Notes and Communications

It's Public Knowledge: The National Digital Archive of Datasets*

PATRICIA SLEEMAN

RÉSUMÉ Cet article décrit l'histoire et le développement du *National Digital Archive of Datasets*, un service offert par le centre informatique de l'Université de Londres pour les Archives nationales de l'Angleterre. L'auteure présente le contexte dans lequel le projet a émergé dans les années 1990, son approche qui diffère de celle des archives de données informatiques traditionnelles, ainsi que la gamme de ses fonctions archivistiques. Finalement, elle offre des réflexions sur le projet dans son ensemble.

ABSTRACT This article describes the history and development of the National Digital Archive of Datasets, a service run by the University of London Computer Centre for the National Archives of England. It discusses the project in light of the context in which it emerged in the 1990s, its departure in approach from traditional data archives, and the range of archival functions. Finally, it offers reflections on the project as whole.

Introduction

In 2001 I escaped the world of data archives and moved to Egyptology while on secondment at the Flinders Petrie Museum in University College London. Their collection held a Ptolemaic census, written in demotic, unearthed by the archaeologist Flinders Petrie in Rifeh in the early twentieth century.[1] It was similar to modern-day census data, tabular in structure with data divided into columns and equally as indecipherable to the naked eye. This ancient example of the gathering of structured data demonstrated to me how long such resources have existed. These are powerful records providing us with a glimpse of a population, or even a household, at the time of the Ptolemies. The

* I wish to make it clear that the views expressed here are my own. I do not speak for any institution. I would like to thank all my colleagues at the National Digital Archive of Datasets for their help and assistance, namely Kevin Ashley, Head of the Digital Archive Department, University of London Computer Centre, and Sally Hughes, Senior Data Specialist. I also wish to thank Terry Cook, Raivo Ruusalepp, and Kate Manning.

1 A language used for a thousand years from the 7th century BC until the 3rd century of our era. Occasional graffiti occur even later.

quality and quantity of such information in a non-tabulated form would be cumbersome and difficult to assimilate.[2] While databases are seen as relatively new forms of records, the essential concept of structured information gathering has existed for thousands of years.

Structured data in electronic format has existed for at least fifty years.[3] The early lack of pre-emptive action in relation to the preservation of the data and its metadata, and the absence of migration programs, has led to its loss, even where active preservation programs are in place.[4] This is no longer an option. National governments and international organizations have been among the first to adopt electronic media for their transactions. Statutory obligations require that they retain these electronic records in accessible form. To do this, pre-emptive action is needed.

The National Digital Archive of Datasets

One of the pioneers in the area of digital preservation is the National Archives (TNA),[5] which in the mid-1990s developed two programs for the preservation of electronic records. The first was the in-house Electronic Records in Office Systems Program, the EROS Program for short. The aim of EROS was to guarantee that the records of long-term value created in office systems introduced by government bodies were available for future access. The second program was the Computer Readable Data Archive (CRDA). This was later renamed the National Digital Archive of Datasets (NDAD), which complemented EROS with its remit to preserve datasets and other forms of structured data which have been used in government departments for some years.[6]

2 Similarly, maps have the capacity to describe spatial and quantitative information.

3 National Archives and Records Administration (hereafter NARA) has data from the 1950s. See Record Group 15: Records of the Veterans Administration; Series: Repatriated Korean Conflict Prisoners of War, 7/5/1950–10/6/1954. See <http://www.archives.gov/>, visited 23 September 2003.

4 An immediate and dramatic example of the repercussions that inaction in the sphere of digital preservation can have is seen in the area of linguistics. Computer scientist and linguist Professor Steven Bird of Melbourne University has noted that linguists are worried because they have been enthusiastic digital pioneers. Attracted by ever smaller, lighter equipment and vastly improved storage capacity, field researchers have graduated from handwritten notes and wire recordings to laptops, mini-discs, DAT tape, and MP3. Now these technologies used to attempt to preserve information on disappearing languages are themselves at risk. "We are sitting between the onset of the digital era and the mass extinction of the world's languages," said Professor Bird. "The window of opportunity is small and shutting fast." Taken from the BBC News Web site at <http://news.bbc.co.uk/>, visited 20 March 2003.

5 The National Archives (TNA), launched in April 2003, brought together two existing organizations, the Public Record Office and the Historical Manuscripts Commission.

6 A dataset is a computer file or related set of computer files, forming part of NDAD, which is organized under a single descriptive title and is capable of being described as a unit in the

NDAD's other remit is to describe, and where possible, provide access to its holdings. While NDAD operates at arm's length from TNA, the selection (appraisal) of records and specifications for levels of processing and services are set by TNA.

Background

Aware of the number of legacy systems in government departments spanning over thirty years, TNA realized early on that the preservation of databases needed action on their part, despite arguments by people such as David Bearman that the vast majority of "second generation"[7] computer systems in government and corporations did not produce the evidence, the reliable documentation of business transactions, needed as proof of activity, rights, commitments, or entitlements.[8] The preservation of office-type documents was a relatively new issue at the time whilst the TNA saw the preservation of databases of urgent concern. As part of a pilot project conducted between 1991 and 1993, TNA undertook a survey of electronic holdings of government departments. After examining the results of this survey, TNA commissioned independent consultants to investigate the cost implications of establishing an archiving facility for machine-readable public records.

The question had already arisen as to whether TNA should use an agent for archiving structured, computer-readable records, in the same way as it had, for some years, used the National Film Archive (now the National Film and Television Archive) to store and make available public records which take the form of moving images (i.e., films and videos).[9] It was recognized that the

finding aids. It may comprise one or more accessions. A database is a collection of information, usually covering subject areas which are related in some way, structured to enable effective retrieval of the information. Databases are organized into a hierarchy of files, records, and fields. A file is a group of related information, such as names and addresses of members of a sports club. All the information about a particular member (name, address, etc.) is stored in a record. A record is a collection of related data items called fields (an example of a field would be a member's name). Definitions are taken from the NDAD online glossary at: <http://ndad.ulcc.ac.uk/help/glossary.htm>, visited 23 September 2003.

7 See Terry Cook, "Easy to Byte, Harder to Chew: The Second Generation of Electronic Records Archives," *Archivaria* 33 (Winter 1991–92), pp. 202–16.

8 These, in Bearman's view, generate data or information instead of transactional "records" and are banished as non-archival and unworthy of the archivist's attention. For more regarding this definition of the record see <http://www.asis.org/Bulletin/Feb-98/bearman. html>.

9 Two schools of thought emerged in the 1990s in response to the preservation of electronic records. These could be defined as custodial and postcustodial. The custodial approach takes Hilary Jenkinson's dictums regarding physical custody and guardianship literally, i.e., unless a record has crossed the "archival threshold" and is in full custody of the archives and archivists, it cannot be an authentic and reliable record. Those repositories which cannot, in terms

level of skilled resources required to preserve and provide public access to these datasets was greater than that then available at TNA, and that specialist expertise and facilities outside TNA would be likely to offer a cost-effective alternative.[10] The consultants' report confirmed this view and provided costing estimates in support of these conclusions. TNA decided, therefore, to place a contract for the establishment of a computer-readable data archive for these datasets.

