Managing metadata: issues and approaches

by Joanne Lamb [1]
Centre for Educational Sociology
University of Edinburgh

Introduction

Over the last few years there has been an increasing interest in metadata and the role it plays, both for cataloguing purposes and to give secondary analysts better understanding of the data they are using. The terms 'metadata' and 'metainformation' are currently used in a variety of contexts. This paper examines some of these contexts and discusses the kinds of metainformation that are relevant to different kinds of data and to the different uses to which that data might be put. Drawing on experience gained in examining data gathered by social science surveys (micro-data) and aggregated data produced by official statisticians (macro-data), it discusses the issues involved in the initial capture and maintenance of these various types of metadata. Particular attention is given to ways of using metadata to inform cataloguing systems and on-line information, and to the interfaces between various metadata holdings. The paper concludes by considering the impact of a more rigorous demand for metainformation on the data capture process.

Background

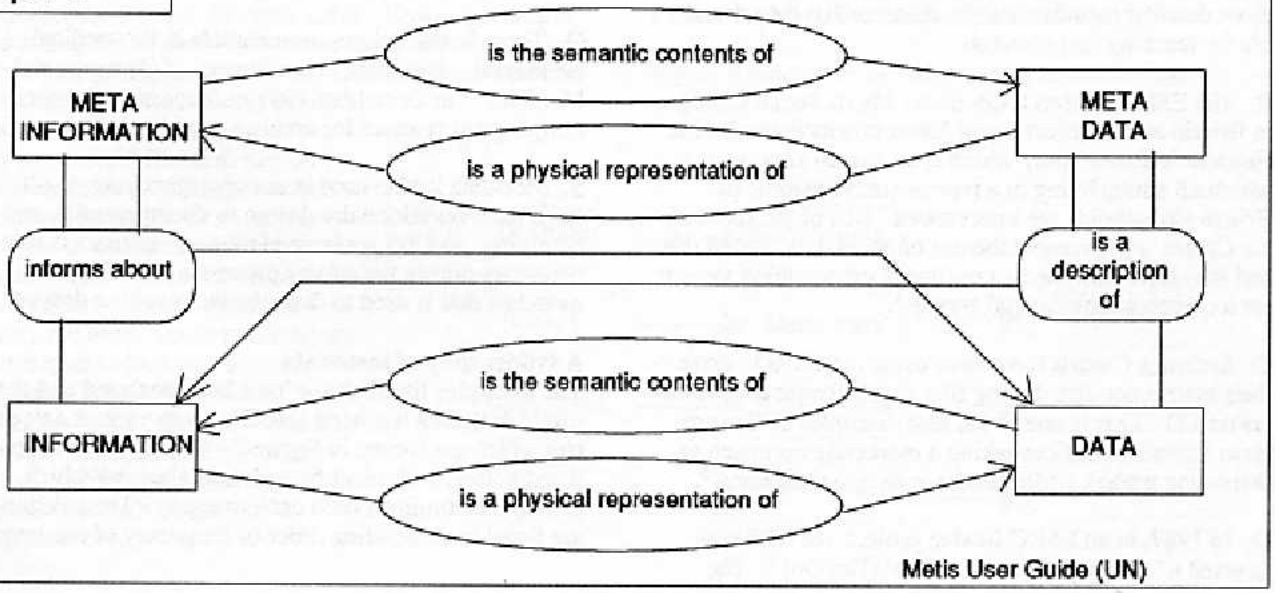
The Centre for Educational Sociology (CES) is a Research Centre of the UK Economic and Social Research Council, situated in the University of Edinburgh. We have been collecting survey data since the 1970s and have been using metadata in survey processing since the mid 1980s. The initial impetus to develop the use of metadata came from a need to rationalise the documentation of a series of related large and complex surveys collected between 1976 and 1985. In 1990 we became members of a European Commission shared-cost project, with partners from Belgium, England, Luxembourg, and Spain, to construct an interface to statistical information.

Metadata and the role it plays

We first need to determine what metadata and metainformation are. At a superficial level, metainformation is 'information about information', and metadata is 'data about data'. This is not as frivolous an explanation as it may seem. Information about information can be seen as a never ending hierarchy, a point that will be discussed later. The METIS User guide[2] has a diagram which succinctly illustrates the relationships between these concepts (Figure 1).

