The Schools' Census and Digital Archaeology

Peter Garrod

UK National Digital Archive of Datasets
University of London Computer Centre
20 Guilford Street
London WC1N 1DZ
p.garrod@ulcc.ac.uk

Abstract

This paper outlines the efforts of the UK National Digital Archive of Datasets (NDAD) to reconstruct information about a major series of UK education datasets, the Schools' Census. Data from this annual survey of schools in England and Wales had been preserved by the Department for Education and Skills and its predecessors, but without the documentation required to place the data in context. Moreover, the data dictionaries which accompanied the datasets were incomplete, and failed to adequately explain the functions of many fields. NDAD circumvented these problems by reconstructing metadata for the datasets using examples of completed Census forms, published statistics, and other outside sources. The paper argues that NDAD's efforts are an example of what Ross and Gow (1999) have called 'digital archaeology': the rescue of neglected and damaged data resources, including cases where the threat is not to the data but to our ability to understand the data. The paper concludes that while 'data intelligibility' may indeed be recoverable, the cost in the case of the Schools' Census was high, and illustrates the practical limitations of 'digital archaeology'. This paper is a case study of the problems encountered by the UK National Digital Archive of Datasets (NDAD) when preserving a series of datasets known the Schools' Census, and how NDAD dealt with those problems. NDAD used reverse engineering techniques to reconstruct missing information about the data, thereby making them more accessible. Although we were not aware of it at the time, our approach to recovering the Schools' Census metadata was an example of what has been called 'digital archaeology'.

The paper explores the concept of 'digital archaeology', and shows how NDAD's work on the Schools' Census is a case study of digital archaeology in practice. In so doing, it addresses the need identified by Ross and Gow (1999) for more case histories about data loss and rescue.¹
The UK National Digital Archive of Datasets

NDAD is an archive of digital data, in the form of datasets, plus original documentation relating to datasets. The data and documents have been created by the departments and agencies of the UK central government. Since its inception in 1998, the archive has been operated by the University of London Computer Centre as a service provider to the Public Record Office (PRO). Our holdings form part of the UK national archives.

NDAD's primary function is the preservation of datasets and documents. In the case of datasets, this often involves converting the data from the formats (proprietary or otherwise) in which they are transferred, into formats more appropriate for long-term preservation. We produce catalogues of our holdings following UK and international archival standards, and provide outreach and support to users. Access by users to open data and documents is via NDAD's web site <http://ndad.ulcc.ac.uk/>, which in effect acts as the archive's search room. Two key archival activities which NDAD does not undertake are the appraisal of datasets and their selection for transfer to the archive. These decisions - and decisions concerning the access status of datasets - are made by the PRO, in conjunction with the departments which created the data.

Digital Archaeology

The concept of 'digital archaeology', as used in this paper, has nothing to do with archaeology per se, but rather is taken from the title of a report by Ross and Gow which forms part of a series of studies on digital preservation funded in 1997/98 by the Joint Information Systems Committee (JISC). The authors never define the term, but its meaning is implicit in their subtitle: 'rescuing neglected and damaged data resources'. Ross and Gow distinguish between two aspects of digital archaeology:

*Data recovery*, where the actual data have to be recovered or reconstructed as a result of damage to the media or the hardware on which the data are stored, or as a result of the obsolescence of the hardware or software required to access the data.

The *recovery of data intelligibility* - where the threat is not so much to the data, as to our ability to understand the data. This can result from weak links in the chain of preserving essential information (or 'metadata') about the data: for example, if encryption keys are not preserved, or if the documentation about encoding strategies is lost. As we shall see, the Schools' Census is primarily an example of this aspect of 'digital archaeology'.

Of the case studies in the report, one which illustrates the recovery of data intelligibility concerns the East German Kaderdatenspeicher: a database of government staff of the former German Democratic Republic. Although one
copy of the Kaderdatenspeicher survived German reunification, much of the information needed to interpret the data either did not survive, or had not been adequately recorded in the first place. For example, there was a lack of information about the different methods which had been used to express dates, and the methods of encoding the ID numbers assigned to GDR citizens.

