Archives and Dinosaurs

by Eric Tanenbaum[1]

Introduction

Dinosaurs and social data archives have a lot in common. When both began their existence they had their respective fields pretty much to themselves. Having almost exclusive control over their environment for a long period, dinosaurs and data archives both swept up material whenever possible and, in time, appeared cumbersome and bottom-heavy. From this state both had to confront a changing environment. However, for all the similarities between the two, dinosaurs differ from data archives in at least one important respect — they no longer exist. Thus while it is too late for dinosaurs to learn from data archives, archivists should consider the dinosaurs' progress if they wish to distance themselves from the dinosaurs' end. This note suggests how they might do so.

Palaeontologists may differ when assessing the relative weight of specific causes of the dinosaurs' demise, but there is common agreement that non-adaptation to changing climactic conditions is important in their undoing. In modern terms it could be said that dinosaurs were frozen out by a changing hardware environment. Archives also confront hardware changes, but their impact on archival work is confounded by concurrent software developments.

This paper describes major changes in several areas which affect computerized data archives. On the hardware side, the paper examines improvements in mass storage capacity and the ergonomics of computers (of all sizes). Software developments, in parallel with these hardware changes, encourage new orientations to social information. From among these the paper focuses on "new" database management techniques and electronic publishing — both have implications for archive growth. Changes in hardware and software are combined by improved communication facilities; the catalyst producing the "alloy" lies in the imagination of information analysts (archive users) whose expectations are aroused by these more elementary developments. The paper describes aspects of the agents of change which are germane to the future operation of archives. An integrated systematic approach to the tasks required ensure that future concludes this paper.

[1] Presented at IASSIST/IFDO International Conference May 1985, Amsterdam. This paper has been previously published in European political data newsletter no. 55:33-44, June 1985.

Mass storage devices

The history of computerized data archives for the social sciences illustrates the evolution of computerized mass storage devices. Cardboard computer cards, or "IBM cards" as they were commonly known, were an early *de facto* standard medium for data storage. The "data archive movement" of the early 1960's was launched when it was recognized that these cards could be banked centrally for subsequent redistribution to other sites which supported this physical standard.

Although computer cards were reproduced and shipped "by the forest", the medium was not ideal. It is clumsy — cards get dropped, insecure — cards get torn, and expensive — bulk reproduction is a resource intensive activity. It also limited the analyst's access to large volumes of information. Clearly faster forms of "data memory" would yield vast improvements in the kind of service that archives could provide researchers.

The magnetic computer tape offered the medium of distribution that data repositories required. It is not as universal as computer cards, for each brand of computer uses a different mode of tape storage. However, almost all archives have computer software that allows them to read and write tapes written in all formats used in their user constituency. Thus, for example, the British ESRC Data Archive maintains a suite of conversion routines that permits it to transform data from its own in-house standard to <u>any</u> form required by British users.[2]

While magnetic computer tapes gave archives a cheap medium for transmitting subsets of their holdings to analysts working at remote sites, the medium constrains the kind of material that can be accessed. First, in almost all cases, it requires that information be stored as sequential files. This immediately limits the scope of data that can be transmitted to a few discrete chunks, if only because of the effort and skill required to reassemble anything more ambitious at the receiver's end. Second, the medium itself has a small finite capacity. Granted the volume of data that can be stored on magnetic tapes has increased dramatically from the 6.4 megabytes feasible with the earliest tapes to a current 210 megabytes,[3] but still requires six physical tapes to hold the results of the 1981 British population censuses <u>after</u> the data have been subjected to complex compression routine. Operationally, this means that the analyst who wants to select census data from points across the nation is involved in considerable tape manipulation. Third, and finally, tapes, which are volatile, offer poor archival security. Ensuring the physical integrity of a tape-resident database is a labour and time intensive task which a central facility can perform because it can take advantage of economics of scale but which an individual would find restrictive.

For these reasons archivists should welcome the recent emergence of new modes of mass data storage, two of which will be described here as a prelude to a later discussion of how they <u>should</u> be incorporated into archival operations.

Several manufacturers have announced the development of disks that use laser techniques

[2] A side benefit of this mode of operation, initially designed to cope with the inelegancies of computer manufacturers' whims about tape standards, is that central archives have protected their, and thus their constituency's, data resources by creating a protective buffer between a single in-house standard to which all data are converted and changing external requirements. Thus, when external technological changes occur the entire database can be transformed to the new requirement by a single routine operation which "maps" the old format to the new.

