## Werk

**Titel:** Maintenance and Preservation of Large Databases

**Autor:** Lievesley, Denise

**Ort:** Graz

**Jahr:** 1996

**PURL:** https://resolver.sub.uni-goettingen.de/purl?514854804_0006|log72

## Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

*European Research Libraries Cooperation:*
*The LIBER Quarterly, 6 (1996), 472-482.*

# Maintenance and Preservation of Large Databases

DENISE LIEVESLEY
*UK Data Archive*

## Introduction to the UK Data Archive

The UK Data Archive is the leading national library of electronic data in the broad fields of social and economic science. The Archive was established in 1967 and is located at the University of Essex. It is core funded by the Economic and Social Research Council and the Higher Education Funding Councils. The Archive incorporates a number of different organisations namely the Economic and Social Research Council Archive, r-cade (the Resource Centre for Access to Data on Europe), the History Data Service (which is part of the Arts and Humanities Data Service) and the Virtual Psychology Laboratory.

The Archive is part of the Institute for the Social Sciences at the University of Essex which has recently been designated a large scale facility in the social sciences by the EU. This is one of only two in Europe. It will provide a centre with good computing facilities together with technical support and most importantly access to the rich data resources of the Archive where researchers from across Europe can come to work on comparative social research.

**Functions of the Data Archive**

The Archive exists in order to promote wider and more informed use of data in teaching and research and to preserve these data so that they continue to be accessible over time. Data are not only provided to the academic community but may be distributed more widely depending upon the conditions specified by the depositor of the particular dataset. The Archive comprises about 7000 datasets and is increasing at the rate of one dataset per working day. About 6000 datasets or sets of documentation were distributed last year

The staff have expanded in numbers quite considerably over recent years and now have reached almost 40. To learn more about the organisation send an e-mail to archive@essex.ac.uk or read the home page http://dawww.essex.ac.uk/

The main functions of the archive are as follows

- establishing user needs
- negotiation and acquisition
- validation of data and documentation
- supplementing documentation
- preservation
- cataloguing and indexing
- reformatting and delivery to users
- promoting use and supporting users

**Communities serviced by the Archive**

The role and value of data archives are expanding due to an explosion of electronic data and the increasing cost of primary data collection, as well as the burden on respondents to provide further information, which makes the use of secondary data an economically attractive proposition. The culture has changed in many disciplines too which means that secondary data analysis is not seen as an inferior method. There is an increasing recognition

of the value of getting data used and of making research transparent and enabling others to replicate or extend one's research. Importantly too as the amount of electronic data has increased so has an awareness that its management has to be active if the data are to be readable in the long term. Specialist skills and equipment are needed for data preservation.

Two communities are served by data archives

- users of data - teachers and researchers
- data producers, owners and funders

The benefits to the users are more obvious. They obtain expensive resources cheaply ; high quality research is promoted as a result of this access; it encourages the re-analysis of data from a different perspective; and the access to data in electronic form permits a level and depth of analysis which cannot be undertaken with published material.

Depositors comprise a wide range of different individuals and organisations and it is more difficult to generalise about the benefits of archives to them since they might have very different aims and requirements. Data are received at the UK Data Archive from the following categories of depositor:

- national and local government departments and agencies
- other public bodies
- non-governmental organisations
- individual academic researchers and academic centres
- independent research units
- commercial sector including market research organisations
- other data archives worldwide

The benefits to depositors of depositing their data can be extensive. One of the most convincing arguments is that their own data will be preserved and will be available for their own use, since few data collecting organisations understand how to manage electronic information over time. As a result of deposit in an archive which actively promotes the use of data, their data will

obtain wider usage and this should result in more citations and a higher profile for those who were responsible for collecting or funding the data. If the depositor requires it then it is possible to give feedback on the use of data and a relationship between user and producer may be facilitated. On the other hand if the depositor prefers, the archive can act as a buffer protecting him from questions from users about the data, which can often anyway be answered by reference to the documentation or are a reflection of the user's inexperience and he or she can be helped by archive staff. The quality of data will often be improved by the checks which the archive may carry out and similarly documentation may be supplemented and sometimes value may be added, say by combining different datasets or adding information on quality. The UK Data Archive operates a policy of allowing the depositor to specify which categories of people may have access to the data and whether royalty payments are to be levied- though at a minimum there must be free access for academic research. Finally an argument in favour of sharing electronic data is that of altruism. The availability of more extensive and higher quality electronic data should result in better trained researchers and a more discriminating user community. Too often university courses are taught using dummy or simulated datasets which results in students having little appreciation of the richness or conversely the limitations of real data.

