METADATA FOR TRANSPORT RELATED DATA – PREPARATION & PRESENTATION & DISSEMINATION

Angelika Schulz[1]

Summary

In terms of statistical data the importance and usefulness of reliable metadata is undisputed and generally accepted. In any case, a profound, professional analysis of empirical data, in particular secondary analysis, demands extensive documentation about the survey itself and the associated data set. This documentation should include background information, sample design and procedures, the applied data collection methods, coding schemes, technical file formats and so forth. Comprehensive metadata are even more important if a certain database is supposed to be used across various disciplines. For example, transport data are not only relevant for transportation research but also for urban and regional studies. From the user perspective a standardized description format is extremely helpful. It not only supports efficient information retrieval via search engines, but also allows a direct comparison of different data sources. Therefore metadata has to be prepared, presented and disseminated in a more or less formal way.

This presentation will focus on the whole process of preparation, presentation and dissemination of metadata within the "Clearing House for Transport Data and Transport Models"[2].
The initial preparation of metadata is a complex time-consuming task. An overall metadata structure has to be defined (for example the DDI metadata standard), already available explanatory material has to be collected, and the relevant information has to be extracted. If necessary, related datasets has to be checked or even recoded. For presentation and dissemination of metadata a web interface was developed. Especially for the presentation of survey data the NESSTAR system is used.

General Background

The "Clearing House for Transport Data and Transport Models" is a rather new system of information access within the field of transport research. It is operated by the Institute of Transportation Research, which is one out of about 30 institutes within the German Aerospace Center.

Transportation research is a highly complex field of research. There are several subdivisions and uncountable research projects, all of them producing results and the respective publications. One specific research outcome are data - in the broadest sense -, which, in turn, are the base for further research, transportation planning processes or political decision making.

The problem is not a lack of data. Actually, there are huge amounts of all kind of data. The point is, that often virtually anybody - except the data owner himself - knows about their existence and possible availability. Even though a certain survey or modeling approach is known by the community, its third party

[1] German Aerospace Center/ Institute of Transport Research, Berlin;
contact: Angelika Schulz, angelika.schulz@dlr.de.

[2] The Institute of Transport Research at the German Aerospace Center operates the clearing house as a non-profit archive on behalf of the German Federal Ministry of Education and Research and in close cooperation with the German Federal Ministry of Transport, Construction and Housing.

usability might be limited due to insufficient documentation, or - if the documentation is well prepared - comparability might still be complicated due to varying formats of documentation.

Facing this unsatisfactory situation, the clearing house intends to bridge this information gap with a central knowledge base, focusing on transport related information. Generally, the clearing house will facilitate both publication and dissemination of relevant research outcomes. Since detailed metadata are absolutely necessary to carry out further professional analyses, its primary task is to provide such information in a standardized format. The provided information should be free of charge and easy to access. Therefore, access via an Internet portal is recommended.

Focus on Transport Related Data and Models

As mentioned before, data are one basic input for transport research and related disciplines such as geography, regional or urban planning. That could be, for example, empirical survey data on every day mobility. Another, more technical type of data are traffic flow data. These are collected in measurement campaigns using technical devices like induction loop detectors or even global positioning systems.

To document such data properly, various components have to be provided:
At first, there are large data sets, normally saved as flat files or in specific statistical software formats like SPSS or SAS. Secondly, these data sets are accompanied by more or less extensive explanatory material, such as codebooks, questionnaires, written final reports and so on. Of particular importance is the information on holdings and the availability status. Finally, there might be related publications or even project websites.

Somewhat different is the situation regarding transport models, which is the second field of interest. Modeling can be considered as basic method in transport research. Models are used, for example, to forecast traffic flows or to anticipate a certain traffic demand in a specific region.

The documentation of transport models has to include the following items:
First, a basic description of the underlying concept, and - if available - formula or source code. Since both the development as well as the use of transport models is dependent on actual data, information about required input data has to be provided. Often - based on the specific model - simulation tools have been developed. Particularly with respect to commercial products, notes on the availability status or license conditions are of importance. Finally, there might be related publications or project websites as well.

Presentation

To present all this information, a website has been developed. It consists of several components:

Its main part is a database for transport data. It includes a number of relevant studies concerning mobility behavior and the respective metadata.

Secondly, a database for transport models is under construction, which will mainly include metadata as well. An early prototype is already available.