[3] The comparison is between a 2400' reel recorded at 200 bits per inch ("bpi") and one recorded at 6250 bpi.

to input and output information at extremely high densities onto small robust platters. Thus, for example, one firm's first release promises the storage of one gigabyte (i.e. 1,000,000,000 characters) on a single side of one physical disk. Using the British population census again as an illustration, it ought to be possible to store the entire set of counts on a single disk.

While at their initial release the disks, which cost about £200.00, are somewhat more expensive than the conventional computer tapes required to store a similar amount of information, the radical impact of these new devices will come both because they allow non-sequential access to data and because they are of archival quality, offering a minimum of ten years' secure storage. Data analysts can now realistically contemplate linking large volumes of information from diverse sources in their pursuit of new connections between and among social phenomena.

In response to this facility, archives have to reconsider how they service their constituency. Eventually, archives will have to meet the needs of analysts who have access to mass storage devices by supplying mass data packages. These will likely be based on diverse data sources which might in turn be linked by "discrete", but otherwise broad, "story lines".[4] This orientation to data, for which the Italian and Norwegian data archives' work constructing ecological databases is a precedent, will have to be extended to many areas of social inquiry and will conceivably require a more active intervention in the work of data archives by subject specialists acting in an editorial capacity.

Optical disks, because of their robustness and cheapness, are amenable to distribution in much the same way as traditional magnetic tapes are. Their local use (by independent analysts) is feasible, as the manufacturers of optical disk drives generally use a standard "interface" between computer and drive. Thus, unlike tape equipment, it is possible that this kind of mass storage will soon be available even for desktop "personal" computers.

However, optical disks have value to the archives' own computer installations. For the British Data Archive, it is estimated that over 80% of its files are sufficiently stable[5] to make it sensible to transfer the bulk of its holdings to these devices. This would have the immediate advantage of simplifying internal operating procedures, even if the Data Archive continues to supply most of its users with copies of data files for access on their local machines. However, if one considers another development in technology, the "networking" of computers which permit individuals to address many computers directly from a single site, these storage devices assume a higher profile in the archives' future landscape, because, with a conceptually, if not technically, "simple" modification, they offer an almost limitless volume of fast access data retrieval.

Physically, optical disks resemble long playing gramophone records. Thus, as with gramophone records, these disks can be stored in a machine similar to a "juke box" whereby a would be listener (analyst) can choose any song (data file) that is available within its confines. No human intervention, other than by the "listener", is required. The songs are permanently on-line.

Suggesting a machine that would keep the "Top 40" data files readily accessible to analysts is

[4] In fact, this is analogous to the approach taken by the British Broadcasting Corporation's Domesday Project, which was described elsewhere during the conference and with which the British ESRC Data Archive is collaborating.

[5] The first optical disks on the market offer a "write once, read many times" facility. Thus, for the moment at least, they are best considered devices for storing stable data. Of course, from an archival perspective, the data security offered by a non-erasable device is a bonus to the mass storage capacity.

not fanciful. In fact, at least one optical disk developer (Philips) supplies a "carousel" option for its "Megadoc" system. Although the system is initially directed to the storage of document images, there appears to be no reason it could not be adapted to numerical data bases.

The juke-box approach to data storage is shared with another recently released mass volume device which is based on densely packed cassette-like tape cartridges. Although these are not transportable in the way that optical disks are, they offer much more storage potential and have to be considered a likely enhancement to the hardware offered by a data library service which wishes to support direct access to its holdings by analysts.

As mentioned, the impact of improved inter-computer communication facilities on archiving is considered later in this paper. For now, it is sufficient to note that the potential for "on-line" access to masses of data which is made possible by the two devices just described will encourage social researchers to explore the use of the developing "network" capability, particularly as improved storage capacity is interacting with a radical change in the overall provision of computers themselves. A brief description of the "new ergonomics" of computer use is a useful prelude to a discussion of the effect of networks on archives.