A large-scale Private Finance Initiative (PFI) procurement exercise was conducted which led to the award of a contract to the University of London Computer Centre (ULCC) to manage the overall service, house the staff, and provide computing resources, with the collaboration of the University of London Library (ULL). ULCC had previous experience with large-scale storage of data, and had met the challenge of converting data to different formats and of providing user access to data. However as well as storage and the provision of access, TNA required the datasets to be described according to international archival standards. ULCC had no expertise in this field and therefore joined with the ULL, which had experience in describing archives and manuscripts and in providing access to electronic sources. Together they made a joint and

---

of resources and technical capability, guarantee that they can safeguard the record in terms of adequate preservation strategy look towards the "non-custodial," "postcustodial," or "distributed management" approach. This approach sees the record left in the custody of their creators, monitored, and policed by the archives in terms of non-alteration, migration, and access. TNA did neither, but rather met both schools of thought halfway. It maintained rigorous archival control in the form of appraisal decisions as well as determined levels of processing and services, while custody of the data as well as the extensive technical processing, migration, and distribution was contracted out to NDAD. See Luciana Duranti, "Archives as a Place," *Archives and Manuscripts* 24, no. 2 (November 1996), p. 252. For further definitions of post-Jenkinsonian concepts see the UBC-MAS project Web site, <http://www.interpares.org/UBCProject/>, visited 23 September 2003. See also Luciana Duranti and Heather MacNeil, "The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project," *Archivaria* 42 (Fall 1996), pp. 45–67.

10 In 1995, both the Finnish and the Icelandic national archives contracted out the function of physical preservation of electronic archives to computer centres. However, it proved unsatisfactory as preservation criteria were met only from a technical and not from an archival point of view. The Finnish Computer Centre was capable of preserving (using migration) records in electronic format in a manner that safeguarded their legal and evidential value. However, researcher access did not compare well with that for hard-copy archives. There were also problems in ensuring that well organized system documentation and inventories were prepared. Thus in 1996 the National Archives of Finland and the National Archives of Iceland assumed direct responsibility for all aspects of the preservation of archives. See Matti Pulkkinen with Tom Quinlan, "Nordic Archives and Electronic Records: Preservation of Electronic Records in Nordic Countries," in Rena Lohan, Mark Conrad, Ken Hannigan, and John A. Jackson, eds., *For the Record: Data Archives, Electronic Records, Access to Information and the Needs of the Research Community* (Dublin, 1996), p. 49.

successful bid. Then began a seven-month period of negotiation, implementation, and testing.

On 24 March 1998 the services and systems put in place by ULCC and ULL for the NDAD project were formally accepted by TNA, bringing to a close an intense period of development and testing. ULCC decided subsequently that certain issues needed adjustment. These included improved navigation of the Web site, modifications to the data browsing interface, review of the transfer process, as well as increased provision of on-line help pages.[11] The service was re-launched in July 1998 during the UK presidency of the European Union at a conference at TNA. This heralded the start of what NDAD hoped would prove to be a valuable and popular service for all those with an interest in digital records.

The confirmation of the new service took place around the same time as the introduction of TNA's Web site. This was already showcasing examples of its work on the AD2001 project, a broad program for the successful implementation by 2001 of electronic delivery of a full range of user services. This and other developments in the work of archives, offering on-line catalogues of traditional archive material and archive material itself, were illustrative of the rapid changes taking place in the way archivists were providing access to the material in their custody. The services provided by NDAD brought together work in three of these areas: on-line searchable catalogues, preservation of born-digital material, and on-line access to archival sources.

Management and Contract

The contract was awarded for seven years with an option to extend it to ten years. The first review followed the two-week trial period in March 1998. Throughout the contract, quarterly and annual review meetings were planned as well as a five-year overall review. NDAD was placed initially within the Records Management Department of TNA, with its own contract manager. The NDAD contract is now managed by the Digital Preservation Department of TNA, which was set up in 2001 to be responsible for the preservation of all born-digital government data.

Nothing New?

Most of the services offered by NDAD, and the technologies behind them, were already on offer in one form or another elsewhere in the UK and the world. On-line access to data is a service which has been offered by many aca-

11 Kevin Ashley and Ruth Vyse, "Final Service Report" (UK National Digital Archive of Datasets, 20 March 1998), p. 7.

demic computing services for a number of years. ULCC began to do this almost thirty years ago, with key social science resources such as the Census Small Area Statistics and the Family Expenditure Survey, accessed through statistical packages such as SPSS and SAS. Libraries in the academic and public sectors have provided on-line catalogues of material for some years. Similarly, on-line archival catalogues are rapidly being made available. The preservation of digital material has been a concern of many groups for almost as long as digital material has existed. Much activity has taken place within industry (the seismic and pharmaceutical industries are particularly active in this area) and within specific scientific disciplines which involve large amounts of unique data, including pictures from weather satellites, or data which is expensive to recreate, such as that from atomic and nuclear physics. In almost all cases, these data sources are being preserved by the bodies which created and continue to use them. A smaller number of centres, of which an outstanding example is the UK Data Archive at the University of Essex, have focussed on collecting information from a variety of sources to form archives of interest to various research groups.

What we believed was new about NDAD was (and remains) the combination of elements in a single collaborative service. We preserve and catalogue acquisitions according to archival standards, provide on-line access to the catalogues, and link these directly to the preserved information itself. The infrastructure delivering the on-line service has been sized so as to be able to handle individual datasets of many gigabytes, and a total data capacity of many terabytes. In preserving the databases and associated material which arrives with them, we must focus not only on the short-term needs of the academic researcher of today, but on the wider needs of others who may want access today, and more particularly on the needs of future generations. Material which we accession is not being preserved for seven, thirty, or even one hundred years, but for an indefinite period of time.

NDAD Staff

NDAD consists of twelve people from four different disciplines. The ability to deliver the service as a whole has depended on bringing together these four sets of skills at ULCC. These include:

- Project archivists: They are the key players in making decisions about how to organize the holdings of the archive, and how to catalogue and index those holdings so as to make them accessible and comprehensible. They guide the decisions of the various computing specialists in data processing so as to ensure the archival integrity of the holdings.
- Archive assistants: They provide backup to the archivists in the organization and preservation of material, which includes the digital scanning of

paper documents received. They also provide general administrative support to the project and co-ordinate the provision of help and advice to the service's users. Archive assistants are often the first point of contact for users.

- Data specialists: They act as the interface between the archival and computing worlds. Each has a background in areas related to database design and use, statistical analysis of tabular data, or user interfaces to on-line data and catalogues. They ensure that the data received from departments is what it purports to be, produce metadata descriptions of the data, convert databases into a form suitable for long-term preservation and access, and validate the converted data both to ensure that its description is correct and that the conversion process has preserved its integrity. Most are also involved to some extent in the production of the Web site and some of the software tools used on it.
- Systems support staff: They are responsible for the construction and smooth operation of the various software and hardware systems on which the whole service runs. This includes the tape robot in which all the data and scanned documentation is ultimately stored, the hierarchical storage management system which keeps track of the data within it, and the Web servers which deliver Web pages to the end user.