The staff from the German Federal Archives who worked on the database had to reconstruct much of the metadata using external sources. These included paper files of the East German Council of Ministers, which had commissioned the database; codebooks which had been shared between the East German data centres; and documentation relating to the staff database of one of the East German ministries, which had been closely linked to the Kaderdatenspeicher. Even then, it proved necessary to employ some ex-GDR data staff in order to access information which only survived in their heads.

Ross and Gow use the Kaderdatenspeicher to illustrate that 'data without relevant contextual documentation has limited value'. The preservation of data must be accompanied by the preservation of the relevant metadata and documentation. Where a data source is of obvious historical importance, it may be feasible to recover data intelligibility, but at considerable cost.

The Schools' Census: background

The Schools' Census is an annual survey of schools in England which has been conducted by the Department for Education and Skills (DfES) and its predecessors. The early history of the Census is obscure. Although education statistics based on annual returns submitted by schools have been published at a national level since at least 1905, the survey known as the Schools' Census is thought to have started around 1945/46. Computerisation of the data had taken place by the mid-1970s. Schools in Wales were included until 1977, by which time the responsibility for carrying out the Welsh equivalent of the Census had passed to the former Welsh Office (however, for reasons which are unclear, Welsh independent schools continued to feature in the Schools' Census datasets after that date). By 2002 the Census had been phased out for most local authority maintained schools in England, and replaced by returns of pupil-level data derived from schools' management information systems. However, some types of schools - particularly independent schools - continued to submit Census returns as before.

The Schools' Census gathered data from individual schools on pupils (such as their numbers, ages, sexes); staff (numbers, types, whether full-time or part-time); classes (numbers of classes and numbers of pupils by type of class); and courses of study (types of examination courses and the numbers of pupils pursuing them). Fifty-nine da-
datasets covering the years 1975-1989 and 1993 were deposited in NDAD in various transfers between 1998 and 2002. This paper focuses on the datasets for 1975-1979, work on which was completed in September 2002. For most of this period there are four datasets per year, covering the main categories of schools: local authority primary and secondary schools, including 'middle' schools (which were deemed to be primary or secondary, for the purposes of the Census); local authority and direct-grant nursery schools; special schools and hospital schools; and independent schools and direct-grant grammar schools. There are no nursery school datasets for 1975 and 1976, for reasons which will be explained below.

The datasets for 1975-1979 were transferred to NDAD as simple files, one data file per dataset, without any database structure. With a few exceptions, the files were in the format of a proprietary software package known as QStat - the tabulation software used by the DfES's statisticians at the time of transfer. This in itself posed certain problems. Extracting the data from the QStat files was not a straightforward process. In the case of one dataset, NDAD was unable to export 1.3% of the records in their entirety; data items are missing from some of the fields in these records.

The data files were accompanied by QStat data dictionary files, which provided a number of vital items of metadata: the names of fields, textual descriptions of fields, and keys to encoded values. However, the department was unable to supply any other documentation which might help to put the data into context. No attempt had been made to systematically preserve specimens of the forms and instructions which were sent out to schools, and which formed the basis of the data in the datasets. The earliest set of Schools' Census forms which the department could supply (and which was only partially complete) was for 1997, although isolated examples were traced from 1993 and 1986. None of these forms was of any use for interpreting the datasets from the 1970s. It also became clear that paper files relating to the planning of the various editions of the Census had largely been destroyed at first review. In 2002 DfES records staff discovered two files relating to sweeps of the Census held in the 1990s, but no records have been traced from the earlier period.

The Schools' Census: problems

The department's failure to preserve documentation relating to the Schools' Census might not have been so significant, if the data dictionary files had been adequate. But when NDAD examined the data dictionaries for the 1975-1979 datasets, we found that they had a number of deficiencies. These included:
Truncated field descriptions. Many field descriptions end in the middle of words or phrases. Truncation appears to have been applied at or around 60 characters. In other words, descriptions which were originally longer than that are now incomplete. IT staff at the DfES suggested that the truncation might have occurred in 1991-1992, when a contractor was employed to migrate the Schools' Census datasets from one data analysis application to another. The way in which the process was automated may have caused the truncation. No copies are known to have survived of the data dictionary files as they existed before this event.