However, we need to be more precise than this if we are to consider the use of metadata. To my mind the basic characteristics of metadata are as follows: it describes data that exists either physically or conceptually; it is

Figure 1:



Metis User Guide (UN)

stored in a computerised form; and it is relevant to a particular task, providing aid in either the processing or the understanding of data. This is a minimalist definition that is deliberately all embracing, but the emphasis is on computerised data and on the fact that the metadata is either used by a computing system, or presented to a user to help in the use or interpretation of the data.

Having accepted that we want to use this metadata, we come to two further questions. How can we make it useful, and how do we represent it in computerised form in order to make it useful?

Contexts for metadata

Before answering these questions, we want to look at contexts for metadata. By context we mean the circumstances under which the metadata is to be considered, i.e. its relevance to a particular task, providing aid in the processing of some data, or the understanding of some data. To illustrate this, we will look at some particular examples of the use of metadata and then identify some categories that are used.

Examples of metadata

This section is composed of a list of some users of metadata, together with a short summary of that usage. The list is incomplete, and the examples have been chosen to show the range of applications using or discussing metadata.

A. The ESRC Data Archive has a remit to make economic and social data accessible for analysis, and therefore they need to describe data at a macro level. Their Bibliographic Information Retrieval Online (BIRON) system [3] allows the user to search for studies in terms of subject matter, and they are also interested in providing more detailed metadata for the datasets that they distribute for teaching and research.

B. The ESRC Research Centre on Micro-Social Change in Britain at the University of Essex conducts the British Household Panel Study which is an annual survey in which all adults living in a representative sample of British Households are interviewed. Part of the remit of the Centre is to promote the use of panel data, and to this end they have designed an on-line documentation system for a complex longitudinal survey [4].

C. Statistics Canada have been using metadata to drive their system for distributing files and software to customers on CD. This is one of the best examples of Government Statistical Offices taking a marketing approach to delivering usable, understandable data to their users [5].

D. In 1987, in an ESRC funded project, the CES constructed a 'documentation database' (DocDb) [6]. The focus of this project was on tracking the way that questions and associated variables had changed over time in a number of surveys which needed to be combined for analysing trends.

E. The EISI project[7], was funded by Eurostat under the DOSES initiative [8]. The object is to help official statisticians use unfamiliar data. We approached the problem from the users' view, and produced an online guide to the existing paper documentation produced by official statistics offices. A further development could be to access the statistical data itself.

G. Because of their special needs, Geographic Information Systems (GIS) and their metadata have been the subject of considerable study. A conference on Metadata in the Geosciences [9] was held at the University of Loughborough in December 1990.

L. The Edinburgh University Data Library is also concerned with macro metadata. A part of the University Computing Service, the Data Library has strong links with the University Library and has played an active role in the development of catalogue standards for computer files [10].

M. The METIS user guide is a UN publication providing a formal definition of metadata for statistical information. It was produced in 1989 by the Statistical Computing Project for the United Nations Development Programme and the Economic Commission for Europe. The aim of the METIS group was to work out procedures for describing existing data within statistical information systems, to develop a tool for the users of the system, and to develop a tool to serve the needs of statistical information management systems.

O. There is also interest in metadata in the medical profession. An article in the Journal of Occupational Medicine [11] in December 1991 outlines, among other things, good practice for archiving epidemiology studies.

S. Metadata is also used in survey processing, in all steps from questionnaire design to documentation and archiving. In CES we have identified metadata that is necessary during the survey processing as well as metadata that is used to describe the resulting datasets [12].

A typography of metadata

The examples listed above have been analysed and their use of metadata has been classified into various categories which are shown in Figure 2. The categories are listed below, with identifying letters showing which example institutions used each category. The categories are listed in descending order of frequency of occurrence.