The Archival Material and the Transfer Process

The material with which NDAD is dealing can be seen in the most general terms as information represented as a set of tables, in which columns contain a particular data item, and rows identify subjects for which data items are recorded. Although the elements in these tables are typically numbers or simple text, they may be pictures or even more complex multimedia items such as sound, documents, video, or digital maps. The key attribute which causes a particular set of computer records to be selected for preservation in NDAD is the tabular nature of the data, and its ability to be processed or analyzed in some way by a computer system. NDAD also contains some digital documents, and digital forms of paper documents. There is often a close link between a database and a number of other documents: these are strongly associated with the datasets and key to their full understanding. They form part of the archive with the datasets proper.

Selection

Although NDAD works on the preservation and provision of access to datasets, it is not involved in the initial selection of data, which is carried out in the same fashion as with other UK public records. TNA provides guidance to government departments on selection decisions and it has personnel

that supervises selection and performs records management work. Staff at NDAD may discover potential datasets which may be considered by TNA for transfer during the course of their work on another dataset, but TNA always decides on selection.

TNA's guidance on selection had not until recently specified the values which underlie its appraisal policy. Operational Selection Policies (OSPs) now apply the criteria set out in TNA's Acquisition Policy to the records of individual departments and agencies or to records relating to a cross-departmental theme. OSPs are normally the subject of public consultation before they are finalized. TNA is at present undergoing a review of the system[12] it uses for the selection of records in which it is developing four strands of work: pilot projects testing new methods of appraisal; research on macro-appraisal; the study of the particular issue of case files; and participation in the work done by the Electronic Records Management (ERM) Development Unit at TNA.[13]

Although NDAD does not carry out appraisal decisions, it is important we understand the reasons for selection. An essential part of our description of the datasets involves their role in relation to the processes and decisions they support. As a result, to ensure adequate description, we need to document the dataset's aim and purpose, thus reflecting the reasons for which it was selected. These systems are, of course, inextricably linked to the functions of the department. It is also important to note any changes to a database. An addition of an extra table or field within a series may well reflect an additional or slight change in the role of the dataset in the creating department. In addition, the complete review of a system may result in a new series and may reflect change in the department's role. This seems to be a type of functional analysis of the data in reverse, from the bottom up. These databases can be seen to be key to the running and decision-making of departments. One of the biggest problems associated with electronic records and their appraisal is that archivists no longer have the luxury of waiting for thirty years to make appraisal decisions. Selection has to be made very near to, if not at, the time the record is created. Identifying the informational value of a record is not always a problem; however, the sheer quantity of records produced may mean that we can no longer micro-appraise, or at least that we cannot rely on this alone. Nevertheless, at the macro-level, hundreds of databases may exist in a department – how to choose?

12 The "Grigg" System has two main elements: a system of timing and procedures recommended by Sir James Grigg in the report of the Royal Commission on Departmental Records, which was the basis of the Public Records Act of 1958 and then implemented by TNA and departments; and the advice and guidance given by TNA on how reviewers assess the value of records.

13 "Reviewing the Grigg System," *Records Management News (PRO)* (March 2003), p. 6.

Typology

Typologies can assist in the process of appraisal. The following have been developed by Shepherd and Smith as well as by Niklaus Bütikofer:[14]

1. Datasets with additions only include the:
1.1 "One-off" dataset, such as a survey or project-related data, where the boundaries of the project and data are clear; and the
1.2 Inactive or static dataset, where the data is added to and held in closed database files and will not be added to after closure.

2. Datasets with amendments ("dynamic") include:
2.1 Active or static data which can be updated and overwritten; and
2.2 Active data which is continually added to, but new data sits alongside older data.

An example of type 1.1 held by NDAD is the one-off AIDS Advertising Evaluation Survey.[15] It is marked by a beginning, middle, and end to the structure and time period of the data gathering and input. Once the survey was complete the dataset was closed. No more data was added. An example of type 1.2 is a dataset of immigrants, whose individual data records are closed once they have been granted citizenship.

Type 2.1 is demonstrated in a banking system, whose records provide details of personal accounts and where data is overwritten with each new transaction. Traditionally, a record of the system would be kept by creating regular snapshots, capturing data content at any given time. NDAD has not accessioned any "pure" examples of these datasets.[16] An example of type 2.2 of our typology is an active database system consisting of cumulative longitudinal data in scientific, medical, or environmental applications. The Anatomy Dataset held by NDAD, for instance, is an electronic version of hard-copy registers, files, and reports similar to those produced by Her Majesty's Inspector of Anatomy since 1832. The database was introduced in 1992, running parallel with paper registers until 1995.[17]

14 Elizabeth Shepherd and Charlotte Smith conducted a study of the application of ISAD(G) to datasets held by the National Digital Archive of Datasets in 1999. See E. Shepherd and C. Smith, "The Application of ISAD(G) to the Description of Archival Datasets," *Journal of the Society of Archivists* 21, no. 1 (2000), pp. 55–86. Niklaus Bütikofer presented a paper on this topic at the ERPANET Workshop on "Long-term Preservation of Databases," Bern, 9–11 April 2003.

15 National Digital Archive of Datasets: CRDA/35.

16 Shepherd and Smith, "Application of ISAD(G)."

17 National Digital Archive of Datasets: CRDA/21.

Just as we expect governments to be accountable to the citizen, so too must archivists be accountable to their stakeholders for the appraisal decisions they make. Functional appraisal, also known as macro-appraisal, developed by Terry Cook at the (then) National Archives of Canada, is a methodology emerging from the study of social theory developed by sociologists such as Anthony Giddens.[18] It advocates that appraisal should be focussed on functions, processes, activities, and transactions, not on subjects and records. Its basic formulation is that records follow, relate to, and support business functions. Sarah Tyacke has observed,

> If the record, its definition, its selection, and its interpretation has become less certain than in the past, whether by virtue of post-modernist influence or by virtue of the instability of digital records ... it does not mean that archivists should abrogate themselves of the responsibility for selecting what we regard as the authentic and reliable record of the past.[19]

Archivists need explicit strategies and criteria to turn worthy and grand objectives into reality. Like it or not, archivists actively shape the documentary legacy of their own time.[20]

Datasets and the Appraisal Issues they Raise: NDAD's Perspective

From NDAD's perspective, certain issues have stood out which I will proceed to describe. This approach is at a micro-level and does not discuss appraisal more broadly.

The fledgling nature of the project meant that initial surveys carried out to identify potential databases for transfer did not adequately describe the systems that existed in departments. It was also challenging to raise awareness throughout government departments with regard to selection of datasets for transfer. The concept of a database as a public record was a new one. It also involved, for the first time, a variety of people spanning the whole process from design to implementation. These stakeholders can include the data owner, system designers, and statisticians, many of whom had little awareness of archives. These people can be extremely important as they provide contextual information about the data, which often has not been noted in any standard format.

18 Library and Archives Canada: Services to Government, "Appraisal Methodology: Macro-appraisal and Functional Analysis. See <http://www.archives.ca/06/06101_e.html>, visited 20 May 2003.

19 Sarah Tyacke, "Archives in a Wider World: The Culture and Politics of Archives," *Archivaria* 52 (Fall 2001), p. 22.

20 Sir Hilary Jenkinson firmly believed that archivists should not appraise records, as this would compromise their role as custodians of documents left by the creator.