**Delivery of data**

The mission of the Data Archive at its most basic level is essentially twofold: to preserve machine-readable social science research data for posterity and to provide these data to researchers. Enabling access to its holdings is, therefore, of primary importance to the Archive.

The Archive provides deposited data and machine-readable documentation in a variety of formats and media to users who

have received the necessary permission from the depositor. Most archived datasets are held in a variety of formats, the most popular at the moment being SPSS, SAS, SIR and STATA. The Data Delivery section translates datasets into different formats and different platforms (Macintosh, DOS, UNIX) for users as needed, and will provide subsets of large datasets when requested. We provide data on a variety of media, ranging from floppy disk to CD-ROM to exabyte and DAT tape. We also have the facility for lesser known media, such as a portable SCSI drive, quarter-inch cartridge and even the old large magnetic tapes. Becoming ever more popular is provision of data by ftp (file transfer protocol) which involves no physical medium at all, but a computer-to-computer transfer.

Over the last three years, the Archive has processed an average of 126 orders per month. This includes access orders, orders requesting direct access via another UK facility, and orders for paper documentation only, all of which take time to process. The orders for data range in size and complexity from requests for one small dataset on floppy disk which can be completed in a matter of minutes, to CD-ROM versions of every single year of a large government data series in a custom-made format which may take several days. The Archive works with depositors to ensure that the documentation which accompanies data is comprehensive and relevant to users. Guidelines evolve as the range of data types increases.

Fast data delivery is an important issue for some users whereas others have a more relaxed timetable and our systems accommodate both with the facility for prioritising urgent orders.

## Control of data

The need to control access to data introduces constraints in their management. The ethos in Europe is not in general one of data sharing and there is little recognition of the fact that data can

be viewed as a public good and that not to use data incurs expense (in terms of less efficient decision making). Indeed increasingly data is being treated as a commodity and sales are used to raise badly needed revenue. The sharing of information internationally can change the perspective of both data users and data providers.

One aspect of the need for control of data means that several different versions - say anonymised and unanonymised versions - have to be managed and good records kept to control distribution rights.

**Preservation of data**

It is vital that we preserve electronic information in a way that permits them to be useable over time. The Archive holds many unique copies of datasets which we must be able to read despite changes to hardware or software configurations or even more fundamental changes to the generation of computing technology. Because of the lack of facilities for electronic data preservation at the UK Public Record Office we have become the de facto official archive for many large Government datasets.

The Data Archive prides itself on being able to handle almost any format and media type for the deposit of data and, on the rare occasion that problems do occur, Archive staff will work with the depositors to reformat the data. The main software dependent formats of data received are SPSS, SAS, SIR, STATA, Paradox, Dataease, Microsoft Access, Microsoft Excel, Microsoft Word and WordPerfect. Many are in simple ASCII format. The Archive maintains a suite of conversion tools in order to move data to current preservation standards. Occasionally the Data Archive assists in projects to retrieve old electronic data. This is not always successful thus demonstrating the importance of archiving data correctly at the time they are produced.

When data are received in the Archive they are processed with an emphasis on ensuring that all files are present, readable

and, importantly, in accordance with the documentation. At this stage any confidentiality problems will be identified and clarified with the data owner. The structure of the data will be examined and data processing staff will ensure that the documentation reflects the structure correctly and fully. User services staff and those staff with responsibility for data delivery are informed of any possible problems which might be encountered in reformatting or delivering these data to users. Attention is also paid to the documentation of derived variables since these are often overlooked by those creating the documentation.

The data and documentation must be sufficient to enable a secondary user to understand and analyse the data, without any major discrepancies in the dataset. However, the data will not necessarily be perfectly clean and the Archive cannot take responsibility for the errors in the data. If problems cannot be reconciled at this stage or if they occur later - say, when a user identifies a discrepancy - then supplementary documentation will be attached to the dataset so that future users are alerted to the problem. User records ensure that those who have already received the faulty data can be contacted.

The Archive's preservation policy and practice ensures:

- the physical reliability of digital data
- the security of the data from unauthorised access
- the usability of the data
- the integration of the data, where appropriate, into - information and dissemination systems
- the maintenance of effective data documentation

The system is flexible in order to meet the demands of changing technology and in order to meet the evolving needs of the user community.