Figure 2

Indicators/Variables/Codebook	A, B, C, D, E, G, M, S
Data files/access	A, C, D, G, L, M, O
Questions, Questionnaire	A, B, C, D, E, O, S
Study/Research design	A, B, C, D, G, L, O
Publications/Bibliography/Report	A, B, E, L, M, O
Keywords/Domain	A, B, E, G, L
Statistical analysis/information	A, B, L, O, S
Coding notes	A, L, O, S
Spatial elements	A, E, G, L
Surveys	B, D, E, M
Documentation	L, O, S
Output tables	E, M, O
Populations	E, L, M
Rules & Algorithms	E, M, O
Statistical units	E, L, M
Classifications	E, M
Methodology	G, L
Organisations	E, G
Periodicity	E, L
Time series	E, M
Users	A, M
Correspondence	O
Glossary	M
Interviewer notes	A
Language	L
Licencing	C
Measurement instruments	O
Protocols	O
Sample frames	E
Thesaurus	M

This table a number of questions. There is a broad consensus on 7 of the 30 items listed. However, there is also a long tail of about half of the items which have only one or two mentions. We have to ask ourselves why this is the case. Are these items solely of specialist need or is it the case that they are more difficult to capture? Are these concepts that have been identified in theory, or have they been put to practical use in a working system? There is now a history of some twenty years of metadata usage, and more recent publications have begun to look at a theoretical approach. Svein Nordbotton [13], Bo Sundgren [14], David Hand [15] and K. A. Fröschl [16] have all recently written papers which concentrate on a conceptual approach rather than a pragmatic one.

When is metainformation relevant?

Having put the study of metadata in context, we return to the practical questions facing practitioners, and ask what relevance this metadata has to them. In response to the question 'Who needs metadata?'. I would argue that all users of data do: the producers, the users and the 'brokers', i.e. the archivists and librarians. However, different kinds of users have different needs. We also have to ask what metadata is needed, what its function is, and how the metadata is obtained. I would submit that its function is threefold, to aid documentation, to improve the production of data, and to inform all users of the information relevant to their particular task.

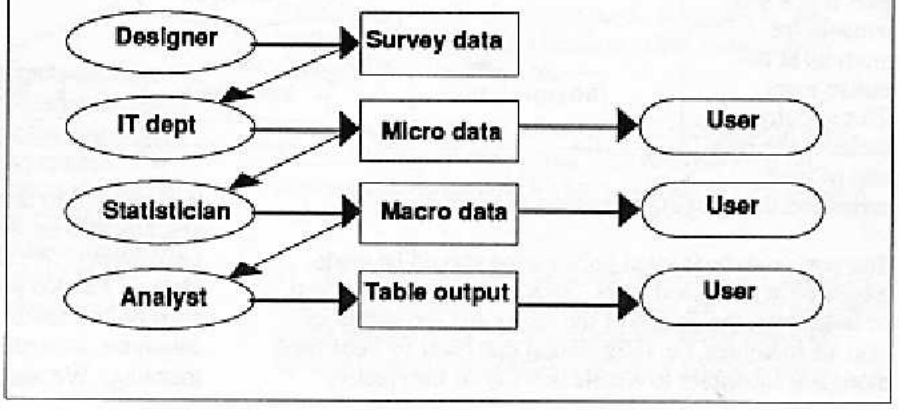
Producing the data

In the following section, I am concentrating particularly on survey data, and have identified four groups or sections who contribute to producing final data. At each stage the data may be available to end users (i.e. people not involved in the production process and who therefore know nothing about the data). Each of the groups contribute to the metadata by supplying some information about the data that is useful to the end user.

Figure 3 represents this process. The data is processed by one section after another, and the metainformation about the data is extended by each section. The designer conceptualises the survey and causes the raw data to be collected. The IT department computerises the data at the micro level. The statistician aggregates the data at the macro level, and the analyst defines printed tables.

Figure 4 represents the information about the data

Figure 3:



gathered at the micro level. The first group is identified by the Designer. This person or team identifies the research question, designs a suitable instrument for collecting relevant data, and has knowledge of the background to the problem. The contribution can be summed up as sample design, questionnaire design and purpose of the survey.

The second group is identified as the IT department, but includes the administration of the survey, coding and data capture as well as design and identification of datasets. This group provides information on practical aspects of the survey process, response rates, coding notes, anomalies discovered, type of analysis package, codebook information, availability and whereabouts of data. This information is summarised as codebook, physical description of the data set and technical report.

