

National Digital Archive of

Datasets News



Welcome to the second edition of Datasets News from the National Digital Archive of Datasets.

This is a quarterly publication, highlighting the latest dataset releases and developments in the service. The National Digital Archive of Datasets has long been unique in the UK in offering an online querying facility where you can analyse and research archived datasets, as well as catalogues and related documents - all of which are seamlessly integrated to provide a full digital service.

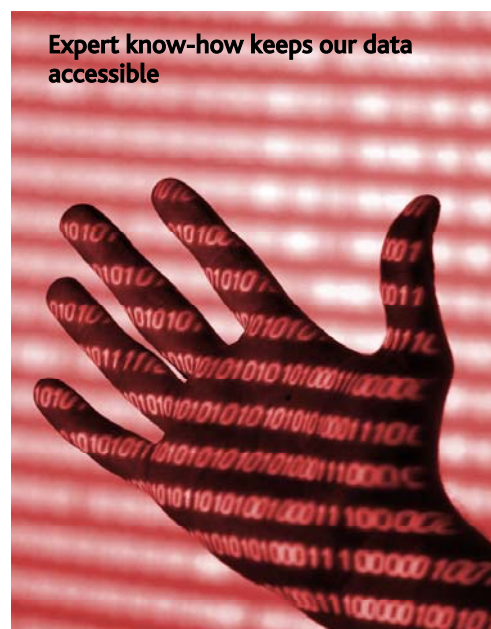
This issue we're focussing on data - search and rescue, with articles by Kevin Ashley and Frances Blomeley, and a look at recent releases.

We hope that you enjoy the Spring issue and are keen to hear your views. Please email us at: digital-archive@nationalarchives.gov.uk We look forward to hearing from you.
The Editor.



Data rescue and the "neural archive"

At the outset of the NDAD service in 1998, we had little idea of what kinds of datasets we could expect to receive from government. Many recently-created datasets tend to be stored in one of a few widely-used database applications, and are usually straightforward to process. However, on some occasions older "legacy" datasets continue to emerge and present new challenges.



Expert know-how keeps our data accessible

Many public records have been generated or stored in electronic format for decades, and in digital format for even longer. The United States began recording and processing Census data using punched cards over 100 years ago. This information was managed mechanically: 'sorters' separated cards according to column values and 'tabulators' printed out information from the cards. Although it would be a long time before punched-card data could be processed with a computer; this innovation still reduced the time taken to process the data from 7 years to 1 year. The value of digital records was established.

In the UK, punched cards were first used to store data in the 1911 Census.

By 1948, 26 punched-card processing machines were installed or planned; and by 1965 computers were appearing, with 45 installed and over 200 more planned for the next decade. The earliest datasets received by NDAD date back to the 1960s and were generated on what is believed to have been an IBM1401 machine with 16000 positions of core storage, six magnetic tape units, a card read/punch unit and an output printer (CRDA/5/DS/1, Primary Births). The procurement of hardware, sometimes the size of a diesel locomotive engine, was a complex and costly operation, and the computers were expected to be in service for a very long time.

In the early days of computing, the marketplace was in many ways a more diverse environment than it is today. Technology resembles biology in that a new design evolves into many different specialist forms, each best adapted to make the most of a particular environment. Over time, competition forces extinctions and convergent evolution. The most powerful, the most adaptable, and the most opportunistic survive.

Continued on page 2.

Continued from page 1.

This constant development means that storage formats and hardware become obsolete, often only a few years after they were created.

Early computers were developed with their own architectures, their own operating systems, and their own character sets. Software for processing datasets was usually a customised application, written in-house. Although there were many computer-programming languages in use, the majority of applications were written in COBOL for business data or FORTRAN for scientific data. The development of databases gradually took over the handling of storage, with programming languages used to handle the interface with the user – data entry, querying, and reports. In later developments, these interface tasks became subsumed into the database application itself.

Since 1998 we have received datasets in a variety of unusual formats, reflecting their diverse origins. These include mixed character/binary format with character data being encoded as EBCDIC (Extended Binary Coded Decimal Interchange Code, an early coding devised by IBM), a disk image of an ICL variable-length-record blocked labelled tape, tapes in Prime MAGSAV format, and fixed-length records in fixed-length blocks in EBCDIC. Some datasets have shown evidence of their travels as they have been migrated from one machine architecture to another; some have been subject to corruption over time during manipulation by different versions of home-grown software.

Some datasets have arrived in fixed-width ASCII format and were readable straight away, but were accompanied by little or no data structure documentation. Rows of numbers are valueless unless we know where the field divisions lie and what they represent. In these cases a range of different techniques can be applied to determine the data structure, depending on the type of application originally used to process the data.

