Global data banks: a wise choice or foolish mistake?

Metz, Rainer

Empfohlene Zitierung / Suggested Citation:
Metz, R. (1991). Global data banks: a wise choice or foolish mistake? *Historical Social Research*, *16*(1), 98-102.
https://doi.org/10.12759/hsr.16.1991.1.98-102

(25) Besonders sind hier die »Zeitstufendaten« der fünften Merkmalsgruppe betroffen.
(26) Z.B. für die Erstellung von Reports etc.
(27) SPSS–9: erstellt formatierte und unformatierte Dateien; Rechenzentrum: Wandlung von SPSS-Systemdateien in Rohdaten mit optimaler Formatwahl und Wandlung in BMDP-Dateien.
(28) David Wishart, *CLUSTAN User Manual*, 3rd ed. (Edinburgh: 1978); und ders., *CLUSTAN – Benutzerhandbuch (3. Ausgabe)* (Stuttgart: 1984).
(29) W.J. Dixon, *BMDP–82* (Berkely: 1983); und auch Günter Bollinger u.a., *BMDP – Statistikprogramme für die Bio-, Human- und Sozialwissenschaften* (Stuttgart: 1983).
(30) Shirrel und Roal Buhler, *P-Stat 78 User's Manual* (Princeton: 1979).
(31) Vgl. Cornelius C. Goeters, »Anhang: Karten«, in: Wolfgang Ribbe (Hrsg.), *Das Havelland im Mittelalter. Untersuchungen zur Strukturgeschichte einer ostelbischen Landschaft in slawischer und deutscher Zeit*, Germania Slavica V = Berliner Historische Studien 13 (Berlin: 1987), S. 463–475, hier S. 464f.

Global Data Banks: A Wise Choice or Foolish Mistake?

Rainer M etz*

The present need for information systems for historical data is already quite large and continues to grow at a rapid rate. Our general concern here involves the integration of existing data into a »system« where they can not only be retrieved and processed comfortably, but also compiled in their totality and combined with other data sets if so desired, thus ensuring rapid access to large stocks of information. Equipped with the appropriate hardware and software, such electronic data banks are ideally suited to take advantage of the latest developments in the world of electronic data processing, where speed, precision and the flexible reproducibility of the obtained results are its key features. In view of the breakneck speed with which both hardware and software are being developed, there would seem to be no major obstacles standing in the way of designing historical databanks. Indeed, their implementation should be able to proceed without

* Address all communications to: Rainer Metz, Zentrum für Historische Sozialforschung, Bachemerstr. 40, D-5000 Köln 41.

further ado. Yet a closer look reveals that matters are not quite that simple, as I will explain in the following. This problem can be best illustrated in terms of the advantages one generally expects from a data bank.

1) One of the main functions of a data bank is to provide rapid information on a particular topic (key word), thus eliminating the tedious task of looking up information in standard reference works and special publications. In most cases, a search carried out with only one key word produces vague results, for the desired information is often too complex to be narrowed down to a single word. A successful retrieval often requires the formulation of several search terms linked by qualifiers such as »and« or »or«. The ability to link search terms logically is the main prerequisite of a data bank and is generally referred to as its »lexicon« or »retrieval« function.

2) In addition, this retrieval function must be flexible and capable of presenting information in a way that conforms to the user's needs, for example, in tabular and graphic (including cartographical) format. Presenting stored information on a 1:1 basis is considered inadequate in most cases, for the user often requires relative or index figures which first must be calculated from the original data. For this reason, a flexible means of presentation should be required for data banks of this type.

3) The features just mentioned in point 2) above should also include the capacity to transfer retrieved information to other software packages. The output should be designed in such a way that it can be easily utilized as input for other software packages, or for at least the more popular commercial ones. The simple reason for this is that no data bank can be expected to provide system components capable of responding to every type of query and research method. Just consider the broad range of features currently available in most statistics programs or the requirements expected of desktop publishing systems. In contrast to commercial uses, this »interfacing« feature is of decisive importance for scientific research.

4) Lastly, a few points regarding the quality of stored information should be mentioned. The speed alone of electronic data processing can often give one the false impression that the available data base represents a certain *totality*. For example, a user who needs information concerning agricultural prices in the 16th century doubtlessly expects to find a comprehensive base of information organized by subject and region. It would not suffice if the data bank only contained the prices quoted in a single publication, for this would rob the user of a valuable base of comparison. Besides this often tacit suggestion of totality, the up-to-dateness of data remains an indispensable requirement. However, up-todateness in terms of historical data suggests something quite different than its normal use. Here

it means the need to include the latest research findings for a given field of study. This feature is highly relevant for updating a databank. As a final point concerning the quality of data, I would like to point out that presenting information on a 1:1 basis is at best an inadequate solution. The retrieved information should not only be documented but also annotated. The user's wish for *information about information* is all the more justifiable when this information is of a quanitative nature. This too has several important repercussions on the design and implementation of a databank.

The considerations outlined in the four points above doubtlessly impose very high demands on the performance of historical databanks. Since the task of compiling and organizing data banks of this type involves a great deal of time and effort, the realization of these demands should be given high priority. Admittedly, this does present some immediate problems in practice. To illustrate the scope of these problems on a practical level, let us examine the present state of development of the »Medieval and Early Modern Data Bank« (MEMDB) and its future plans. (1) The MEMDB project was set up at Rutgers University in 1982 with the aim of establishing a data bank of information covering the economic history of the Middle Ages and early modern times. While historians were put in charge of outlining the data bank's contents and compiling data, the project's technical engineering was carried out by computer specialists from Stanford's Research Libraries Group (RLG). This two-track design strategy turned out to be quite a boon for the data bank, especially with regard to its userfriendliness.