In terms of the **physical reliability** of data, the Data Archive has migrated data through several changes of computer systems. Four years ago, the data were moved from a ½ inch tape based system to one based on a jukebox of optical platters. At present,

the data are being migrated again onto DLT tapes to improve capacity and enable the integration into more advanced database systems for dissemination and searching. At the same time, data will be maintained on at least one, and normally two additional media. For example, the data are being stored on optical media (either CD-ROM or optical read write) and two types of cartridge backup. These copies are in physically separate places, with a copy being kept in a fire vault on another part of the campus. Additionally, one off-site copy is currently maintained in London.

In terms of the *security* of the data, the Data Archive has a strict policy of ensuring that only authorised users within the organisation can have direct access to the data. In addition, the data writing capacity is restricted to one carefully controlled account. The growth of networking means that external access is being further controlled and monitored by the installation of a firewall, using the latest and most secure technology.

The *usability* of the data is ensured by storing the data in ASCII as well as common formats. The ASCII version ensures that any future system will be able to read the data easily. It does, however, have implications for the functionality of the data, as some features of the data may be difficult to reconstruct from this version. In the absence of internationally agreed standards for data description it is important to maintain this lowest common denominator. Additional formats of the data are kept in cases where there is concern about the complex structure of the data and in cases where a large number of users will require the data in a common format, and this has been the policy in recent years. Such formats contain more detailed descriptions of the data and are stored in a system portable version wherever possible. These procedures ensure that the data will be easily portable across systems. Typically, these formats correspond to the format in which the data were received. However, this is not always possible, for example if the data were deposited in a rare or redundant format, and in these cases the data are integrated to

the nearest corresponding but widely used format. Thus the Archive's emphasis has been on standardisation and migration rather than emulation of past systems.

The physical evolution of the storage system which is in progress at present will enable the Data Archive to carry out a logical restructuring of the whole data collection. We have created a common system for the data in order to exploit automated management tools. For example, the various validation and testing programs developed by the Data Archive can be implemented more systematically and without the same need for manual intervention as a result, facilitating the *integration* of the data into information and dissemination systems.

It is essential too that the preservation of documentation is taken seriously. In order to improve the storage of documentation and its delivery to users, and to take advantage of new technological ways to integrate data and documentation, the strategy of the Archive is now to hold documentation in digital form. Where possible documentation is acquired in digital form - usually as a word-processed document - but when this is not possible documentation is being scanned. The Higher Education Funding Councils have provided funds to enable the scanning of paper documentation for the retrospective collection. Digital documentation is being stored in image format. Adobe Acrobat format is being investigated since it maintains the documentation structure and appearance. In addition, it is planned to use optical character recognition software to convert priority documentation to text-based versions. As for data files, conversion to other formats is kept under continual review and will be employed when it is deemed desirable.

## Challenges for the future

A number of challenges face us in the attempts to improve the service to users. Data are increasingly being deposited in a

software specific format which cannot be disentangled without destroying the usability of the data. On the other hand many users want the data in software specific format and thus there is a tension between the archival service and the provision of access to researchers today. Data are rarely static, even if they are not continually updated it is likely that usage will result in changes to them, and this raises problems as to what should be preserved. It can also raise problems of authentication - who is permitted to make changes?- and of version control.

It is going to be increasingly important to address issues of quality and the related problem of liability in delivering information. Since quality is defined as fitness for purpose and users by definition may be carrying out a wide range of different sorts of research, it is extremely difficult to know what is relevant to tell them about data quality and thus more research is needed into the needs of users for information about quality. In fact great improvements generally are needed in the form and content of data documentation. Data producers are the best people to create most of this but they need guidance and encouragement. The increased use of the Internet for data delivery provides challenges as to how we can integrate the data and documentation for ease of access, as well as exploiting the documentation as a resource in its own right to a greater extent.

Most of the data we hold at the Archive has already been anonymised before being deposited with us in order to protect the identity of individuals and thus maintain pledges of confidentiality. However as a result of this the full unanonymised versions are not being preserved for posterity. A further problem is that the procedures to anonymise data are not straightforward and different systems may be best for different usages.

**Improvements to our work**

I should not end on a pessimistic note about the challenges which face Data Archives. In fact there are a large number of improvements which will make our work much easier in the future. To name a few as examples:

- there are major improvements in the longevity and reliability of optical storage media.
- developments in optical storage and retrieval devices enable improved data management and access
- there are better interfaces between software and more use is being made of standards
- improved data encryption techniques are helping in data authentication
- I could have chosen many more examples of ways in which technological advances and sharing of information are aiding the work of preserving and distributing electronic data. Data archivists and librarians who manage electronic resources share many of the same problems and do have a great deal to learn from one another.