Technology has changed the way that government operates.[21] So while datasets can be selected for the important data they hold reflecting government policy and administration, they also represent interesting innovations, either technological or organizational, in the British government. Computer systems that changed what was possible, rather than just re-implemented manual processes, are of great historical interest. Striking successes of the use of IT in government are good candidates for preservation, as learning tools for the future if nothing else. Existing processes are obviously speeded up,[22] but new processes evolve from the use of technology.[23]

Metadata is another important issue to be considered during selection. If there are no codebooks or interpretative metadata, the data may be meaningless. In particular, it should be remembered that even "readable" data is not self-evident. Often, even if the data can be extracted from its software and the storage medium on which it was created, it cannot be understood unless key documentation is available. This documentation can include technical information such as the meanings of encoded values in the data and descriptions of

21 An example of this is the manner in which information technology has enabled government to increase public surveillance and social control using a variety of systems. See James A. Rule, *Private Lives and Public Surveillance* (London, 1973). Another study by Helen Margetts establishes information technology as a vital feature of public administration, exploring in detail its real impact on the central governments of Great Britain and America since 1978. It reveals the two governments' information systems, the struggle to keep pace with technological development, and the battle to fulfill the grand promises. The author places information technology at the centre of public policy and management. Four case studies demonstrate how information systems have become inextricably linked with the core tasks of government organizations. The key government departments examined are the Internal Revenue Service and Social Security Administration in the US, and the Inland Revenue and Benefits Agency in the UK. See Helen Margetts, *Information Technology in Government: Britain and America* (London, 1998). I am grateful to Kevin Ashley for these references.

22 An example of the re-implementation of manual process is the 1880 US Census. Counted by hand, it took some seven years to complete. When Herman Hollerith's machine (the first electromagnetic punch card system) was used to count the 1890 census it took the US census department a mere six weeks to finish the job.

23 "The MAFF Coastal Defence Survey (National Digital Archive of Datasets: CRDA/10) is one dataset in NDAD that resulted from a contracted-out process: not just the computer system but the entire data collection and interpretation process. Perhaps this would be a good test for the theory about de-skilling of the civil service as a result of outsourcing. We have certainly ensured that we have preserved documents that describe not just the data itself, but the contracts that governed how it was collected and how quality control was to be carried out. They certainly don't seem to show any lack of relevant skills in the department, quite the reverse, in fact. However, they don't throw much light on why the survey has not been updated. The insurance industry is certainly concerned that we don't have current information on the threat of flood and subsidence from weakened coastal defences: the unquantified risks are reaching very large amounts." K. Ashley, "Process Re-engineering: A Brief History of Government Computing," NDAD News <http://ndad.ulcc.ac.uk/news/ndad_news/>, visited 23 September 2003.

the functions of fields. This information should come with the data in the form of data dictionary files and lookup tables. However, sometimes in the case of legacy systems supporting documentation may not have survived.[24] It is equally important to have contextual information, which often will not be recorded in any standard way, such as why and how the data were gathered. If the data were gathered in a survey, it can be very important for future analysis that we know the questions which were asked in the survey and its methodology. These questions can also reveal much about society at the time.[25] Even the layout of the questionnaire may be important for understanding the responses. In the case of either legacy systems or existing systems such risks need to be assessed in establishing the amount of essential metadata that will be needed and must be available to explain the data, to ensure its reliability and authenticity, and to render it valuable to the researcher.[26] NDAD has a system of validation checks it performs on the datasets it accessions which go some way to determine essential metadata. Such issues have caused dilemmas, as it is important to ensure a record's validity. This can be difficult with old systems (as well as some present-day systems) which do not have a method of audit trails. But difficulty should not be an excuse for inaction and again risk assessments must be taken in such instances to ensure as much as possible that we accession a true and reliable record.[27]

As indicated by Shepherd and Smith, the intellectual and physical boundaries which exist in paper frequently do not appear in their electronic counterparts.[28] Electronic records can exist in uncontrolled, unstructured environments in which it is frequently difficult to identify filing systems. Electronic records and in particular datasets very often do not have clear boundaries compared to traditional paper-based records. This may affect selection decisions: where does this system begin and where does it end? Can we select part of a large dataset and disregard the rest? What is the record in such a context?

24 Risk assessments of legacy systems are done by TNA on a case-by-case basis on each individual legacy dataset. They include the number of existent copies of the data as well as the acceptable loss rate.

25 An example is the AIDS Advertising Evaluation Survey questionnaires, which demonstrated current attitudes towards homosexuality and AIDS in the 1980s.

26 "Advanced technology is making it easy to fool people. It would be well if technology also devoted itself to producing forms of records, photographic, printed, sound-recorded, which cannot be altered without detection, at least to the degree of a dollar bill. But it would be still more effective if the code of morals accepted generally rendered it a universally condemned sin to alter a record without notice that it is being done." This is an optimistic quotation of 1967 from famed computer pioneer Vannevar Bush in William G. Carlton's essay "Pax Atomica" from Richard Rhodes, ed., *Visions of Technology: A Century of Vital Debate about Machines, Systems and the Human World* (New York, 1999).

27 For definitions of post-Jenkinsonian terms, see the UBC project Web site at: <http://www.interpares.org/UBCProject/>.

28 Shepherd and Smith, "Application of ISAD(G)."

This has caused us to pause and think about our definition of basic archival terms such as "records" and "recordness," a sort of postmodern angst about accepted values. Research programs have dedicated themselves to undertaking this task of "requalifying" terms. In a sense, electronic records have caused us to question much of our profession's accepted norms.[29]

Capturing Data for Transfer

Another issue to consider at time of selection is how the data should be captured. The data are provided to us usually as a snapshot taken at a given time. The frequency of snapshots is dependent on the frequency of data modifications and deletions and also on when legal or business requirements necessitate major deletions. Snapshots should also be taken before major changes in the logical structure (schema) of a dataset.

Snapshots of data are taken usually at the end of a financial year or at the beginning of the calendar year, whatever is more meaningful to the dataset. For example, data containing financial information will be suited for a snapshot at the end of the financial year. A dataset operated by a university may require snapshots at the end of the academic year.

Snapshots, of course, work well for those "first generation"-type datasets, which are not dynamic and do not change or are not overwritten. At the time of appraisal, decisions need to be taken with regard to whether snapshots are adequate and whether they adequately capture the information in the system.

In relation to transaction files, however, snapshots do not fully tell us when changes occurred and thus certain data can completely disappear. An example of this is the Indian Registration System, a database/register of Indian births and deaths in Canada.[30] When the snapshot of this system was taken at the end of the calendar year, a baby born in January and who died in May of the same year did not appear in the subsequent year's snapshot. The result was the loss of key mortality data regarding indigenous peoples in Canada. In this case the transaction files must be captured. However, it is important to note that just because we have the technology to capture transaction files we do not always have to do so. This decision should be made when appraising the files. What is clear however is that none of these decisions can be made without clear appraisal rationale and methodology.

29 For an introduction to postmodernist ideas and archives see Terry Cook, "Fashionable Nonsense or Professional Rebirth: Postmodernism and the Practice of Archives," *Archivaria* 51 (Spring 2001), pp. 14–35; and Terry Cook, "Archival Science and Postmodernism: New Formulations for Old Concepts," *Archival Science* 1, no. 1 (March 2001), pp. 3–24.

30 Library and Archives Canada, Records of the Department of Indian Affairs and Northern Development, R216–21–8–E, Indian Registers and Lists Series, Indian Registration System.