Duplicated field descriptions. All of the data dictionaries, to varying degrees, contain cases where the same field description is applied to a number of separate fields. The fields usually contain different data, indicating that despite the identical descriptions, they do not relate to the same thing. Truncation is often the most likely culprit: many duplicated descriptions are close to 60 characters (in some cases interrupted in the middle of words), suggesting that the extra text which originally distinguished one field description from another has now been lost. However, other duplicated descriptions are much less than 60 characters: e.g. the 1979 dataset for primary and secondary schools contains four fields with the description 'Classes size 31-35 taught by 1 teacher'. Such cases suggest that some of the deficiencies in the field descriptions may date from when they were originally created.

Unexplained encoding. The data dictionaries usually explain encoded values, but not always. For example, the datasets for primary and secondary schools all contain a field named PSMO (described as 'School type indicator'), which contains values of 1, 2 or 3: no explanation of these values is provided.

Unexplained abbreviations. At times the field descriptions can be highly telegraphic. While many abbreviations (e.g. 'FT' for 'full-time', 'FTE' for 'full-time equivalent') can be easily deciphered, others are less obvious. For example, the 1975-1979 primary and secondary datasets contain large number of fields which are described as relating to 'E' classes, 'N' classes, and 'O' classes. Nothing in the data dictionaries indicates the meaning of these categories. As a result of the work which NDAD has done to reconstruct information about Schools' Census, we know that these categories only relate to primary schools. 'E' classes were special classes for disabled pupils, 'N' classes were nursery classes, and 'O' classes were all other types of classes.

The Schools' Census: solutions

The Schools' Census is one of the largest series of datasets held by NDAD, and has obvious research potential, especially given that the statistics derived from the Census would have informed decisions about education policy. When the problems with the data dictionaries became apparent, we were concerned that this might limit use of the 1975-1979 datasets. Researchers would not have had a clear picture of the data if the only information provided was that
derived from the original data dictionary files. To circumvent these problems, NDAD decided to reverse engineer the metadata by turning to a number of outside sources.

We located examples of completed Schools' Census forms from the period, plus copies of some of the instructions for completing the forms. This was done with the help of Gloucestershire Record Office, which had preserved copies of the forms which schools sent to Gloucestershire County Council in its role as the local education authority (LEAs were required to collect the forms from LEA-maintained schools in their area, and send them to the Department of Education and Science). One local authority primary school in Essex, and a number of independent schools - contacted through an appeal to members of the Headmasters' and Headmistresses' Conference - were able to supply copies of their returns from copies held in their own files. Through these means, we amassed an almost complete collection of examples of the main categories of Schools' Census forms for each year in the period 1975-1981. Of particular value was the fact that the forms were \textit{completed} copies - not blanks. Using the identifiers on the forms, we could match the forms to the records for the schools in the datasets. This allowed us to see how the data in the datasets related to the information which had originally been supplied by the schools; how questions on the forms corresponded to fields in the datasets; and which fields represented derived fields, where the data had been calculated by the department after the survey.

We also turned to the annual volumes of the Department of Education and Science's \textit{Statistics of Education} series, in which the results of the Census were published during the 1975-1979 period. These volumes provided much useful background information: e.g. explaining the different categories of teaching staff, and the distinction between 'classes as registered' (classes on the registers of a school) and 'classes as taught' (classes being taught during a selected period of the school day). We validated the datasets by checking totals which we had calculated from the data against figures in the published statistics. The report of the Warnock Committee (1978) on the education of handicapped children was particularly useful for the special school and hospital school datasets, as it explained the statutory categories used at the time to classify pupils with disabilities.\textsuperscript{8} We also made considerable headway by simply carrying out our own searches and cross-comparisons of the data: e.g. we deduced how fields relating to pupil : teacher ratios had been calculated by the department using data in other fields.

