The work of the National Digital Archive of Datasets (NDAD)
Submission to Government Computing magazine
By Ed Pinsent 2006

Introduction
The National Digital Archive of Datasets (NDAD) preserves and provides online access to archived digital datasets and documents from UK central government departments. Datasets are collections of structured digital information and often take the form of surveys commissioned by Government departments. This service is provided by the University of London Computer Centre (ULCC) under contract to The National Archives (TNA). All the datasets are public records, which means there’s a legal obligation to preserve them and make them available. The NDAD service began as a unique combination of skills, involving IT experts and archivists working together to provide digital preservation. Every dataset we work on is different, with a new set of challenges.

What NDAD does
The NDAD service captures information from databases, usually by a process of export or downloading, to ensure the core data is captured. We then preserve that data, and its associated metadata, permanently. This preservation isn’t quite the same as backing-up data every night, as we’ll see. NDAD also makes a presentation version of the data, and hosts it on our website, thus rendering it searchable online. Where appropriate, in accordance with the Freedom of Information Act 2000, we will close fields, tables, or even entire datasets if, for example, they contain sensitive or confidential information.

We also offer (digitally, of course) a range of supporting documentation, such as copies of technical manuals, which explain more about how the original database operated. And each dataset has a complete descriptive catalogue which puts it all in context – which department used the database, when, how, why, and what results they got out of it. All of NDAD’s systems and solutions are geared towards the efficient provision of this service.

What NDAD doesn’t do
We don’t provide management or archiving services for electronic records of Government in the form of Office documents or e-mails; these are covered by other TNA services, including the Seamless Flow programme and the work of the Digital Preservation Department. We’re not a hardware / software museum, so we don’t keep copies of programs or old computers. We don’t emulate programs, or their functionality. Unlike conventional record offices, we don’t normally facilitate visits from members of the public. Government Departments, of course, are also entitled to view their documents. But for the most part, our users are online users. And we don’t select our holdings; TNA does the selection and nomination, deciding which datasets should be deposited with NDAD.
What use is a dataset?
Many datasets created by central Government are statistical; they contain large amounts of statistical facts and figures, gathered from across the United Kingdom. Historians, researchers, social scientists and academics find this factual data invaluable. These datasets can be reused for numerous research purposes, sometimes unrelated to the reason they were originally created. Like other public records, datasets are also evidence of organisational activity in Government. They even represent a slice of technological history, showing the success (or otherwise) of particular IT projects, or the take-up of certain systems, across Government.

All of this secondary, re-use value comes into its own when a database is nominated for permanent preservation. What NDAD offers to users is a diverse collection of statistical and technological information, from the last 40 or so years of the UK, whose historical value is growing every day. And users can manipulate, search, and compare the data in any way they wish.

How a transfer happens
There are numerous large-scale database and data management projects happening in Government all the time. Client Managers at TNA, each charged with managing the records from a range of departments, recognise that a department's digital assets are public records too, and every bit as significant as the filing cabinets full of paper which they regularly select for transfer to Kew. When a dataset likely to hold information of historical interest to the public is spotted, the Client Manager works in collaboration with the Departmental Record Officer in the selection and appraisal process, leading to its nomination for permanent preservation.

At this point, NDAD get involved in the process, and we open three-way negotiations with key staff in the department: the Data Owners, staff who collect and use the data; IT people, who facilitate the transfer of the database; and the Departmental Record Officer. All these players are important in the process. The DRO and/or Data Owner can tell us who used the database, and why; the IT person can tell us about its internal structure, and how to export a copy of it successfully.

Are you a ‘data owner’, an IT manager or a DRO? Here’s what you can do to help preserve your database:

Plan. The potential requirement for archiving of the data in any database should be included at the specification stage for any new system.

Keep supporting records. Key documents (high-level business case for the system, specification of requirements, other systems documentation, user manuals, post-implementation review) will need to be transferred with the dataset. Preserve these records within your Department - ideally on registered files.

Remember small is beautiful too! Potential transfers to NDAD could vary from department-wide, long-term shared systems to one-off surveys carried out by an individual scientist.