If the data had been processed using a programming language and we have a copy of some program source code, then knowledge of the language can

enable us to follow the manipulation steps. For example, we can see records being converted from one form to another, read from a different device, or processed to produce reports. Data statements may provide informative labels for array names, report headings, etc. If we are very fortunate, programs themselves will contain numerous and helpful comments.

We may also be able to divine the meaning of a particular field by seeking correlations with another known field; we may be able to compare our data with published reports; or we may be able to talk to someone involved in the development and use of the dataset. We may look at the range of frequencies for a particular field and recognise data patterns, for instance dates and years; and knowledge of a particular subject may allow us to make an educated guess on the basis of contextual information.

It is worth pointing out that the determination of the structure of a dataset can't be guaranteed, particularly if no associated material exists; but we're very happy to try!

Other examples of some of the challenges of obtaining intact data from old media, and of tracking down dataset metadata ("digital archaeology") are described in a 2004 Ariadne article on NDAD by Jeffrey Darlington.

Datasets can be transferred in a native database or other application format from which tables are extracted by NDAD. These have included: Cardbox, DataEase, dBase, FirstPoint, FoxPro, Lotus, Microsoft Access, Microsoft Excel, Microsoft SQL server, Microsoft Word, Navidata, Oracle, Qstat, SIR, SPSS, and Symphony. To extract the data from a proprietary application, those applications which are widely used - for example Microsoft Access - do not usually present any serious problems, as many of these packages can also import data from other applications. Where this is not the case, it is sometimes possible to find utilities which will perform direct conversions between different formats. If there are no other options, we have been able to purchase a copy of the application and

learn how to use it to export data.

As time goes on, the chance increases of encountering a dataset for which the original application no longer exists, or does exist but is not compatible with earlier versions; but it is unlikely that the data will be completely inaccessible given the wide range of conversion tools available. To continue the biological analogy, there are also many fossils on display in the computing equivalent of online natural history museums.

These examples illustrate the importance of what in expert systems is referred to as "know-how" – not only the ability to recognise a pattern or process and know how to deal with it, but also the ability to generalise – that is, to make an educated guess based on experience of similar patterns and processes. If you have been dealing with tapes, or computer programs, or software applications, for many years, you will have absorbed considerably more than can be gleaned from a user manual.

ULCC has been involved in the provision of information technology services for nearly 40 years; cumulatively, we have literally centuries of staff experience across the range of hardware, operating systems, programming languages, application software, and data formats. When the NDAD service began, we had no real idea of what to expect in terms of the kind of datasets that would be transferred. In the event, one of the service's most valuable assets has proved to be this pool of "know-how" at ULCC: what might be termed the "neural archive".

Frances Blomeley
Data and Applications Specialist

Jeffrey Darlington, A National Archive of Datasets, Ariadne Issue 39 April 2004.
<http://www.ariadne.ac.uk/issue39/ndad/>

See also:
Kevin Ashley, Process re-engineering: a brief history of Government computing, NDAD newsletter No. 8, June 2000.
<http://www.ndad.nationalarchives.gov.uk/news/nl/pdf/ndadnews008.pdf>

Jon Agar, The Government Machine: A Revolutionary History of the Computer.
Cambridge, MA: MIT Press, 2003.

Data searching tips

How many schools in England are named after saints? The answer is 5630 - at least it was in December 2000, according to the data in the Register of Educational Establishments held by NDAD as CRDA/47. To find this out required the use of one of NDAD's least well-known and most powerful means of searching data: the MATCHES verb.

NDAD's data browsing facility allows you to use simple database query functions to view only certain records in a database, available via the

Selection criteria

tab when you view

data. All of the functions work by allowing you to filter data based on the values in specific columns, and most are simple tests: is that column greater or less than a specific value, or does it contain a specific word? Often this is enough to find data of interest, but when it is not, the MATCHES function is often of help. It's harder to use than some types of search, but the effort spent using it often helps us to find nuggets of interest or to count what we want quickly. This brief article just illustrates what can be done. A fuller explanation of usage is available in the NDAD help, at http://www.ndad.nationalarchives.gov.uk/help/data_browsing/regular_expressions.html

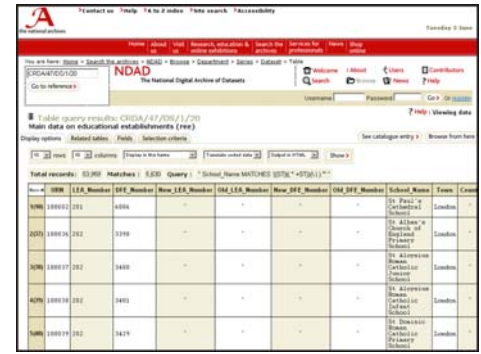
When we ask for a column to MATCH something, we supply a pattern - using a special language called a regular expression - to specify what we want to see in that column. In the first example, we are looking at the column called School_name in the table 'ree' (viewable at <http://www.ndad.nationalarchives.gov.uk/CRDA/47/DS/1/20/quickref.html>).