This combination of research and reverse engineering clarified the functions of a number of fields with truncated, duplicated, ambiguous or otherwise defective descriptions. The results were incorporated into our catalogue descriptions of the datasets, in a way which distinguished between the original field descriptions supplied in
the data dictionaries, and the additional information created or reconstructed by NDAD. The following examples illustrate the process of metadata recovery:

(1) The 1978 and 1979 primary and secondary school datasets both contain a field with the following, obviously truncated, description: 'Nursery Assistants - Full-time women in nursery classes, Cl'.9 We know from the contemporary volumes of *Statistics of Education* that nursery assistants were employed in primary schools to help teachers in charge of nursery classes.10 Examples of primary school forms from 1978 and 1979 indicate that data was gathered on nursery assistants according to whether they were employed in nursery classes or other classes, and by whether they were 'Class I' or 'Class II' assistants.11 Based on the data and the other field descriptions, it is highly likely that the field relates to Class I assistants - and that this is the part of the original description which was lost. Instructions issued to local authority nursery schools in the 1976 Census indicate that Class I assistants were those with a certificate from the National Nursery Examination Board, or an equivalent qualification.12

(2) Each of the datasets for primary and secondary schools has four fields with an identical description - 'Classes of 1-5 pupils' - but different data.13 These fields were found to have a complex range of functions which changed over time. In the 1975 and 1976 datasets, the first three fields only contained values greater than 0 in records for secondary schools (excluding middle schools deemed to be secondary). The data appeared to relate to a section of the form in which secondary schools were asked to record information on classes as taught according to the age range of the class. The first field recorded the number of classes of 1-5 pupils where the age range of the class was under 14, the second related to classes of the same size where the class was aged mainly 14 and 15, and the third covered classes aged mainly 16 and over. The fourth field provided a total of classes of 1-5 pupils in all age groups, in secondary schools, and also contained data for middle schools (both middle schools deemed to be primary and middle schools deemed to be secondary).

The fields continued to play this role in the 1977-1979 datasets. However, they now had extra functions, indicated by the data and comparison of the data with the forms. In addition to classes as taught in secondary and middle schools, the fields also recorded information on classes as taught in primary schools, according to whether the class of 1-5 pupils was an 'E', 'N' or 'O' class. NDAD has a copy of the covering letter which primary schools received in 1977 with their Schools' Census forms, which gives a reason for this change in the data. It explains that the 1977 Census would require primary schools, for the first time, to record data on classes as taught, a requirement which previously had only applied to secondary and middle schools.14 We would never have guessed the true nature
of these fields, or the complexities of their roles in the datasets, if we had relied on the original field descriptions alone.

(3) We found, at times, that the original field descriptions could be actually misleading. Field AG_15 in the 1975 and 1976 primary and secondary datasets had the description 'Total pupils in N classes (primary schools only)'. In fact, it contained no data for primary schools, and related to classes as taught in secondary schools. In the 1977-1979 datasets the same description was used for the same field, and was partially correct: the field recorded pupils in 'N' (nursery) classes in primary schools taught at the selected period. However, it also continued to record the numbers of pupils in classes as taught in secondary schools (specifically, total pupils in classes aged mainly 14 and 15). We have no explanation for why a field description which appears to have been mainly relevant to the later datasets was used in the data dictionaries for 1975 and 1976.

As the above examples suggest, most of our efforts have focussed on detecting and correcting deficiencies in the field descriptions in the datasets for primary and secondary schools, which are by far the largest (the number of records ranges from ca. 26,000 to ca. 28,000). However, an interesting anomaly has come to light in regard to the 1975 dataset for independent schools. 618 of the 3155 records in this dataset are categorised as relating to 'independent nursery education establishments recognised as efficient'. This is far too many, because the published results in *Statistics of Education* indicate that there were only 8 schools in this category at the time.\(^\text{15}\) It was meant to be a residual category applying to only a handful of schools.