Tell your DRO. Any readers aware of a dataset which might be a candidate for preservation should contact their Departmental Record Officer.
Capturing a dataset
NDAD aim to secure a copy of the data in the most effective way possible. If it’s a small, static, or semi-current system, it may be possible to simply copy the data files all onto a disk. For systems which are still in daily use and continue to have records added to them, a ‘snapshot’ of the data will be taken. Over time, NDAD aims to accumulate regular snapshots of such projects, thus building up a record series of database copies that, to some extent, replicates the complexity of a dynamic, constantly-changing database.

Processing a dataset
Here we come to the heart of NDAD’s approach to archiving databases. First, we make preservation copies of the data we received, in its native format; these copies are held ‘untouched’ in secure server storage. Next, we convert copies of the data tables into a format that doesn’t rely on proprietary software. The most common format we use for this is .csv (comma-separated value) tables, widely regarded as the most ‘neutral’ and robust format for the storage of tabular data. Table-level metadata, for example data dictionaries, are also saved into usable forms that we can operate in our systems.

This conversion allows two powerful results. First and foremost, it enables long-term preservation of the data, because these formats are not dependant on the changing vagaries of software development; and, for preservation purposes, it is easiest to migrate data to and from such formats over time. This archival action, informed by the very latest thinking in digital preservation standards, isn’t the same thing as relying on your system administrator to do backups of the dataset every night. With NDAD, preservation means active and permanent!

Secondly, the conversion enables NDAD to put the data tables online, in a simple browsable and searchable format, for our users; tables, or even entire datasets can be downloaded. If the tables are all not too large and the user has time, then they can download all the tables for a dataset for free. Otherwise (and particularly for tables over a certain size), the user can order a copy, for which we charge a small fee.

NDAD is an essential and valuable part of The National Archives’ remit (“It is responsible for looking after the records of central government and the courts of law, and making sure everyone can look at them”) in the digital era; we’re preserving important digital records for the long term, and making them available to the public.

Documentation and rendering
NDAD also offers supporting documentation for each dataset. We identify and select any documents which help users interpret and better understand the dataset, who created and used it, when it was used, and what it was used for. Documentation includes materials like technical manuals, blank forms for gathering data, and annual reports of published data extracts. Like the datasets, this information is preserved and rendered. We convert digital documents, putting them into common and safe storage formats such as PDF
or plain text. We convert paper documents, by scanning them; we make plain
text versions and scanned images available, and put everything we can
online. Even pages from a Government website, if relevant, can now be
preserved and rendered as documentation.

**Features of the NDAD service**

From its inception in 1998, NDAD was an innovative solution, begun at a time
when everyone knew what the problem was but didn’t know how to solve it. In
facing up to the problem, The (then) Public Record Office showed far-
sightedness and setting up the NDAD service was quite a bold step for them
to take. Aspects of the service have since been imitated by other digital
preservation institutions. For example, the role of the ‘data specialist’ which
originated in NDAD is one which didn’t really exist in IT circles 15 years ago.

NDAD was founded jointly by an archivist, applying traditional archival and
preservation standards to modern digital records; and an IT technical expert,
devising new and imaginative solutions. Keeping these professional standards
on an equal footing is what has made NDAD work, unlike some projects which
fail when they are too strongly IT-focussed (or conversely, too archivally
focussed).

Today, the NDAD team is multi-disciplinary, comprising a mix of data
specialists for data analysis and processing; archivists, preparing catalogues
and advising on preservation standards; systems designers, for developing
the tools we use; and digital conversion technicians. The strong co-operation
between these disciplines ensures success. The pioneering work at NDAD
has built a strong foundation for digital preservation.

The skills developed for NDAD also transfer well to other aspects of digital
preservation and information management; bringing together elements of the
NDAD service in different combinations allows us to provide a wide range of
other digital preservation services. The team are currently involved in projects
concerning the selection and archiving of websites, using remote harvesting
software; the delivery of a specialist digital preservation training programme;
developing an e-repository for University clients; and the development of a
tool for assessing the preservation needs of digital assets. Our shared roles
and common understanding of preservation issues enable these possibilities
to become reality.

http://www.ndad.nationalarchives.gov.uk/
http://www.ulcc.ac.uk