Transfer

Once a dataset is selected for transfer, the relevant client manager at TNA for the particular department notifies NDAD to take custody of the data and to assist the department in identifying associated documents which should accompany it. Two transfer forms are dispatched to collect technical information and non-technical information respectively.[31] The Departmental Records Officer (DRO) in the creating department facilitates the dissemination and collection of the forms. This allows for the gathering of as much metadata as possible relating to the creation and operation of the system as well as other crucial contextual information. The aim is to gather adequate information at this stage in order to keep to a minimum the number of times we need return to the department with further questions after transfer. We also need to identify any possible technical problems before the dataset is transferred.

Throughout the process NDAD endeavours to obtain information from individuals involved in all stages of the life of the dataset: creation, upkeep, use, and interpretation of results. This involves dealing with both technical and non-technical staff within and outside the department. The information we must gather as part of the accessioning process includes the boundaries of the dataset and the functionality of the original systems involved. Boundaries are often not an issue, but in some cases one dataset may actually form part of a larger computing system; in others, a complex dataset may be presented in different ways to different user communities and the intention of the original selection decision may not be clear. Functionality of systems is one of the key elements we seek to describe in the eventual catalogues. To be able to interpret the records in their original context, one must be cognizant not only of the data contained within them, but also of how the data could be retrieved, interpreted, and displayed when used in its original form. Current computer systems may often make it possible for us to derive information from a dataset in a way that would have been impossible for its original users.

When comparing the transfer of datasets to the transfer of traditional records certain differences can be noted. For example, dataset documentation is very often crucial to the understanding of the data. Codebooks or data dictionaries held in paper format may have been discarded or may be missing while the encoded data has been migrated. NDAD has found that despite our detailed transfer forms, particularly with first-time transfers, initial acquisition of the data marks only the beginning of discussions with and clarifications

31 The transfer forms were originally organized as one composite form. It was decided to divide it in two so we could target more directly the data owner as well as the Departmental Records Officer and include the recurring questions which have caused us to return to the department. There will always be reasons to return to the department but it is hoped that with a well completed form that this will become increasingly unnecessary.

from the department. Very often system/dataset documentation does not exist and we need to rely on the "oral tradition" of information gathering.

Transfer Issues

Because the project was new, involved the transfer of a new type of record, and brought in ULCC, outside TNA yet intrinsically part of it, we did not really know what to expect. Traditionally, the transfer of archival material takes place and there is no need to return to the department to ask further questions concerning the records. This is not what we experienced in relation to datasets. Referral back to departments occurs for several reasons: the partial transfer of data; the need for clarification on access issues; the transfer of corrupt or unreadable data; the lack of sufficient dataset documentation vital for interpreting the data; the need for general information about the data such as input and output, and the lack of field descriptions, etc. Paper records once over thirty years old are usually open unless extended closure is deemed necessary. The records we deal with are often contemporary and as a result we often have questions concerning access.

Raising awareness of NDAD in large British government departments can be quite a challenge. The novel nature of the model we use to preserve datasets means that even with preliminary notice sent by TNA to departments it is necessary to review the rationale for NDAD prior to discussions relating to transfer. This needs to be handled delicately in order to ensure that the department develops trust in the service we offer and often requires meetings with the relevant department. This is coupled with various outreach approaches.[32] NDAD has organized a series of NDAD Open Days in which information is targeted at the various stakeholders involved in the creation, transfer, and use of data within government departments.

The prioritization assigned to NDAD by transferring departments can also be an issue. Often DROs are not archivists or records managers or they have a dual responsibility within their department. This, along with the rapid turnover of DROs, can hamper progress in terms of transfer and can result in a rather dislocated link between NDAD and the department. Promoting the concept (and reality!) of a dataset as a historical record has also been an important issue especially when dealing with people who have never been involved in the transfer of public records.

Essentially our experience has taught us that the transfer of data involves many diverse stakeholders with whom one must communicate adequately in order to build a relationship of trust and efficiency. The nature and cross-section of the expertise at NDAD has assisted in developing these relationships.

32 The lack of awareness of records management issues is not a problem that is specific to digital records but the involvement of other actors, such as IT departments, makes its impact keener.

Preservation[33] and Cataloguing[34]

Whether it arrives by post or e-mail, the first task to be performed with incoming digital media is to create "bit-wise" copies of the source data.[35] These are placed on our secure server designated for incoming storage. All subsequent processing of the data is to be carried out using these copies as input, rather than the original source media. The exact process to be followed differs depending on whether the input media is file-structured or not. File-structured media effectively fall into one of the following classes:

- An ANSI-labelled tape[36];
- A tape or disk containing a file archive in tar, cpio, ZIP, MS-DOS, or VMS backup format[37];
- A DOS-formatted, MACos-formatted, or VMS-formatted floppy disk containing named files[38];
- A CD in ISO-9660 or High Sierra format.[39]

For datasets, the fundamental choice the data specialists need to make when selecting a preservation format is whether to use a format consisting completely of characters or one which uses an entirely binary or mixed character/binary format. Whatever format is chosen, character data must always be converted to ISO 10046 (also known as extended or 8-bit ASCII) if this is not its source character set. Standard conversion tables are available for conversion from the extended Windows character set and from standard EBCDIC.

The preservation strategy adopted by NDAD is that of migration. The goal of the data transformation process is to produce data in one of a number of standard formats suitable for access for browsing by users of the system and for generating copies of data for use by researchers in current computing systems. The encoding and storage formats chosen must be amenable to future conversion to different formats as well as provide convenient access for today's users. It is also essential that the transformation process preserve the content and the intellectual ordering (as distinct from physical ordering) of the original dataset as far as possible. Whatever steps we take to transform the

33 Preservation issues of a technical nature dealt with in this article are taken from the in-house NDAD procedural manual. Kevin Ashley, *Computer Readable Data Archive Digital Preservation Manual* (London, 1997).

34 Drafting finding aids.

35 Departments are responsible for providing secure transportation of archive material to the NDAD premises.

36 American National Standards Institute (ANSI) compliant tape.

37 These are methods for "packaging" files.

38 These are various operating systems.

39 These are standard formats for writing and placing fields and directories on a CD-ROM.

data, it is essential to record precisely what actions we performed. Data specialists evaluate the completeness of this recording process using simple criteria: if the process had to be repeated by the original staff or others with similar skills from the beginning, using only the original data, the description, and the metadata accompanying the dataset, could it be done? If the answer is "no," then more detail needs to be added to the records of transformation actions.

Data are converted to one of three standard formats for permanent preservation. All are essentially flat files, two entirely textual and one binary. The form chosen is that which is most appropriate given the original form of the data. A wide range of international or publicly-available standards are used, from the text codes of ISO 8859,[40] the data storage formats of IEEE 754,[41] the image storage formats defined by TIFF,[42] down to the standard date format defined by ISO 8601.[43] A key element of the accessioning process is the creation of metadata describing the dataset and its constituent tables. Depending on the data source, some of this metadata may be created automatically from information in the source database or application. In other cases it is created by hand, after an inspection of accompanying technical documents. A single metadata file is created for each table and performs multiple roles: it controls access to material (by defining fields which must be anonymized); it also controls the display of data; drives the menu system which allows users to make queries of the data; allows automated conversion of data to new formats now (for export to other systems) or in the future; and automates the generation of piece-level catalogues which describe each individual field in a dataset. The time it takes to create this file varies. If, for example when dealing with SPSS files, we do not need to add metadata (as it is already often present in the original SPSS file), this process can take a short amount of time, even minutes. If there is no metadata present, this process can take much longer, because we have to create the metadata ourselves, either by tracking it down from documents, or by re-contacting the department to verify our results with the data file. Some datasets can hold a single file with sixty fields while others can hold up to 400 files. The variety in size and complexity of a database means that it can take from one to several months to process an entire dataset.