In fact, rather than being independent schools, these 618 establishments are likely to have been local authority and direct-grant nursery schools. Firstly, the number of schools matches the total of LEA-maintained nursery schools (612) and direct-grant nursery schools (6) in 1975, according to the published statistics.\(^\text{16}\) Secondly, NDAD holds a copy of a Census form which was completed in 1975 by a nursery school run by Gloucestershire County Council. There is a record for this school in the dataset; the data in the dataset matches the information on the form.\(^\text{17}\) So for unknown reasons, the 1975 independent schools dataset also includes records for public sector nursery schools, which have been misleadingly categorised. Although NDAD has not yet completed work on the 1976 independent schools dataset (transferred to NDAD in early 2002 to fill gaps in the series), there are initial indications that the situation is identical to that for 1975; i.e. LEA nursery schools are included in the same data file as the independent schools. The DfES was unable to trace any nursery school datasets for 1975 and 1976; we now know where those data are located.
Conclusions

At times it felt as though, the more we delved into the Schools Census metadata, the more problems we uncovered. This raises the question of how those problems came about, and what lessons can be learnt about pitfalls to avoid in the preservation process.

It seems likely that much of the original metadata was poor even at the time when it was created. Some fields were not adequately described, and there was a failure to create field descriptions which distinguished between the functions of certain fields. Too much reliance appears to have been placed on users in the department understanding the context in which the data were created. These initial problems were later compounded by breaks in the documentation chain. Documents - such as paper files and copies of survey forms - which might have been crucial for understanding the data were not preserved by the department, and have had to be unearthed from other sources. No record appears to have been kept of processes which affected the data, such as the inclusion of LEA and direct-grant nursery schools in the same dataset as independent schools, in 1975. There was also a clear failure to curate the data dictionary files in a way which would preserve their integrity. This was compromised when the field descriptions were truncated, and parts of the metadata were permanently lost.

The fact that these problems have occurred raises questions about the extent to which the DfES and its predecessors have actually used the data from older editions of the Schools' Census, once the original analysis of the data has been carried out and the results published. The department has preserved the datasets, but it is difficult to believe that the problems noticed by NDAD would not have been spotted (and perhaps corrected) if an attempt had been made to re-analyse the source data. Put bluntly: could this be a case of 'out of sight, out of mind'?

This paper has argued that NDAD’s work in reconstructing and creating metadata for the Schools' Census can be seen as an example of what Ross and Gow call "digital archaeology". Was it worth it? We believe so, but the costs of the approach which we have taken need to be recognised.

Mary Feeney, writing about the Kaderdatenspeicher, says that the efforts of the German Federal Archives show that while it is possible to recover information about data where the documentation is missing and incomplete, 'this is a painstaking, time-consuming - and expensive - exercise'. NDAD’s experiences bear out this conclusion. The 1975 and 1977-79 datasets were accessioned in a single batch in April 2000 (of the three 1976 datasets, one was transferred and processed in 1998, and the other two were received as 'strays' early in 2002). From April 2000 until September 2002, when work was finally completed, these 15 datasets were the responsibility of two members of
staff: an archivist (the author) and a data specialist. Although the work was intermittent and interspersed with other things, it took us much longer to process the datasets than we would have liked. NDAD certainly could not take the same detailed approach in every situation. Our contract with the PRO requires us to receive at least 36 datasets in the 2002-2003 contract year. 'Digital archaeology' is a luxury which it is difficult for us to afford.

It is easy to draw conclusions from cases like the Schools' Census about the need to create and preserve metadata, but such conclusions are of little use for the large amounts of legacy data which are flowing into archives like NDAD. We have to deal with material which was not created or curated following ideal procedures. Some legacy datasets which are defective may simply not be worth attempting to recover, even if recovery of the data or metadata is feasible. As more government datasets are identified by the PRO, it seems certain that we will continue to face hard choices about where to concentrate our efforts.