By Kevin Ashley, Head of Digital Archives

You might think we could just search for columns that contain the word 'Saint', but it isn't that simple. The word is invariably written as the abbreviation St, so we get St Thomas and St. Anne. It's sometimes followed by a dot, and sometimes it isn't. If we search for all names containing "St" we'll find far too many spurious matches, such as Hampstead School.

We need to look for schools whose names contain the letters "St", followed by a dot or a space, and either preceded by a space or by nothing at all. The query **School_name Matches** `'((ST)|(.* +ST))(\. |).*'` does this for us. The pattern breaks down like this:

<code>((ST) (.* +ST))</code>	The brackets enclose a set of alternatives, separated by a vertical line.
<code>ST</code>	The first alternative is ST, right at the beginning of the name
<code>(.* +ST)</code>	The second alternative is any characters, followed by at least one space, followed by ST
<code>(\.)</code>	After the ST, we want a dot or a space (the dot has a special meaning, and the preceding \ makes it stand for just a dot)
<code>.*</code>	And after the ST and its dot or space, any number of any characters can follow



The syntax is not immediately obvious, but it is quick to pick up. Using this has helped us exclude names like "1St Steps Playgroup" whilst picking up "Our Lady and St Philip Neri Roman Catholic Primary School", which would have been missed if we simply looked for all schools whose names began with "St".

Once you have the hang of it, you start asking questions you don't really need to know the answer to, such as:

How many head teachers have surnames that begin with a vowel ? (answer: 3611)	Surname MATCHES <code>[aeiou].*</code>
How many head teachers have surnames between 4 and 6 letters long ? (answer: 26,159 - very nearly half of them)	Surname MATCHES <code>[a-z]{4,6}</code>

Finally, a more challenging example and one that would be very difficult to do in any other way. Returning to our original example, how many head teachers' names contain vowels which appear in alphabetical order? (That is, names like Davis or Brown, but not names like Morris, where an o comes before an i.)

The answer surprised me, but if it intrigues you, you will have to find out for yourself. Go to

<http://www.ndad.nationalarchives.gov.uk/CRDA/47/DS/1/20/query.html>,



Continued from page 3.

pick the fields "School_name" and "Surname" to display and try the query:

Surname MATCHES

```
'[^aeiou]*a?[^aeiou]*e?[^aeiou]*i?[^aeiou]*o?[^aeiou]*u?[^aeiou]*'
```

The online help for using MATCHES explains how to build up patterns in a step-by-step way, and gives more information on the standards it is based on.

It's well worth reading if you want to know more about them.



Millennium Commission

Data and website preserved for the future

The National Archives is pleased to announce the release of key data from the Millennium Commission. There are 2 datasets in the collection, along with the website, which has been preserved separately. The Commission's website is still active at time of writing, though the Commission itself ceased to exist in November 2006. This highlights the value of preserving this resource in dataset form. The National Archives has taken an archive 'snapshot' of the Millennium Commission website. This is the first time that a website has been actively selected and archived alongside the data it presents and is a great example of how the UK Government Web Archive and NDAD can integrate and compliment each other.

The archived website is available here: <http://www.webarchive.org.uk/tep/15128.html>

Millennium Commission: Awards Scheme Database (AMIS) (Final database: 2004): CRDA/65/DS/1

The Awards Scheme, administered by the Millennium Commission, was the first of its kind to allow individual people to benefit directly from a National Lottery grant. 'Millennium awards' (i.e. small grants of between £2000 - £5000) were allocated to individual people for projects that were of recognised benefit to themselves and their community. Projects supported by the Scheme ranged from the setting up youth groups on inner city estates, to leading neighbourhood clean-up projects and establishing schemes to tackle racism. Since its launch in 1996, nearly 32,000 people from all over the UK have received grants of around £2,000 for projects.

This new dataset, the Awards Scheme database (also known as the Awards Management Information System or AMIS) served as the main repository of information related to the administration of the Awards Scheme. Through this electronic resource, the Millennium Commission was able to monitor, analyse and report on key aspects of the Scheme.

