The Royal Statistical Society Working Group on Archiving Data

by Hilary Beedham[*]

Abstract
The Royal Statistical Society has recently established a working group to create standards for the collection and preparation of data in readiness for preservation. The working group consists of members of key organisations that are involved in both the collection and preservation of statistical material. The working group includes representatives from the private sector; the Office for National Statistics, (ONS); the Public Records Office, (PRO); the National Centre for Social Research, the UK National Digital Archive of Datasets, (NDAD); and the Data Archive. The representation of these organisations brings to the group a wealth of experience in both the collection and preservation of data from a range of sources including historical and administrative records, survey data and spatially referenced data.

The goals of the group are as follows:

- To define the extent to which materials, including questionnaires, data coding dictionaries, instructions for computations, working drafts and definitions of terms should be archived for future use.
- To establish a code of best practice for doing this
- To suggest how data creators, custodians and users can co-operate to ensure that best practise is observed.

The paper will explore the need for such standards and will describe progress to date with a view to stimulating debate and eliciting wider opinions on some of the key issues that the group will be addressing.

Why establish a working group on the archiving of statistical material?
In July 1998, the Royal Statistical Society convened a meeting, 'Archiving statistics: challenges and prospects'. The meeting was opened by Dr. Tim Holt, the Director of the Office for National Statistics and was well attended by over 60 data custodians and archivists, data producers and both public and academic researchers with interests in a diverse range of subject areas. In his introduction Dr. Holt recognised the importance of recording the processes by which statistics have been produced and acknowledged that the approach to preservation of such material within government has been inconsistent and varied between departments. Indeed, the Government Statistical Service had no overall policy on the archiving of the statistical material it generates. Dr. Holt also recognised the influential role of the Data Archive[1] in demonstrating what could be achieved in the preservation of such material and drew attention to the recent establishment of the National Digital Archive of Datasets (NDAD). He welcomed the meeting and hoped that it would lead to improved procedures that would be agreed between the various sectors with an interest: data producers; data custodians and archivists; and data users.

All of the speakers recognised the importance of preserving those materials that explain the research or data collection process in order to allow fully informed used of the statistical material for future historical use and secondary analysis. Consequently, the speakers all contributed to the key aim of the conference: the stimulation of discussion about which paper and electronic materials are needed for the informed use of published statistics and how these can be preserved. There was general agreement that such material should include the contextual material associated with a data collection exercise. The list of possibly relevant material is potentially extensive and can include original questionnaires and data; coding notes; instructions for the creation of derived data; working drafts and definitions of terms. It can even be extended to include policy documents explaining why a particular set of data were collected or compiled in the form they were and at a particular time. The discussion included not only the provision of material associated with the collection of statistics through surveys but also the preservation of material produced during the collection and collation of administrative statistics such as birth and death counts or unemployment figures.

A recurrent theme of the meeting was the recognition that data producers need to ensure the implementation of good practice throughout the data collection exercise. In particular, it was recognised that it is critical that this need be met from the earliest stages of any project involving data collection. Ideally, guidelines needed to be established as a

reference tool both for the funders of data collection exercises and their project managers, to enable them to build preservation requirements into their management procedures at the development stage of any project. The application of such guidelines should then facilitate the collection and collation of all the relevant contextual material in readiness for archiving and preservation, once a project is completed.

Thus participants at the July 1998 meeting were unanimous in calling for a need for a coherent approach and defined guidelines for data preservation. Speakers in turn noted the loss of historical material, the need to preserve the contextual material relating to data collection exercises and, associated with this the need to ensure that a complete historical record is captured. There was general agreement that although these are recognised and worthy goals, the lack of a set of standards and guidance on the collation and preservation of such material is a major factor in the failure to meet the goals. In summary, there was an acknowledged need for action in this area. Thus, with the support of the President of the Royal Statistical Society, an RSS working group was proposed which has subsequently been approved by the RSS Council. This group is now well established.

The RSS Working Group.
Following an invitation to those attending the July conference, to express interest in participation in the group, the inaugural meeting took place in October 1998. Its composition reflects the breadth of interest demonstrated at the conference itself, including representatives from the spheres of custodian and archivist, data producers and data users. Thus, the committee includes data providers from the Office for National Statistics (ONS), the Home Office and the National Centre for Social Research. Data custodians are represented by the Public Record Office, (PRO), the UK National Archive for Datasets (NDAD), the Data Archive and Qualidata and users through the dissemination role played by the data custodians.