Computers: a changing style

Traditionally, social science data archives could assume, reasonably, that their catchment area comprised all computer using social researchers. As computer use in social science was intricately linked to a quantitative orientation, computer users were numerate and usually shared a kit of tools that were applied to research tasks. Moreover, the "conventional" computer oriented social investigator, who was most attracted to the "calculative" power offered by computers, was adequately served by the existing provision of computers in research environments. The computer, physically located in a central position in the institution, was fed numeric data, manipulated them, and then supplied the results of the manipulation. Data archives, which were also centrally located, were well-suited to this mode of computer access and in most countries developed strong institutional ties with the providers of computer services used by the research community. In this way, archives could minimize the technical barriers which inhibited researchers' access to their holdings.

The recent growth of desktop computers, cheap enough to be purchased by individuals, threatens the homogeneity of the computer-using community. A cursory glance at the "micro-computer" marketplace suffices to show that the main appeal of these machines is not that they are superior calculators but that they are remarkably sophisticated typewriters which manage to combine a keyboard, an electronic scissors and a truly non-spill gluepot.

More important, though, for archival development, these desktop machines are changing the prevailing view of what constitutes "machine-readable" data. It does not take long with a "word-processor" to recognize that semi-structured textual information often is more easily organized, manipulated and analysed with the help of a computer than it is manually.

Not surprisingly, facilities offered by desktop workstations are also changing the "traditional" computer analyst's orientation to computerized functions. Granted, quantitative analyses of large data sources are still best done by large, central "mainframe" computers, but the post-processing of the results for research reports is now best accomplished with the software (and sometime hardware) facilities offered on micro-computers. Thus, for example, the survey researcher will continue to manipulate the survey's data with a large computer to produce summary information. These results will be captured on the machine on which the report itself is composed. While this might be only to avoid re-typing tables or matrices, the analyst will likely also wish to apply micro-computer facilities like graph processors or spread-sheets to the reduced dataset for further refinement.

Both of these instances of expanded computerization demonstrate an increasing integration of information processing which replaces the earlier compartmentalization of computers by specific tasks. Clearly, computer use is no longer the exclusive prerogative of the specialist in quantitative techniques. This has profound implications for data archives.

First, archives are bound to encounter a new community of users who regard them (archives) as just another source of information. These researchers will have been in contact with electronic publications of other types (e.g. bibliographic search services or reference publications) and will not immediately consider the traditional data archive as being in any way different. Nor should they. It would be odd if the oldest purveyors of computerized information could not service the needs of the newest seekers of that kind of material.

Of course, data archives do not generally provide the summary (digested) style of information that most reference seekers want. However, often that information is available to archives who, however, reject it because it is not their normal stock in trade. In the future, if archives are to serve this new market (which will include a significant proportion of their older market) they will have to make this kind of information accessible.

Second, the integration of information handling practised by the archives' traditional users will affect what these users expect archives to provide. They too will be less inclined to halt their work progress to permit conventional archive practices to work. They will demand more immediate access to these services, requiring that the archives' input to their information needs be much more transparent. Archives will (and likely should) become visible only when (infrequent) hitches develop which require intervention.

These developing expectations can be traced to the influence of desktop workstations. However their fulfillment can only be realized by archives if the archives have access to the large scale storage capacity described earlier and if the archives can offer access to these facilities to remote users. Communication networks offer the link.

Computer networks

When computers first became available to researchers employed in the British academic sector, they were provided by individual institutions. In time, an informally organized system of resource sharing developed, wherein some institutions assumed a responsibility to service some of the larger needs of institutions in a particular region. By the end of the 1970's it was likely that a computer user in any given institution would have access to a larger regional centre as well as to a local center. Indeed, communication with the larger machine was often via the smaller local computer.

Although this arrangement allowed researchers to use much more powerful installations than their own institutions could afford to provide had they stayed totally independent, they still offered a limited access route to the country's entire community of computers. For the British Data Archive this meant that there was little need to develop a facility that enabled direct enquiries to its holdings by external users — most users continued to depend on a magnetic tape based service and the Archive was best advised to devote its efforts to improving that mode of data dissemination.

Recent developments in data communication have changed this aspect of the computer user's working environment. In many countries, most computers are now functionally no further away from the user than the nearest keyboard. In the British academic sector, for example, the Computer Board–sponsored Joint Academic Network ("JANET") offers researchers in that sector an appropriate inter–computer link which facilitates communciations among university computers in Great Britain (as well as with other "networks" in Great Britain and abroad).