The dataset provides details of all the individual projects supported through the Millennium Awards Scheme from 1996 until 2004. It specifically holds data on the following key areas: Award Partner organisations; Award schemes; Award scheme finances; Award winners; and Award winner projects. The Commission worked with over 100 Award Partners (i.e. established charities and other non-governmental grant organisations) who operated their own Award schemes and were responsible for individually distributing the Awards. The Partners offered Award winners an important source of internal support, knowledge and expertise in order to help see their projects to completion. Award schemes were established under several major themes: Young people; Older people; Community; Arts; Environment; Health and Education.

From the first round of Lottery grants distributed in 1996 until the last Award approved in March 2004, the Millennium Awards Scheme gave out a total sum of £92.7 million. The dataset is important evidence of the UK Government's management and regulation of its grant funding, offering research insights into,

for example, the varying range and distribution of individual grants allocated across the voluntary sector and throughout many varied constituencies. It also provides some important case studies of individual projects that have the potential to facilitate wider learning and exchange. A number of Impact Studies commissioned by the Millennium Commission (2001-2003) have already explored the social impact of the Awards Scheme e.g. the extent to which it has provided benefits to individuals, irrespective of age, background or ability.

Following the dissolution of the Millennium Commission, UnLtd - The Foundation for Social Entrepreneurs - was chosen to act as a Trustee to continue the work of the Millennium Awards Scheme by ensuring that Awards continue to be made available to future generations. Visit their website at <http://www.unltd.org.uk/> See <http://www.ndad.nationalarchives.gov.uk/CRDA/65/quickref.html?jump=1> for further details.

Millennium Commission: Grants Database (PROFESA) 2006: CRDA/66/DS/1

The Grants database PROFESA provides details of grants awarded to all the organisations supported through the Millennium Awards Scheme, which allowed individual people to benefit directly from a National Lottery grant; the Millennium Projects scheme; and the Millennium Festivals scheme.

PROFESA contains basic information about all the awards, projects and festivals supported by the Millennium Commission, including applicant, project outline, cost, and progress.

The content covers three main areas: Awards, Projects, and Festivals.

Millennium Projects were the most visible part of the Commission's work, including large-scale buildings and environmental schemes accounting for over £1.3 billion of National Lottery money. The Millennium Commission only had enough funds to support a tenth of the applications it received. At the time of completion, there were over 215 Millennium Projects on around 3,000 sites throughout the United Kingdom.

We have also captured the Commission's **Image Library**, a collection of photographs of projects, awards, and festivals, which was an integral part of the original system.



Glasgow's Hogmanay from the PROFESA Image Library

The images were originally available to users of PROFESA and to the public via the Millennium Commission website,

<http://www.millennium.gov.uk>

It was possible for members of the public to search for Millennium Projects, Award Schemes and Festivals, retrieve a thumbnail of the relevant image, and a link to the larger image.

Both datasets are open to the public, with the exception of certain contact fields (for example recipients' names and addresses), which may represent a data protection risk if exposed. Parts of the documentation collection are redacted for the same reason.

See

<http://www.ndad.nationalarchives.gov.uk/CRDA/66/quickref.html> for further details.



Cardiff Millennium Festival from the PROFESA Image Library

More releases...

We have recently released a second dataset in the National Lottery Awards Database series. This provides information about awards to Good Causes made from the proceeds of the National Lottery. It captures information about grants awarded by the various lottery Distributing Bodies from data they supplied to the Department for Culture, Media and Sport (DCMS). The database is maintained by DCMS as a central source of information that can be interrogated in response to ad-hoc queries about lottery awards. The database is available in a limited form on the DCMS web site allowing visitors to the web site to search it remotely. NDAD holds two snapshots of the database: the first contains details of awards made between 1995 and 2001; the second covers grants made up to and including 2 September 2006. The 2006 snapshot contains details of more than a quarter of a million lottery grants totalling more than £18 billion.

These range from 24,000 grants to small community groups and individuals of less than £1000 to 133 grants in excess of £10 million for major capital projects. Some of these major grants were for projects that were intended to celebrate the Millennium such as the Eden Project in Cornwall and the Millennium Stadium in Cardiff.



The Eden Project in Cornwall

Examples of more recent large grants include £13 million to the Cutty Sark Trust in 2005

and £16.8 million to the National Museum of Scotland's Royal Museum Project. The National Lottery Awards Database series also relates to the Millennium Commission datasets detailed above and the Local Heritage Initiative project directory covered in our previous issue.



Millennium Stadium, Cardiff

See the Series Catalogue

<http://www.ndad.nationalarchives.gov.uk/CRDA/39/detail.html> for further details.