Terms of Reference.
The first meeting agreed the following terms of reference that have been subsequently agreed by the RSS.

- To define the materials, including questionnaires, data coding dictionaries, instructions for computations, working drafts and definitions of terms that should be archived for future use.

- To suggest how data creators, custodians and users can co-operate to ensure that best practice is observed.

- To establish a code of best practice for achieving this.

Existing literature
Subsequent meetings have been held in November 1998 and in February 1999. At the first of these the group established the need for a project plan which is now in place The first task of the working group was to discover existing material that might be relevant and to review this. We have set ourselves a fairly daunting task since the breadth of statistical material under consideration is great. We are considering, amongst others, survey material, administrative records such as health records, observational data such as road traffic counts, census material and geo-coded data. The inclusion of contextual material extends the range of material significantly and we had extensive discussion about precisely what material needs to be preserved.

There was a general recognition that there are a number of initiatives which may well feed into and influence the work of the group and that there are a number of organisations which have, over many years, established their own guidelines for data collectors. It would be foolish to ignore this work: there are no benefits to re-inventing the proverbial wheel. Nor have we any desire simply to reproduce any existing document that potentially provides the standards in a given area. During late December and early January, therefore, members consulted with colleagues and trawled the Internet for papers and documents. A list of relevant documents was then compiled and each member was allocated material for review.[2]

We approached the review systematically, asking the following questions for each document:

- What is its purpose?

- Who is the audience?

- What type of material has been targeted?

- How detailed is the information?

- Is the document prescriptive or for guidance only?

The review confirmed that there is a lot of material available that relates either to the deposit of material for further use or to the preservation of such material. There is also a great deal of technical information available relating to file and transfer formats and a lot of information relating to areas such as respondent confidentiality and copyright. There is also a significant body of work that gives guidance on contextual material. All of this work has been carried out by experts in the particular field and cannot be ignored. For example, the ICPSR Guide to Social Science Data Preparation and Archiving, was described during the review as “so sensible and universal, and the manner of its offering so persuasive that it could be accepted as a ‘mandatory’ standard”.

Following this review, it was clear that although much has been written about the preparation of statistical material for preservation, there is no one document which offers a complete set of guidelines for all types of material and all data creators. Whilst many, such as the ICPSR guidance, provide sound advice, each has been designed for a select community of data providers. Understandably, then, documents tend to emphasise either the particular data type with which the organisation is concerned, for example, qualitative material, or information, such as acceptable deposit formats or media, which are specific to the organisations own procedures. A further distinction was evident in the material whereby existing recommendations can be loosely divided into two types. The first is those documents provided by institutions with whom data creators have a legal or contractual remit to deposit data, and the second are those that have been written by groups or institutions, only some of which have a custodial responsibility and are acting in an advisory capacity only.

Problems and resolutions for the working group

When determining the style, structure and content of the guidelines, a number of points were agreed to be self-evident.

There is agreement amongst the group that the most efficient and beneficial use of standards is to apply them at the data creation stage but we also recognise that preparing material to agreed standards for archiving imposes a financial burden on the data provider. These costs are incurred whether or not the provision of the material is mandatory or voluntary and whether or not the provider is a public or private organisation. It is a burden that is likely to affect the quality and quantity of material that is prepared for preservation and is regularly cited as an obstacle to archiving. One of the greatest challenges that the working group will have to overcome is the need to convince data producers that they will accrue significant benefits from the preparation of their material to agreed standards.

The group expects to recommend three approaches to this problem. Firstly, we are planning to include a section in the guidelines that will give advice on the potential costs incurred by preparing data for preservation. It is hoped that this will encourage those who commission data collection exercises to build realistic costing for preservation into their budgets from the outset. If we can achieve this, data collectors should be relieved of the budgetary constraints imposed when they are expected to send data for archiving. Careful thought will be needed in the presentation of this advice. Our current thinking is that it will need to be presented in terms of man-hours, for example, since information based on currency costing will not be relevant across national boundaries and will quickly become outdated.

Secondly, the group will seek methods of promoting the known benefits and often hidden cost savings of preservation of statistical material. For example, data collection is becoming increasingly costly. It is also becoming increasingly frequent as a means of discovering more detail about social and economic phenomena and, in the case of survey data, for example, respondent resistance is said to be an increasing obstacle to effective data collection. It is only sensible then to ensure that we get the maximum benefit from the statistical material that is collected. We can do this by promoting the re-use of material, for example where time-critical data are not essential.