A special feature will be a test suite for transport models. It will offer the possibility of testing a certain choice of models with particular reference data sets. There is also a first prototype available.

An additional component will be a directory of cross references leading to additional information sources such as other archives or research institutions.

Information Retrieval and Access

Necessarily, the available information has to be structured to facilitate both easy retrieval and access. Therefore, some catalogues have been developed so far. They are supposed to support all kind of searches such as for specific topics, regions or time frames. Additionally, a site specific search engine has been implemented.

The result of a search will be a list of available metadata open for browsing and download. If available, the related datasets are accessible via NESSTAR[3], which is a software environment to publish statistical data.

Provision of Metadata

The main part of the provided information will be metadata. The provision of metadata includes several steps, which are preparation, publication and dissemination.

The basic input will be data as output from empirical surveys or measurement campaigns. In most cases, these data will be accompanied by the original documentation, such as written reports and codebooks. Sometimes, related publications are already available.

What sounds pretty simple, becomes complicated when getting into details. In terms of data, one have to deal with various software formats such as flat ASCII files or proprietary statistical formats. In terms of documentation, there will be different levels of quality: information might be more or less extensive, it might be complete and well structured or divided into several incoherent parts.

Steps of Data Preparation

The complexity of data preparation depends on the quality of incoming data files. The process normally includes several steps:

1. If not available anyway, the original files have to be imported and converted in an appropriate working format (e.g. SPSS).
2. In the case of flat files, categorial and sometimes variable labels have to be added. For that step a complete codebook is mandatory.
3. If the dataset's quality is unknown, formal checking procedures are recommended. Performing simple frequencies or cross-tabs may help identify uncoded values or logical inconsistencies. Sometimes additional variables have to be included for processing reasons.
4. Recoding might be considered in order to harmonize various datasets or to support search algorithms. A special case is the NESSTAR software, which sometimes requires specific variable formats.
5. Finally, the processed dataset has to be saved in various formats: for example as SPSS portable file, ASCII flat file or EXCEL file. The NESSTAR format is needed for internal publication purposes.

Steps of Metadata Preparation

The preparation of metadata is just as complex as the preparation of huge data files:

[3] NESSTAR: **N**etworked **S**ocial **S**cience **T**ool **A**nd **R**esources. Homepage: http:// www.nesstar.com

1. As basis of everything as much information as possible about the survey and the related dataset has to be collected. This might include the original documentation, questionnaires, a CATI interview master, related publications or even hyperlinks to project websites.
2. Then the relevant information has to be extracted and summarized.
3. In order to provide the information in a harmonized, well structured format, a codebook has to be compiled. Technically, the codebook is an XML file according to the DDI Document Definition[4].
4. To compile this XML codebook, templates might be used. If necessary, these templates might be modified.
5. Finally, metadata documents have to be produced in various formats for publication and dissemination purposes (i.e. Word, Acrobat).

DDI Codebook

The DDI metadata standard is a comprehensive specification to describe quantitative social science data. Using this standard all relevant metadata information can be documented in a well defined structure. One important thing is the separation of content and layout. The underlying tag structure allows computer aided processing of the inherited information.

Technically the DDI codebook is an XML document type definition defining numerous metadata elements and attributes. It is a rather complex structure consisting of five main sections:

1. Document Description (docDscr): describes the electronic metadata document (metadata about metadata)
2. Study Description (stdyDscr): describes the underlying study
3. Variable Description (dataDscr): describes the structure of underlying statistical data
4. Data File Description (fileDscr): describes the respective data file, e.g. a SPSS file (*.sav)
5. Study-Related Material (otherMat): includes references to supplementary material

For further details concerning the DDI standard see the DDI homepage (http://www.icpsr.umich.edu/DDI/index.html).

Metadata Publication

Once everything is prepared and formatted, it has to be published via the website. This will include the update of some static HTML files, the upload of new datasets and metadata files to both the web server and the NESSTAR server, followed by a reboot process in order to update the incorporated databases. Especially in terms of restricted datasets, the NESSTAR access control unit has to be configured.

The online access includes the downloadable metadata documents, hyperlinks to related publications and websites and – if available – hyperlinks to the respective datasets themselves.

Currently, datasets are distributed by regular mail, sometimes via e-mail. The distribution of confidential data generally is subject to the owner's permission.

[4] DDI: **D**ata **D**ocumentation **I**nitiative (http://www.icpsr.umich.edu/DDI/index.html)