JANET, as a communication path, meets the need for a facility that allows one computer to talk to many computers. Its more important contribution, however, is to hide the intricacies of network use behind a facade which makes it simple for computer users to address multiple computers with little more knowledge than that required to use their local computers. It accomplishes this with "protocols" which standardize message transmission between sites. At the level of the network, messages may be commands to join a remote computer site as a "local" user or to retrieve information to the user's own site.

It does not take too much imagination to see what the effect of this new facility for communication with remote sites might be on the operating procedures of data archives. At the least, many analysts will want to interrogate a catalogue of archival holdings to determine which, if any, files contain information of interest to their projects. However, having located pertinent data sources many will wish to select only those parts that are relevant. They may then be happy to "download" the data to their own installations for analysis but, in theory, they could just as easily (or perhaps even with greater ease) analyse the data at the archive's site, particularly if the archive had implemented specialized software tools to facilitate secondary analysis.

The last paragraph contains an implicit research agenda of projects that are necessary to build the interface for on–line access to archival holdings. An integrated approach to these essential tasks is mentioned in the conclusion to this paper and so the individual tasks need not be dwelt on here. However, at this point it is appropriate to note that the tasks that face an archive also confront any computerized information utility that wishes to encourage direct access to its wares. As these utilities increase in number and as demand for them grows pressure will develop for a coordinated approach to information management on a national (and possibly international) scale which will transcend particular subject orientations. Data archives will have to join these integrated systems — they should be in the forefront of developments. In any event, whatever their institutional inclinations two related features of the environment in which archives now work, the demand for multiple sources and the acceptance of the "relational model" of data management, will push archives in this direction.

Multiple sources

For years, advocates of secondary analysis as a research strategy for the social sciences, and thus supporters of social data archives, have argued that only this form of research allowed linkage of diverse data sources which was necessary to fully explore social phenomena. However, the records of data archives' use patterns suggest that these multiple linkages are rarely made — the majority of secondary analysis seem to be of single data sets.

There are several reasons for this. The first might relate to the difficulties entailed in merging large masses of data supplied on magnetic computer tape. As described earlier in this paper, the user of archival material often had to request much more than was needed, largely because there were no facilities available for obtaining the subsets that were really required. At this level, it could be that the potential of direct user-archive access will be sufficient to encourage users to pre-process archival material before analysing it.

However the availability of many different kinds of information which are not only in computerized form but which, often, are in only machine readable form will force, or at least teach, researchers to address multiple sources of information. At the onset many of these will be "reference" works which are of interest because they yield independent "facts". However as experience is gained in locating "facts" from several (many?) places and retrieving them for assembly on a single computer, researchers' perspectives will become more ambitous. They will begin to want the same capability of addressing multiple numeric data files which they will tailor to a form which is adequate for reassembly into a purpose-built whole.

Besides the potential for network access and the increasing availability of multiple sources of computerized information, one more development on the computer landscape will have a great impact on archive users' expectations of the type of service that an information utility should provide. Fortuitously, this development, the widespread acceptance of a relational model of data management, also provides users with the tool necessary to take advantage of multiple sources addressed on-line.

The relational data mode

Further commentators on the penetration of computers into "everyday" life during the 1970's and 1980's will highlight the provision of "easy" database management techniques which permit researchers to take multiple logical perspectives of particular group of data. Although several different modelling strategies are available, the "relational" approach to database management offers the most exciting and attractive prospect for social researchers because, among all the alternatives, the relational model most closely replicates the way analysts think about data. It thus offers a tool for analysts interested in analysing substantive problems rather than itself becoming the goal for which analysts strive.

While it is not possible to delve into the details of the model here, it is worth noting that the model's strategy of simplifying the association between discrete sets of data supports the exploitation of multiple data sources when analysing a phenomenon. Most importantly, from an archive's viewpoint at least, it suggests that only those data that are required must be retained when assembling a file for analysis. As suggested earlier, this runs counter to the conventional archive practice of "user takes all", with its demand that the analyst cope with a massive body of unnecessary data.

Strangely, given the esoteric nature of database management, this is the change which could have the greatest single impact on the demands put to archives in the future. The elegance of the relational approach to data management has attracted many micro-computer program developers. Consequently, social researchers who were first introduced to computers via these machines will have experienced "quasi-"relational management systems and will have grown accustomed to applying their power. Moreover, as micros have (until recently) offered only limited data storage capacity, these new computer users will have learned to work within the confines of these machines. They will not appreciate that moving to larger machines, as they will do when accessing central information utilities, permits a more relaxed view of data storage.