Up to now, the MEMDB project has only developed a PC-based prototype whose data base consists of over 13,000 currency exchange quotations published in Peter Spufford's »Handbook of Medieval Exchange« in 1986. (2) In general, the data have been incorporated into the data bank unaltered, i.e. as originally published, and indexed according to a certain number of reference terms by which the data can be retrieved. »Advanced Revelation« is the name of the the data bank's software package. Used in conjunction with the skillfully assigned index entries, it provides the user with a powerful tool for information retrieval. Data can be analyzed, retrieved and organized in a easy and flexible manner with the use of only ten commands. (The relatively complicated and clumsy SQL commands do not seem to be essential). Future plans involve the installation of the data bank in Stanford's main-frame computer which can be accessed through existing US and European networks. Main-frame implementation is considered necessary for at least two reasons.

1) Even on a relatively fast personal computer, a query consisting of only two search terms coupled by a single qualifier leads to unacceptably long search times.

2) The demands placed on global data banks in terms of comprehensiveness and up-to-dateness cannot be met by any PC, either now or in the near future.

The next problem is how to maintain the high level of performance currently offered by MEMDB when data sets of a different structure are added to the master data. The problems most likely to arise here can be illustrated by a simple example. Plans are being made to expand the MEMDB master data set, for example, by incorporating additional currency quotations. These data are already available. For example, the data of my book »Coins, Money of Accounts and Price Movements. The Lower Rhine Region in a European Context: 1350 - 1800« (3). If these data were to be incorporated, entries having the form »Cologne, 1442, 1 Gold Gulden equals 24 Albus« would be added to the existing ones. However, Spufford's entries in the data bank bear a different notation, e.g. »Cologne 1423, Rhinish Florin equals 24 Weisspfennig«. In this case the data bank fails to recognize and assign identical objects, such as »Rhinish Florin equals Gold Gulden« and »Albus equals »Weisspfennig«. This method of storing information evidently does not allow for a categorical linking of different »strings«. A correct linking would require the help of a concordance or thesaurus available as a separate module in the data bank.

This simple example alone demonstrates that no search or sort function will produce the desired result unless it is supplemented by additional reference aids. This problem will occur not only when processing data sets taken from diffenent languages but also when only one language is used. The inexactness (»blurriness«) of historical sources - partly due to orthographical variants - represents a particular problem in data processing. While in the above case the structure of information is identical (currency quotations), its presentation is not. Thus a data bank designed in this manner is incapable of transforming or incorporating new sets of data; new information must therefore be indexed according to standard terms or categories. The time and intellectual effort required by such a task should not be underestimated.

It is conceivable that this kind of transformation, or labeling, of information is possible when data have a similar structure. But what about data which are entirely different in structure and category? For example, what is the point in storing household data for Lübeck and currency quotations for the Lower Rhine region in one and the same data bank? As far as I can see, none at all! The sensible solution is not to set up global data banks in which »everything« is stored, but to establish separate data banks for special subjects. For example, a data bank of German historical statistics(4) would be feasible only if it incorporates data of related categories. How can data concerning the economical structure of Württemberg in the 16th

century be combined in an intelligable and research-oriented manner with power economy statistics from the 19th century or with public health figures from the 19th century? These considerations clearly show that there is no point in trying to store »everything« in a single data bank. Furthermore, they show that storing data on a 1:1 basis is a completely inadequate solution. Instead, data should be labeled according to additional categories relating to time period, region and subject matter. The categories neccessary to accomplish this, i.e. the overall index structure, should not be gleaned from each publication but should reflect a specifically designed research concept. Inasmuch as global data banks are bound to fail due to the expense involved in their implementation as well as to their inability to carry out necessary data transformations, much attention is starting to focus on data banks designed with a clearly defined data content. The concept of a single »global« data bank is being replaced by one of »modular« data banks. The success of this concept requires a unified software package (or at least a system of compatible software) and efficient interface components. These components should be capable of transferring data not only to other data banks but also to various analysis systems.

One large problem which accompanies this line of thought is doubtlessly the need for compatible software and a universal user interface. Once this can be realized however, there is no stopping the move from global to modular data banks. For in their entirety, modular data banks will ultimately prove to be a flexible and comprehensive means of information management.

Notes

(1) The addresses for information about MEMDB are: Prof. Rudolph Bell, Department of History, CN 5059, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08903 USA; or: Dr. Rainer Metz, Zentrum für Historische Sozialforschung, Bachemerstr. 40, 5000 Köln 41.

(2) Peter Spufford, Handbook of Medieval Exchange (Royal Historical Society Guides and Handbooks 13), London 1986.

(3) Rainer Metz, Geld, Währung und Preisentwicklung. Der Niederrheinraum im europäischen Vergleich: 1350- 1800 (= Schriften des Instituts für bankhistorische Forschung e.V. 14). Frankfurt 1990.

(4) Andreas Kunz, Eine Datenbank zur Historischen Statistik. In: N. Diederich, E. Holder, A. Kunz u.a. (Hg.), Historische Statistik in der Bundesrepublik Deutschland (= Bd. 15 der Schriftenreihe *Forum der Bundesstatistik*, hrsg. v. Statistischen Bundesamt), Stuttgart 1990, S. 159 - 163