Preservation of Dataset Documentation

Dataset documentation, if it is transferred in hard copy, is preserved as TIFF

40 This ISO standard provides a full series of ten (and soon even more) standardized, multilingual, single-byte, coded (8 bit), graphic character sets for writing in alphabetic languages.

41 Standard for Binary Floating-Point Arithmetic.

42 Acronym for Tag(ged) Image File Format. It is one of the most popular and flexible of the current public domain raster file formats.

43 This ISO standard specifies numeric representations of date and time.

files. These are converted on the fly to other image formats when access is via the Internet. Electronic documentation is also made available in plain text. This will usually preserve a greater amount of the structure and format of the original document, but will not retain the original format or layout of the text.

Before converting any digital documentation from any proprietary format, metadata is extracted which can be found in the document. Metadata may not always be present, and may not be meaningful if it is present. Some systems, such as Microsoft Word, will always create certain pieces of metadata such as a document title whether or not the author chooses to provide it. Often this information is taken from a document template, or from the first heading in the document. Metadata which should be preserved include creation dates, modification histories, author, title, keywords, and document abstracts or comments.

Once one or more preservation copies of each item of digital documentation is made, the accessions system is used to introduce it to permanent storage in the same way as with a dataset. This accessions system has reserved locations and numbers for the documentation which tie it to the dataset as a result of the registrations made at the time of transfer.

Description

One of the unique facets of the project is the use of archival standards in the management of databases. The use of ISAD(G) as a standard was an obvious choice for cataloguing data.[44] Using ISAD(G), NDAD created a finding aid structure which includes an administrative history for each department and agency which has transferred material to us. These contain at least the name and date of creation and (if applicable) dissolution of the current department and all predecessors; the description and dates of legislation establishing the current department and establishing and disestablishing predecessor organizations; the administrative hierarchy; the purpose of the department; its current and past functions; the history of sections of the department which have trans-

44 Other initiatives and standards were considered. The Data Documentation Initiative (DDI), developed by the international social science data community through its association IASSIST, was considered by TNA but it was still in development at the time of contractual negotiations. It was considered much stronger in relation to technical metadata but not as strong in relation to contextual description. Other initiatives which have since emerged include the baseline authenticity requirements for preservation identified by InterPARES; NARA, with the San Diego Supercomputing Centre, as well as its own provision of on-line description and access to some of its born-digital holdings; and CEDARS. The CEDARS (curl exemplars in digital archives) project ran from April 1998 to March 2002. It explored digital preservation issues ranging through acquiring digital objects, their long-term retention, sufficient description, and eventual access. It was based at the University of Leeds. See <http://www.leeds.ac.uk/cedars/>, visited 5 January 2005.

ferred records to NDAD; and the datasets transferred by the department to NDAD.

NDAD's holdings of datasets and documentation are not arranged by record group/fonds or by sub-group/sub-fonds. Administrative histories play the role of preserving provenance at an intellectual level. An administrative history is linked to as many series catalogues as are relevant to that administrative history. A series catalogue, in turn, can be linked to one or more administrative histories. This was modeled on the TNA Current Guide at the time of the development of the project. A flexible structure circumvents the problem of datasets and documentation with multiple provenance, such as databases created by one department and continued by a successor; or "interdepartmental" databases where data is shared between more than one department simultaneously. It also allows for seemingly regular mergers and separations of UK government departments. Researchers can access our catalogues either from our site or via TNA's online finding aids, where datasets are embedded at a series level in connection with their associated records.

The finding aids also include catalogues for each series of datasets, or what we at NDAD have established as a series, such as annual versions or "snapshots" of an ongoing census or survey, or an individual dataset where the dataset is a one-off and not part of a regular process. An example of an artificial series is the North Sea GIS, for which a snapshot is taken at the end of every financial year for transfer to NDAD. In the case of the Schools' Census, an annual survey of schools conducted by the former Department for Education and Employment (DfEE) and its predecessors covering schools in England and (up to 1977) Wales to gather data on topics such as pupils, teaching staff, classes and examination courses, the collection of annual surveys constitutes the series. An example of a series based on a one-off survey is the Children's Difficulties on Starting Infant School Dataset, which provides descriptive information on the nature and extent of children's problems on starting infant school, and the extent to which they were affected by external factors.

The series catalogue includes the administrative history of a dataset, conditions of access to the dataset, copyright in the dataset, system attributes, and logical structure and schema. Sub-series may also be used where necessary. At the file level, individual datasets are described as part of an annual or regular series. Dataset-level catalogues describe information peculiar to a specific instance of a dataset (such as any elements added or removed from a particular year's survey). Item-level catalogues describe the individual tables within the dataset and the fields which they contain. These table catalogues are generated automatically from low-level metadata. The information provided for the latter includes descriptions of the data types for each field, attributes and constraints of each field, and field names. Finally, an on-line thesaurus is provided which acts as an authority list for keyword searching and for the finding aids.

The dataset documentation, which assists with the understanding of the data and includes system documentation, is described in a documentation catalogue attached to the series-level catalogue.[45]

Following TNA's adoption of XML and EAD, NDAD has been retrospectively converting its catalogues from HTML to EAD through the use of a system designed in-house called CERES. This enables archivists to input catalogues into a database structure to create EAD files of the catalogue on the fly for export to TNA. As a result, our catalogues will eventually be available in XML as well as HTML.

ISAD(G)

The catalogues as a whole are produced in conformity with ISAD(G), using the elements which are considered essential for international exchange of descriptive information and others which are relevant to datasets.[46] These include:

- reference code;
- title;
- date(s) of creation or date(s) of accumulation of the material in the unit of description;
- extent of the unit of description;
- level of description;
- access conditions;
- copyright;
- accruals; and
- scope and content.

Hierarchical description, a keystone of the archival profession, was developed to reflect the arrangement of paper records. When applying traditional archival practise to the description and arrangement of datasets, certain issues inevitably arise. Electronic records can exist in uncontrolled, unstructured environments in which it is frequently difficult to identify filing systems.

45 System documentation typically includes such things as the source code of computer programs and descriptions of data produced by and processed used with the original system. Just as the dataset documentation helps place the dataset in its administrative context, the system documentation sets it in a technical context, allowing researchers with the appropriate skills to understand more precisely the nature of the data processing in question, and the particular systems and methodologies used by departments when these computer systems were originally designed and implemented. (Definition supplied by Richard Davis, data and applications specialist, NDAD.)

46 For greater detail on the application of ISAD(G) in NDAD's descriptive work, see Shepherd and Smith, "Application of ISAD(G)."

Their existence is not as tangible and boundaries are not as set in place as with paper records. But the application of archival arrangement to electronic records can be achieved at certain levels and the administrative function of the databases can help identify the higher level of description.