Figure 4:

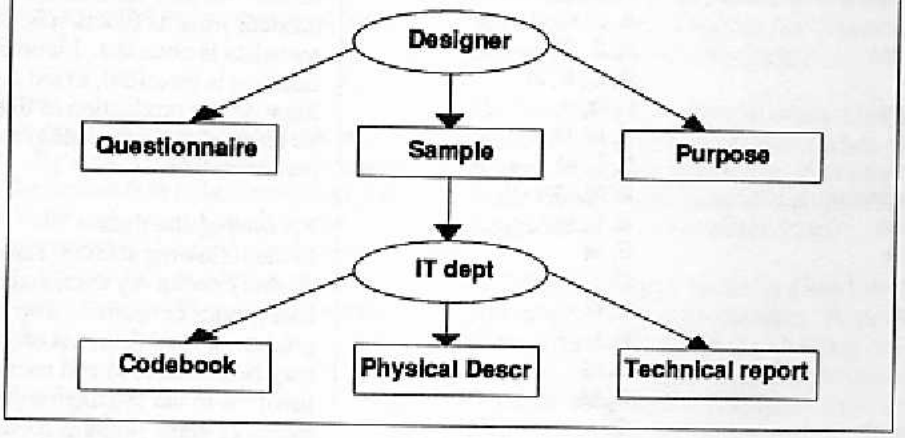
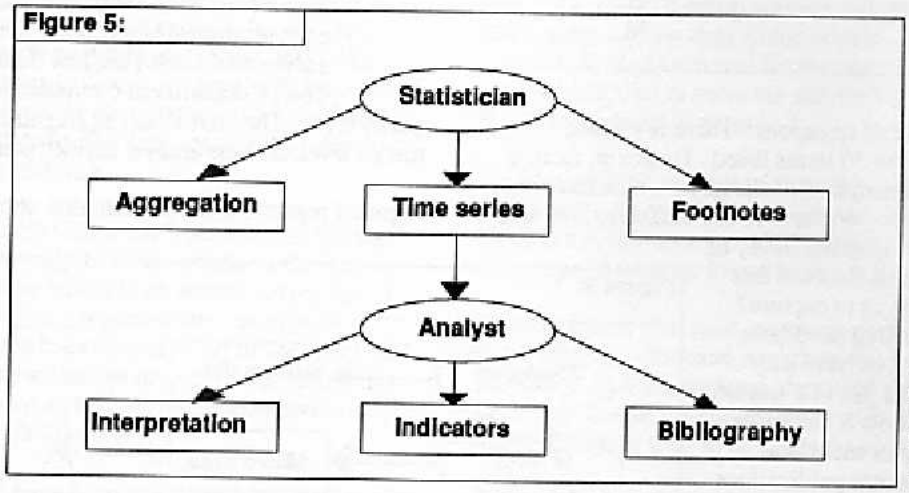


Figure 5 represents the information about the data gathered at the macro level. The group represented by the statistician aggregates the data, defines indicators and derived variables and puts it in a form suitable for analysis at the macro level. This activity includes the creation of time series and the merging of several surveys.

Figure 5:



The analyst defines what information should be made available in published form. This includes the selection of indicators, the design of the tables and the identification of footnotes, i.e. information that must be published alongside the tables to enable them to be interpreted correctly.

Representing the metadata

So far we have discussed some examples of uses of metadata and have seen that a crude classification system can be applied to it. We have also seen that some classes are more widely utilised than others. We then looked at the process of collecting survey data and identified the kind of information (metadata) that is associated with the four main steps in the process. We now need to consider how this information can be captured in a usable computerised form. In order to do this effectively we need to establish a common framework. This means some more fundamental work on the nature of metadata itself.

At the beginning of this paper 'information about information' is described as an infinite hierarchy. The potential amount of information is overwhelming. It is therefore important to bring order to the confusion. Metadata needs to be classified, ordered and associated with its function. Only then can we maximise its potential. To do this we need to draw on the cataloguing skills of librarians, and also on the concepts taken from software engineering. We need to consider essential models. The Object Oriented paradigm which associates data and purpose is a useful way of approaching the problem. In summary, we need to categorise metadata by use and by meaning. We also need to consider how rapidly each

category of metadata might change.

