants we were to produce a computerized data base of Old Babylonian letters: this was accomplished over a period of 5 years, but we were then faced with the problem of distribution. Printouts of the basic sign concordance ran upwards of 10,000 pages; morphological concordances which we had completed for sections of the data base were even larger: the sheer physical volume was such that any type of standard publication was out of the question. Thus I kept making available to interested colleagues tapes and portions of the printouts upon request, and gave much thought to ways of compressing the data through editorial compacting of one type or another. I started within Undena a book series under the title *Cybernetica Mesopotamica* which was to make available the intended results, and began publishing portions of the data sets as well as some of the analysis.

With the advent and the widespread use of microcomputers, the situation changed drastically. We could plan on distributing data on magnetic media, which would not only be much more realistic economically, but also allow the interactive type of utilization of the data that constitutes a major difference vis-à-vis paper-base

products. Through the major support of the Ambassador International Cultural Foundation I was able to begin using microcomputers for archaeological field work, and in fact we were the first to bring "microcomputers" to Syria for data entry in the field. This was back in 1981, when micro-computers were all but micro in size (although they were certainly so in power): unimaginable as it may seem today, IBM was not yet in the market at the time. The work on archaeological materials has resulted in a very complex and comprehensive system of coding and analysis especially for stratigraphic information: this changes in essential ways the mechanism for arriving at strategy decisions during the excavations, by providing much greater capillary control on stratigraphic details; it also adds a whole new dimension of objectivity to the record, since it makes it possible to publish, if one so chooses, the full range of observations made in the field, without the selectivity that has otherwise been necessary in archaeological publications. This is what I call the "global record," which we are in the process of implementing at both Terqa and Tell Mozan—but which I have no time to illustrate here.

If I have referred to it at all, it is because the archaeological component is as essential, conceptually, to the overall scope of *Cybernetica Mesopotamica* as the philological one, and shares in much the same way the technical problems attendant to the formalization of the data and the manner of distribution. I will however concentrate, for the rest of my paper, on the philological dimension alone. Work on this has been made possible recently through the support of the David and Lucile Packard Foundation: under the terms of this grant we are currently concentrating on the texts of Ebla, and hope to be able to continue subsequently with the Old Babylonian letters (prepared especially by John Hayes, Paul Gaebelein and Yoshi Kobayashi), the Middle Assyrian corpus (prepared by Claudio Saporetti), and the Akkadian of the West (prepared by John Hayes and Thomas Finley)—all of which were already available on the mainframe computer and have to be harmonized and revised thoroughly for distribution on floppy disks.

The study of the Ebla texts is of obvious significance. My own work on these texts along the lines described here follows the mandate I was given within the framework of the International Committee established by the University of Rome, and will grow apace with the work done by Alfonso Archi and the epigraphic staff of the Expedition. In this respect I benefitted especially from

the presence of Lucio Milano at UCLA where he served for two years as Visiting Professor. Through our collaboration with him and our other Italian colleagues we have been adding texts as yet unpublished to our data base. As a result, not only is the preparatory work for publication made easier for the editor, but also we can hope to have the data pertaining to future volumes of the *Archivi Reali di Ebla* very quickly available in the format which I am describing.

The second part of my paper deals with the general concept of distribution disks. To disseminate data in disk form is an obvious step to take, given the current general availability of personal computers. Yet there are some significant considerations to be made. While computer use is widespread, it is not necessarily so as yet for computer literacy. By this I mean that in a broad sense computer use is generally limited to its lowest common denominator: in an undergraduate class of 60 which I taught recently, some 40 students produced their papers on word processors; but when I made class notes available on disk, only two students took advantage of the opportunity. A similar ratio might obtain among the intended users of *Cybernetica Mesopotamica.* In distributing data on disk, we aim therefore not only to serve, but also to stimulate, a need.

At this point, my first goal is to make the *primary data* available in a *cohesive and well-structured* format. The second goal is to provide *programs* that will allow an effective *interactive use* of the data.