Relationships between the tables and fields within a database help determine the lower levels of description. As Shepherd and Smith argue, the biggest discrepancies occur at the series level, due to the "amorphous nature of datasets."[47] This can be seen when accessioning active datasets which are continually updated and which are captured once yearly in a snapshot. This is described as an artificial series.

Other issues noted when applying the use of ISAD(G) to datasets include the date elements. The complex nature of datasets leads to difficulties when indicating the dates both of creation and content. Creation can mean the actual design of the system, i.e., the structure which holds the data, as well as the creation of the contents. Other dates include date of last input, last access, and dates of active use.

Issues also arise with regard to the creator element. Due to the two-dimensional nature of datasets, there most likely will be two creators: the database designer and the compiler. As stated already there are many stakeholders involved in the creation, maintenance, and compilation of a database. As a result of Shepherd and Smith's study we have adopted a "Statement of Responsibility" to facilitate multi-provenance.

Within the Scope and Content area of ISAD(G), the "System of Arrangement" element was removed. Instead we have the element "Logical structure and schema," which allows a description of the relationships which exist between tables and fields of datasets.

Description: Expansion and Change

In addition to expanding ISAD(G) elements, we have added several new elements which reflect the particular requirements of describing datasets. These include a "System attributes" element which describes the software and hardware, and "Validation" elements which describe how data input to a system was validated and how we have validated our own transformation of the data. The logical structure and schema element describes the tables within a dataset and their relationship to each other. We have also expanded the date section of ISAD(G) to include dates of creation of datasets and dates of contents. Our use of ISAD(G) has shown that it is possible to use the standard to describe a specialized kind of record or records such as electronic records. It is an effective and adaptable "meta data" standard for discovery and preservation which

47 Ibid.

can be applied to electronic records. However in order to do so, it is necessary to add extra elements to cover the specific needs of that kind of record as we have done.

In addition, it is questioned whether the principle that a record shall belong to only one fonds or sub-fonds can still be valid. The ever-changing nature of UK government departments coupled with cross-departmental workgroups as well as complex systems used and shared by many sections, suggest that this can no longer operate successfully, and in fact is highly problematic. The virtual nature of the environment in which Cook's third generation of electronic records emerges, with resources shared through the Internet, challenges these "set-in-stone" assumptions.[48] The series approach adopted by NDAD permits the description of records of multi-provenance, allowing us to link to one or more fonds.[49]

Web Site

The catalogues and the datasets themselves can all be accessed via the Internet.[50] Access to NDAD is also available from TNA's on-line catalogues: NDAD emerged at a time when TNA was developing its on-line catalogue (PROCAT). A situation whereby the NDAD catalogue descriptions were merged in PROCAT, and the researcher connected with NDAD only when wishing to access the data, would allow better integration of our holdings with those of TNA.

Access to the catalogues is unrestricted (except where there are closure requirements on parts of the catalogues themselves), as is access to all other parts of the Web site apart from the archives themselves. Access to the archives (either datasets or their documentation) requires registration. Registration is free (as is use of the archive), and can be carried out via the Web site. The home page leads to a variety of resources, including on-line help, background information on the service and newsletters, as well as the archive catalogues themselves. A visual map of the site is provided to help users navigate through what, at times, can be a complex structure. The catalogues can be searched using either free text retrieval (with Boolean and proximity operators) or via thesaurus-assisted keyword selection.

48 See Cook, "Easy to Byte, Harder to Chew."

49 A survey of traditional archival records held in the Australian National Archives as early as 1974 indicated that 27 per cent of the series were of multi-provenance. See Peter Scott, "Facing the Reality of Administrative Change – Some Further Remarks on the Record Group Concept," *Journal of the Society of Archivists* 5, no. 2 (1974), p. 94.

50 NDAD was one of the first data archives to provide on-line access to data. NARA started in 2003.

Thesaurus

The thesaurus contains a list of subjects, places, and personal and corporate names mentioned in the administrative histories and catalogues and covered by the datasets. The subject section of the thesaurus is based upon the 1995 edition of the UNESCO thesaurus. Terms from the UNESCO thesaurus are used to index all NDAD catalogues, i.e., administrative histories, series-level, dataset-level and dataset documentation catalogues. We do not index the data itself, as this would allow the researcher to access the data without reading through the contextual information provided. The purpose of the UNESCO thesaurus is to index literature in the areas of education, science, social and human science, culture, communication and information, and it was chosen in preference to others as it contains terms corresponding to the kind of material to be described and is easy to use. Also, it was already in use by at least one other data archive. Wherever applicable terms in the UNESCO thesaurus can be found, they are used to compile the thesaurus for NDAD catalogues. These terms together form the NDAD version of the thesaurus. When terms are added to the NDAD thesaurus they are allocated their appropriate micro-thesaurus broad, narrow, preferred, and related term entries in accordance with the UNESCO thesaurus. Terms additional to those in the existing UNESCO thesaurus are added only where necessary. These terms are also allocated an appropriate micro-thesaurus heading and relevant broad, narrow, preferred, and related terms in conformity with the structure of the UNESCO thesaurus. New terms added by NDAD are noted so that a record can be maintained of such changes. Thesaurus entries, together with Web page links, are maintained in a database, and the contents of the database are exported each night to build the components of the thesaurus search system. NDAD is also developing an authority list for place, personal, and corporate names.[51]

Design of the Web Site

It is interesting to note here that the structural model of the NDAD Web site reflects the archival levels: fonds, series, sub-series, file, item, and piece. The administrative departments and the sections within those departments represent fonds and sub-fonds, or sub-fonds and sub-sub-fonds of a central government fonds. The series is equated with one or more datasets originating from a

51 Their use is based on the National Council on Archives' *Rules for the Construction of Personal, Place and Corporate Names* (1997). The development of this excellent means of searching with the on-line thesaurus is one of the many examples of the constructive and co-operative working relationship which exists between the data/application specialists and the archivists in NDAD. My colleagues Peter Garrod and Louise Craven of TNA are actively engaged in work in this area.

single government department. The sub-series level and the file level represent a sub-series of datasets or a single dataset. The item level is used to describe tables within a dataset, and the piece level equates to fields within a table. This hierarchy underlies NDAD's classification system, so it was hardly surprising that it should also provide the framework for the Web site. It also helps users, particularly those with an archival background, to understand our site. While navigating up and down from higher to lower levels of description, the researcher is provided with visual indicators of where exactly they are in relation to the hierarchical structure of our finding aids.

When browsing the datasets themselves, the user can exercise control over what is displayed in a number of ways. Particular fields for display can be selected from a given dataset table. In addition, queries may be used to limit which rows of a table are seen. The queries can be entered directly using an SQL-like syntax, or built up by using menus of field names and operators. Queries can compare single fields against specific values or a range of values, as well as perform certain other special tests, and can be joined with a variety of Boolean operators. In the absence of a query, data is simply displayed in the order originally present in the table. Users browse through the data, page by page (the page size being selectable), and may alter the fields being displayed or the query being used at any point during the process. Simple tabular presentation is used for most forms of data. If required by the transferring department, the system can blank out certain fields from a table containing sensitive information (such as names or addresses) or can deny access to certain tables altogether. Whilst viewing the data itself, links are provided back to the descriptions of the data and most browsers will allow simultaneous display of both sets of information.