Maintaining metadata

The preceding section discussed theoretical and conceptual questions associated with metadata. This section looks at the more practical issues. We highlight a number of questions about the nature of metadata and the implementation of systems using it.

If we accept the notion of an infinite hierarchy of information at the intellectual level, we still need to examine the concept pragmatically. We need to consider if there are practical reasons for distinguishing between data and metadata, for example in relation to existing analysis packages, and whether there is an intuitive difference suitable for the kind of data we are handling, i.e. is there some kind of metadata (e.g. unit of measurement) which humans expect to see 'closer' to the data than others.

A second consideration is that of physical storage. Should data and metadata be stored in one system, or are the structures and uses such that they are better held separately? If they are stored separately, how do we ensure that the data and metadata are kept consistent?

There are practical questions concerned with the definition of a dataset. Is there such a thing as a definitive dataset? Should the data be considered independent of the statistical package in which it is held? We also need to consider metadata for related datasets and suites of datasets. Are we describing physical or conceptual datasets? Can we have subsets that are valid studies? Can we merge datasets into valid studies? There are special problems associated with time series and longitudinaldatasets. For example, how do we describe changes in the real world?

Next we need to consider upgrades and versions. We need to know how to handle modified or restructured data. What happens when we add new derived variables? Should a dataset be static or dynamic. If we resolve these problems, we need to consider how the documentation can reflect these decisions.

Conclusions

In conclusion I first want to reflect that metainformation is probably more difficult and more expensive to capture than the data it describes. It is also the case that metainformation is generated at all points in the system. For these reasons, metadata essential to all users should be identified, and captured once at the most suitable point in the process. Having captured this expensive commodity, metadata should be made to work. It should also be held in a flexible structure so that it can be transferred between systems.

Finally, there is a great deal more thinking to be done on the nature of metadata and how it can be used to ensure that the data we all use is accessible and can be interpreted with the minimum of error. This means that time and resources need to be given to the theoretical and conceptual questions that are unresolved. The study of metadata needs to be seen as a valid intellectual activity in its own right and not only as a by-product of a particular statistical system. Only then will we be able to implement standards which can have sufficient validity to be widely accepted in the heterogenous and fast moving world that data libraries are trying to serve.

1 Presented at IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993.

2 United Nations Development Programme (1989). Users Guide to Metainformation Systems in Statistical Offices.

3 Annual Report 1990-91 of the ESRC Research Centre on Micro-Social Change to the Economic and Social Research Council.

4 The BIRON System: the Archive's on-line catalogue and subject index. ESRC Data Archive University of Essex Wivenhoe Park, Colchester, Essex CO4 3SQ, UK.

5 Deeker, G; Scott Murray, T.; Ellison J. (1993) On providing client support for machine readable data files Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg

6 Design of a conceptual model for a documentation database, ESRC R000231293, 1989-1990.

7 Expert Interface to Statistical Information, EEC DOSES programme project no B34, 1990-1992

8 Development of Statistical Expert Systems: information pack, EUROSTAT, Batiment Jean Monnet, Plateau de Kirchberg, B1907, Luxembourg, March 1989.

9 Medyckyj-Scott, D., Newman, I., Ruggles, C. and Walker, D.(1991).(Ed) Metadata in the Geosciences Group D Publications Ltd.

10 Burnhill, P. (1991) Metadata and cataloguing standards: one eye on the spatial (Eds Medyckyj-Scott, Newman, Ruggles and Walker). Group D Publications Ltd 13-37

11 The Chemical Manufacturers Association's Epidemiology Task Group (1991) Guide-lines for Good Epidemiology Practices for Occupational and Environmental Epidemiological Research, Journal of Occupational

Medicine, Vol. 33 no 12

12 Lamb, J. (1993) Metadata in survey processing Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg

13 Nordbotton, S. (1993) Statistical Meta-knowledge and data Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg

14 Sundgren, B. (1993) Modelling meta-information systems Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg

15 Hand, D.J. (1993) Data, metadata and information Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg

16 Froschl K.A., (1993) Towards an Operative view of semantic metadata Proceedings of the Statistical Meta Information Systems Workshop, Luxembourg