As for the data: The most basic format of our distribution disks is that of graphemic transliteration. While this may appear simple enough, real problems arise as soon as one tries to build a data base that goes beyond one type of text: harmonization is critical if one is to develop a uniform approach to data retrieval, as for instance with sign concordances. In addition, standard transliterations are hybrid systems which include other levels of representation besides graphemics. We have tried to give as full and as precise a representation of the graphemic level as possible: the resulting encoding system seems to us sufficiently transparent to be accessible to Assyriologists, but at the same time sufficiently differentiated to account for all relevant graphemic phenomena. These data can be used with any conventional word-processor, utilizing at a minimum such simple but powerful functions as word or

character searches. To minimize potential problems, only basic ASCII characters (essentially, those that are visible on the keyboard) are used in the primary encoding, so that for instance special characters such as *šin* are represented by a sequence of two characters (s and circumflex in this particular case). Separate programs are provided for those who wish to change their data to either show on the screen or print on paper special characters, including cuneiform.

Certain sets of data will be coded for aspects other than graphemic, and issued as such on separate disks. For instance, the royal letters of Babylon had been fully encoded on the mainframe computer for morpho-lexical analysis by John Hayes, myself and others, and have also been analyzed for historical categories by Patricia Oliansky and others. Morpho-lexical analysis of the personal names from Ebla had been started under the supervision of I. J. Gelb, and will now be continued in collaboration with Alfonso Archi, Pelio Fronzaroli and Lucio Milano by James Platt, Joseph Pagan, Mark Arrington and myself. The encoding manual for morpho-lexical categories, which I have prepared, corresponds in some measure to a compacted version of Akkadian grammar.

The full use of the data will however be done in an interactive environment through some type of programs. The format of our data files is so clearly defined that those adept at it can more easily use them in their programming. But we will at the same time make available a number of our own programs which will allow various types of manipulation of the data. Some are simple utility programs which allow to format or manipulate the texts in different ways for various types of personal use. Some are more complex interactive programs that allow, for instance, to derive selective indices within or across disk boundaries. More complex programs yet will allow in depth analysis with, for instance, a statistical base. Finally, we also recommend certain commercial programs that seem especially useful for interfacing with our data. Some programs will be made available routinely on data disks, but program disks as such will also be issued periodically, and will contain updated versions of all major non-commercial programs available directly from us. One direct result of this approach will be, I hope, an effective contribution to real computer literacy, in the sense, for instance, that these programs may be used not only for our own data disks, but also for data being prepared for publication or simply being studied by individual scholars and students.

An ancillary type of disks will contain secondary literature. On the one hand, we will distribute on disk the text portion of some of the volumes which are being published in book form by Undena. The first example is *Mozan 1*, which will be available shortly. These disks will contain exactly the same text as the printed version, but without formatting commands, so that most word processors will be able to access them. They will thus serve primarily the purpose of greater convenience in scanning through the text portion. A different type of secondary literature will consist of compilations of secondary literature prepared specifically for disk use, whether or not they are also made available at the same time in paper format. One such example is an extensive annotated bibliography on Akkadian grammar, which I have in advanced state of preparation.

As for some practical aspects. Each disk contains one major body of texts, generally corresponding in scope to a book size publication. We are restricting for now our distribution disks to MS-DOS format. Besides the data themselves in graphemic format, there are a number of introductory files, and supplementary files—which I will explain in a moment in connection with three distribution disks which are currently available. While we retain the copyright on the disks to protect them from abuse, there is no limitation on copying and at $1 per disk even the initial distribution price is nominal. Even though in some cases there is room left on the disks, we will in general maintain this correlation between disks and bodies of data: the cost of a disk is a negligible factor, and distribution of, as well as bibliographical references to, disks are easier if such a correlation is maintained. Where necessary, we will use archival compacting to store larger data files that belong together on one disk. At any rate, disks should be considered on the same level as book publications, with an autonomous bibliographical validity of their own.