Access Issues[52]

Datasets which are public records are subject to the same criteria as paper records in the public domain. In the UK this means that they are traditionally closed for thirty years, until their transfer to an archival repository, but this closure period does not apply to records opened to the public prior to transfer. They obviously must remain open. The Lord Chancellor has discretion to open records, with the concurrence of the minister concerned, earlier than the specified period. This is termed accelerated opening. Records may be kept closed to public access longer than thirty years to prevent a breach of good faith. They also may be retained in departments beyond the thirty-year period. Retention is primarily a matter of the physical whereabouts of a record, but

52 The access situation will change with the entry into effect of Freedom of Information Legislation on 1 January 2005.

when departments are authorized to retain records, a decision will be made whether to permit access. Records may not be closed or retained without the concurrence of the minister of the department concerned, and of the Lord Chancellor, who receives the views of his Advisory Council.

Copyright

Most of our records are subject to Crown Copyright. Registered users are granted permission to access this Crown Copyright material and to download brief extracts of the Crown Copyright material onto electronic, magnetic, optical, or similar storage media, provided that such activities are for private research, study, or in-house use only. They must also not distribute, sell, or publish any Crown Copyright material taken from a Web site. Such use of the material requires formal permission of the Controller of Her Majesty's Stationery Office.[53] Other copyright issues arise with datasets when the department has used the services of an external non-government company to design or develop a system. The company may try to guard the copyright at times even after the system is in use.

The User: Access

This was the first time in the world that a national archive provided on-line access to born-digital material. Similar projects at the time preserved databases but did not provide on-line remote access to their holdings. At the time of development of the project, little was known about how researchers would react to remote access to archival materials. A test was conducted with "tame" users drawn from archive schools, museum studies students from the University of Leicester, as well as students from the Institute of Historical Studies at the University of London. Many of these people were not familiar with the use of the Internet and had to be shown from scratch how to use the site. Once they familiarized themselves with it, they were assigned certain tasks to ascertain how easy it was for them to complete the task. The results were noted and conclusions applied to the Web site. Four years later, after significant changes to the user community, coupled with our expanding holdings, it was decided that further reconnaissance in the area of user services was required.

User Services

One of the more obvious drawbacks of having an on-line archive is that one is no longer in touch with the users on an immediate basis. One can gauge what

53 <http://ndad.ulcc.ac.uk/popup/copyright/crown.html>, visited 3 December 2004.

datasets are being used from the user logs, but these do not indicate to what extent or to what purpose the data is being used. Our on-line registration form asks our researchers to register their field of interest. Even though this section is frequently completed, we knew that we needed to find out more about our users. In the fledgling field of on-line data archives there has been some (but not much) investigation into secondary users and uses of data. An interesting example is the Center for Electronic Records' examination of their users and the nature of their research, which resulted in fascinating study conducted by Margaret O. Adams, as yet unpublished.[54] As traditional archives are now providing many of their holdings on-line, we were not alone in wondering who our users are.

An important part of our outreach drive was a publicity campaign. The result of planning this campaign was a total re-branding of the project using professional design. The image to be conveyed was one that reflected the digital nature of our holdings but emphasized our "archivalness." We produced posters and information leaflets that would be useful to both depositor and archivist alike. This material has been and is being sent out to various interested institutions such as corporate and public libraries. An on-line questionnaire and an electronic mailing list have also been developed. In this way we plan to collect information on the secondary use of our data. It is important to clarify that our holdings are on the whole *not* historical as they are in the main contemporary records. Nonetheless we have records dating from the 1960s. The traditional user of a data archive should be made aware of our holdings, but a certain amount of capacity building needs to take place with traditional users of archives to break down any misapprehensions they may have in relation to data.[55] Part of our outreach remit will extend to NDAD open days where we demonstrate to depositors and users what we hold and how to use the data.

Conclusion

TNA has charted a pioneering course, recognizing that while at the outset it may have had few answers to the problems facing the archival community, it was not an option to do nothing about the preservation of databases. Through the establishment of NDAD, it was in the position to ask some interesting questions, knowing that any solution would not be perfect, but also that it is easier to understand the issues after making the effort to do so.

NDAD considers that it has done well in defining what it means to preserve

54 Heard by the author at the Society of American Archivists conference in 1998.

55 "All those interested in studying society, past or present, need to take charge of quantitative data: to command it rather than to be the slave of a seeming authority of numbers emerging from documents or the writings of a small body of numerically inclined researchers." Pat Hudson, *History by Numbers: An Introduction to Quantitative Approaches* (London, 2000), p. 5.

a database. Due to the novelty of the process it needed definition as well as implementation by NDAD. Though other approaches can be used in the preservation of databases, the approach we took is still relevant and universally accepted.[56] NDAD has not employed exotic technology or difficult techniques to achieve its aims, although we believe we have brought them together in a unique way. The only exceptional element in our infrastructure is a storage server with a large capacity, 300 terabytes, although this is standard technology used by thousands of banks, oil companies, and other industrial concerns. Collaboration was the key to the successes of the project and it is now understood that the archive profession cannot afford to shy away from the management of digital media and would be misguided to believe that it can cope with the problem alone without the technical expertise and experience which data and applications specialists have to offer. Initiatives such as the Digital Preservation Coalition (DPC),[57] the National Preservation Office (NPO) based at the British Library, and CEDARS are strong partnerships which work with the tremendous cross-section of stakeholders involved in the area of digital records and their preservation. Equally, data and applications specialists would be ill-advised to try single-handedly to preserve such media, as simple physical access or its virtual equivalent is not enough. Additional information and explanations are needed in order for researchers to understand and make use of collections and materials.

We can now make sense of large amounts of information gathered in a format which is easy to manipulate and re-use and all at a speed beyond the Ptolemies' wildest dreams. We can now imagine qualitatively different and hitherto unimaginable ways of doing research. But we have yet to develop a capacity to keep up with the technology we produce. Digital preservation is one of the casualties. TNA, a pioneer along with other institutions, must be lauded for its attempt to act in this area. To re-word Professor Steven Bird's testament, we are sitting between the onset of the digital era and, if we fail to act, the mass extinction of the world's memory.

Data is key to government operations. It provides crucial support to many of its creators' processes and decisions. It is often the raw material from which countless documents, reports, and statistics (often quoted during endless hustings) are produced. If we do not keep this record, sometimes deemed "unwor-

56 "[M]igration has been suggested as a primary method of digital preservation and it has been more widely reviewed than any approach." Seamus Ross, "Changing Trains at Wigan: Digital Preservation and the Future of Scholarship," National Preservation Office, London, 2000.

57 The Digital Preservation Coalition (DPC) was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure our global digital memory and knowledge base. See the DPC Web site page entitled "About the Digital Preservation Coalition," at: <http://www.dpconline.org/>, visited 20 February 2003.

thy" of the archivist's attention, how can these documents be proven to be authentic, reliable records of a government which must be accountable to its citizens? Can data librarians, professionals who emphasize informational content over context, meet this objective? The archival profession has centuries of experience and extensive skills to offer in presenting not only information but the context with which to view this information. Perhaps records are not just evidence of business transactions but evidence of human activity?