As these new computer users represent the "growth" potential for data archives, their influence on archival develoment cannot be ignored. Thus it is appropriate that a description of the impact of "new technology" on archives conclude with a speculative note on the most powerful driving force for change, the new user community.

Changing people

It will be evident from the remarks earlier in this essay about prevailing archival practice that users of social science archives almost invariably came from a small segment of the social science community. Oriented toward "quantitative" social research, they grew up with archives and, like archives, learned to accept — and perhaps even like — the "user hostile" environment in which computer users were expected to work. The ethos of computer use has changed and new entrants will be unware of the need for a hairshirt. Archivists, who tend to be of the old school, will have to adjust their expectations of users to correspond to what their expanded catchment area expects of them.

This new generation of computer users will treat computers with the same ease as they did typewriters a decade ago. For them, the computer is a general utility for a wide range of tasks, among which is information gathering. People accustomed to interrogating a bank account on line or ordering furniture from a direct access shop, will not consider that assembling cross-national data on the association between class and political participation is sufficiently special to warrant the cumbersome hurdles that now impede access to archival data. The archivist must be aware that barriers which reflect past contingencies will direct a major portion of their user community to other information services which offer more flexible access to social information.

Having said this, it must be recognized that the transition from dinosaur to butterfly will not be an easy metamorphosis. One feasible route toward the changeover is described as a conclusion to this paper.

From dinosaur to butterfly: an easier metamorphosis

There is a danger that the earlier discussion which related technological developments and current practice will leave the mistaken impression that archives are unresponsive to change. In practice, archives have worked to incorporate most technological advances into their operating procedures. In the area of networking, one could cite the EEC-sponsored ACCESS project which is designed to produce a cross-national, integrated, bibliographic, on-line data base. The longstanding development of the CESSDA Study Description Scheme fosters the bibliographic control crucial to the identifying sources of comparable data. There have been many examples of archival use of centralized mass storage facilities — for example, the British Data Archive's distributed arrangements for the supply of the 1981 population census.

However, for all these individual projects, the breakthrough to a comprehensive information service still seems a remote prospect. Although part of the problem is related to archival practices, a significant share of the difficulties are attributable to more general features of computer use. As these affect all information providers, the removal of these encumbrances on efficient information access requires the development of an integrated system. It is to this joint effort that archives should devote their resources.

In most countries, the computer user can give an empathetic hearing to the tale of the Tower of Babel. While communication utilities like JANET mask the intricacies of making connections between computers, they do little to improve users' access to different computer systems. In effect, the computer network gets the user to the computer's door but, in most cases, that door is locked against the user's entry unless the user possesses privileged knowledge, to say nothing of privileges.

Prevailing computer practices, which reflect a period when each institution offered its own computer power and each computer manufacturer devised its own operating system, throw up the greatest barrier to a "butterfly-like" access to information. Until this artificial restriction on computer use is overcome, archives and users will be forced to work in an environment in which flexible approaches to information sources are blocked.

However, the obstacle could be removed with a central computer-based facility, accessible to all by network communications, which shields users from different computer environments and protects the environments from many different users. This facility would offer a classified catalogue of all available information sources in the United Kingdom which contained information about the substance of each source and technical information about access arrangements. More importantly, the user would only use the database for subject searches — the technical information, which would be kept transparent to the user, would be used to "automatically" invoke the dialogue necessary to access the host information sites.

Social data archives should be promoting the development of a utility like this. It requires more than a unilateral venture from any single sector and demands more resources than archives themselves can expend. Social data archives, nonetheless, have a privileged role among information providers for they were among the earliest to be computerized. Thus they offer a rare perspective from which to view the changes described in this paper to those with whom they might cooperate.

A central utility like this would benefit social researchers because the only "new" specialist skill required relates to the bibliographic search procedure, which would be common to all sources. It would reward social science because it would allow the exploitation of technological advances which would otherwise be barred to it. It would be attractive to social information providers because they could work to one common standard. It should appeal to current data archives because it promises to provide the protection against the technological "chill" that spelt the dinosaurs' demise.■