I will come now to an illustration of specific examples of distribution disks, using three titles which are currently available. The *first* one (*CMT* 1) contains all texts from Terqa excavated through the fourth season, as well as a new edition of all known Khana texts: the texts have been established and annotated by Olivier Rouault, Amanda Podany, and myself. The *second* disk (*CMT* 2) contains the Middle Assyrian Laws as established by Claudio Saporetti. The *third* disk contains indices of all words, numerals

and names from *ARET2*, and has been established by James Platt, Joseph Pagan and Mark Arrington.

I have subdivided the distribution disks into five major categories: the data, subdivided in turn into texts (*CMT*) and archaeological materials (*CMA*), the programs (*CMP*), the secondary literature (*CMB*) and the indices (*CMX*). Within these categories, disks will be numbered sequentially without concern for further subtypes, as I had tried to do originally in the first prospectus of *Cybernetica Mesopotamica*. The reason for the elaborate structure envisaged there was to bring out the structural relationships among the various components, which I had at that time anticipated would be numerous. The kind but firm criticism which has recently been made by Edzard of this approach misses, I think, this point and yet, I could have shown validity of the overall system only if the various parts had in fact come to fruition. Now that we can rely on disks for a much more effective and inexpensive distribution of the data, the paper base collection *Cybernetica Mesopotamica* will be limited only to a few representative volumes of data, and otherwise to analytical titles. The electronic files on the other hand will consist exclusively of data and of programs which, when combined, will produce interactively the kind of results that were envisaged as separate series within the paper base collection.

Disks will be further labeled with a letter that will mark the particular edition or "generation" represented in that particular disk. Because of the ease of updating, and because distribution disks will be produced on demand thereby including always the latest "generation," it is important that this be made part of the formal disk "title." The table of contents of a disk, or "directory," is shown as Fig. 1. This is actually a sort of template, a list of all possible files that may occur on a disk, but in practice only some will be found on any given disk. The first file can be activated by pressing an exclamation mark on the keyboard: this will provide basic start-up information, and will proceed to read, as desired, all other introductory files on the disk. Introductory files are marked by a prefix enclosed within two hyphens: files with the prefix C contain information about the system *Cybernetica Mesopotamica* as a whole; prefix D refers to the data contained on that particular disk; prefix E refers to encoding rules, i.e. it provides the encoding manual for the data given: prefix F refers to the file formatting characteristics used in the disk; prefix G refers to general overviews of the data, such as statistical summaries or compositional analysis;

prefix I refers to indices and prefix P to programs: while separate indices files are redundant because they can be generated through the use of programs provided, they are occasionally included for the benefit of those who are not as yet fully familiar with the operation of the computer. (It is for the same reason that I have avoided using subdirectories and have resorted instead to the use of file prefixes.)

The second set of files, each beginning with a letter T for text (or A for Artifacts in the case of archaeological data) contains the data. Normally, each file will correspond to a single text, although this is not required. Files are labeled sequentially, and the label is repeated on the first line of the file itself, followed by the actual bibliographical reference to the text itself.

As indicated, utility programs allow various types of reformatting. For instance, one program merges the individual data files into a single large file, and adds the bibliographical reference in front of each line: while this makes the size of the file considerably larger, and makes it less manageable for certain operations or certain word processors, it may in other cases serve a useful purpose, for instance for global searches within the corpus. Another program changes the two character clusters (e.g. s and circumflex for shin) into single characters while yet another (not fully implemented yet) will provide the possibility of converting a normal transliteration into standardized cuneiform characters, without reentering the data. Two cuneiform programs are planned to work in conjunction with two commercially available programs which already provide a number of other advantages a relatively low cost (Lettrix and Multi-Lingual Scholar).