Thirdly, the group does not expect to place the entire cost burden onto the data commissioners and collectors. Some of the costs will have to be bourne by the custodians. We expect that as long as standards can be agreed and adhered to, data custodians will take some of the responsibility for converting material to the archival format. One possible way forward with this is to capitalise on the Data Documentation Initiative[3] by making maximum use of the data type definition. This should enable data custodians to write and share conversion routines to convert data into a preservation standard. Work of this nature is currently being done at the Data Archive, the University of Essex and as part of the ddi/dtd beta test. With this in mind, the working group is currently reviewing the dtd as a potential generic starting point for a set of guidelines.

The group has also been involved in discussion about the presentation of the standards. Our aim will be to present the standards in a way that is acceptable to a wide audience and we must avoid the danger of producing a volume that is dense and not easily navigated. Current thinking on this is that it may be appropriate to provide an overview document that contains very basic guidelines with information that is relevant to the providers of all types of data. This document might include information on providing cataloguing records and on the costs of archiving. It might also include an index to sections of a fuller document or references to a series of individual documents that relate to specific types of data or cover complex topics such as respondent confidentiality, in depth.

Review of the Data Documentation Initiative (DDI)

Having identified the DDI as a potential, generic starting point for a set of guidelines, the working group is now reviewing the associated DTD for it's suitability for this purpose.

The review is at an early stage but the DTD does have a number of acknowledged strengths and the group felt that it might provide the core for a set of guidelines that could be applied across data types. Its greatest strength is that it is intended that the DTD should be accepted as a standard. Combined with the composition of its' committee and the inclusion therein of representatives from several continents,

it is realistic to think that the standard can be agreed internationally. The committee also comprises recognised experts in the field and the Initiative is being led by ICPSR, which the working group has already identified as providing excellent material in the field.

It needs to be noted, however, that at this stage the DTD does have weaknesses as a potential standard for the purposes of the working group. In particular, it has been designed as an exchange mechanism and at this stage it is not clear whether it can yet be used as an archival format. Nevertheless, as part of its current beta testing exercise, the DDI committee has invited comments on its potential use as an archival format. There is also ongoing discussion about how well the DTD accommodates aggregate data files and hierarchical files. This is also of concern to the working group but the DDI committee is actively considering it and the Data Archive is directly involved in the development of the DTD in these areas. The links between the working group and the Data Archive will enable the working group to keep up to date on progress and developments in these areas.

The group has three advantages that we anticipate will work in our favour and allow us to contribute to the future development of the DTD to accommodate a wider range of statistical material than it does at present. Firstly, the status of the group, with RSS support and a highly professional and respected membership, will allow us to speak with authority and make informed and respected representation to the DDI committee where we consider the DTD might be developed to meet the required standard. Secondly, the group is fortunate in having members whose interests cover a broad range of data types and statistical interests. So, for example, we have one member with an interest in Geographical Information Systems who is reviewing the DTD for its appropriateness to GIS material. Another member has an interest in textual material and open coded questions whilst a third is interested in individual level data where respondent confidentiality is a particular issue. Finally, the Data Archive is represented on the DDI committee, which has welcomed a dialogue with the working group and is keen to draw upon its expertise.

Possible ways forward.

We are not yet in a position to make definitive statements about the final model for the standards although we are clear on some issues. We do want to capitalise on the significant amount of high quality material that already exists. We also want to take account of the budgetary constraints of data producers and we want to offer standards that can be realistically adopted and maintained.

Nevertheless, it is possible to make some suggestions as to how the standards recommendations are likely to develop. It is most likely that we will adopt a position that there is already a great deal of material that could, with agreement from interested parties, be adopted as part of a formal set of standards. The group might then produce a document that directs producers, custodians and users of different types of material to organisations that have established an appropriate and agreed standard.

A second approach might be to encourage the expansion of an existing standard, such as the DTD, to include areas that it does not yet support.

In practice it is most likely that a combination of these two options will be adopted.

For more information on the RSS working group or if you would like to discuss the work of the group, please contact the author by email on beedh@essex.ac.uk.

1 The Data Archive is housed at the University of Essex, Wivenhoe Park, Colchester, England, CO4 3SQ. http://dawww.essex.ac.uk

2 A list of the documents covered can be obtained from the author at the University of Essex or email beedh@essex.ac.uk

3 DDI – co-ordinated by the International Consortium for Political & Social Research at the University of Michigan.

* Paper presented at the IASSIST Conference, May 19, 1999, Ryerson Polytechnic University, Toronto, Ontario..
Hilary Beedha,The Data Archive, The University of Essex, UK.