-C-FILES

FILES ON CM DIRECTORIES (March 21, 1987)

CM directories may contain any one of the following files

!.EXE	gives first orientation and reads introductory files
-C-	(prefix for general files about Cybernetica Mesopotamica)
-C—PREF '	brief preface to system as a whole
-C-FILES	describes introductory files (=i.e., this file)

-C-INTR	describes Cybernetica Mesopotamica as a system
-C-SIGNS	cuneiform signs with unknown reading, not in standard lists
-C-TITLS	catalog of titles in Cybernetica Mesopotamica
-C-UTIL	utility programs available for further use of data bases
-D-	(prefix for files dealing with data on this disk)
-D—PREF	brief preface to data base on this disk
-D—VERS	synopsis of characteristics of version on this disk
-D-BIBL	list of references sorted in bibliographical order
-D-CATEG	identification of data items by provenience, date, type, etc.
-D-EPIGR	documents as artifacts (field numbers, archaeological settings)
-D-HARMO	harmonization principles and changes from established edition
-D-INTRO	general introduction to data base included in this disk
-D-NOTES	notes on specific, unique passages
-D-REF	list of references sorted in the order of the data
-D-REFCO	concordance between data files and references
-E-	(prefix for files with information about encoding)
-E-TG	encoding manual for texts, graphemic format
-E-TM	encoding manual for texts, morphological format
-F-	(prefix for files which explain file format for data files)
-F-G	data entry format for texts, connected graphemic version
-F-GI	output format for texts, alphabetical list of items
-F-GO	output format for texts, alphabetical list of occurrences
-F-GT	same as -F-G, but with volume label on each text line
-G-	(prefix for general information outside current scope)
-G-COMPS	compositional analysis of basic data unit (e.g., text outline)
-G-TAB	tabulation of numeric data (e.g. summaries of entries by type)
-I-Xxxa	(prefix for files with indices to data files) X - CME data files (A, S or T) xx - volume (1 though 99) a - version (a through z)
-I-Xxxa#.GI	numerals: lexical items with totals
-I-Xxxa#.GO	occurrences with references
-I-XxxaD.GI	divine names: lexical items with totals
-I-XxxaD.GO	occurrences with references
-I-XxxaG.GI	geographical names: lexical items with totals

-I-XxxaG.GO occurrences with references
-I-XxxaN.GI names (various): lexical items with totals
-I-XxxaN.GO occurrences with references
-I-XxxaP.GI personal names: lexical items with totals
-I-XxxaP.GO occurrences with references
-I-XxxaW.GI words (other than above): lexical items with totals
-I-XxxaW.GO occurrences with references

-P- (prefix for program files)
-P-xxxxx.XPL explains goals and procedures for program "xxxxx" (e.g., ADDT)
-P-ADDT adds publication titles to individual records
-P-NDXO establishes lexical items by lexical items
-P-NDXI establishes lexical items by items
-P-REMT removes publication titles from individual records

Xxxa(-zz). data files labeled as follows:

- X CME (Cybernetica Mesopotamica, Electronic files)
  - A - archaeological materials
  - S - secondary literature
  - T - texts
- xx sequential number of CMX disk (from -1 to 99)
  - A1 to T99
- a generation of CMX disk (from A to Z)
  - A1a to T99z
- (-zz) sequential number of files for given disk (optional)
  - A1a-01 to T99z999
- -(—) extension identifying type of data
  - G - graphemic version
  - GO - graphemic index of occurrences

T1A-01.G

- T CME (Cybernetica Mesopotamica, Electronic Files)
- 1 disk number
- A generation
- -01 text number
- .G extension specifying format

SPRING 1990 VOLS. 5 - 6 ISSN 0149-5712

MAARAV

A JOURNAL FOR THE STUDY OF THE NORTHWEST SEMITIC LANGUAGES AND LITERATURES

SOPHER MAHIR:

NORTHWEST SEMITIC STUDIES PRESENTED TO STANISLAV SEGERT

EDITED BY EDWARD M. COOK