

Herausgeber:

Zentralarchiv für Empirische Sozialforschung
Universität zu Köln

Das Zentralarchiv ist Mitglied der GESIS

Direktor: Prof. Dr. W. Jagodzinski

Geschäftsführer: E. Mochmann

Postanschrift:

Postfach 410 960
50869 Köln

Hausanschrift:

Bachemer Straße 40
50931 Köln

Telefon:

Zentrale 0221 / 4 76 94 - 0
Telefax - 44
Redaktion - 50

Redaktion:

Franz Bauske

E-mail: bauske@za.uni-koeln.de

Internet: <http://www.gesis.org/za>

ISSN: 0723-5607

© Zentralarchiv

Die ZA-Information erscheint jeweils im Mai und November eines Jahres.

Sie wird kostenlos an Interessenten und Benutzer des Zentralarchivs abgegeben.

MetaDater: Towards standards and tools for the description of comparative surveys

by Uwe Jensen and Ekkehard Mochmann



MetaDater Metadata Management and Production System
for surveys in Empirical Socio-economic Research

A Project funded by EU under the 5th Framework Programme
Key Action: Improving the Socio-economic Knowledge Base
- Data infrastructure – HPSE-CT-2002-00122

Project Consortium

- ZA Germany
- DDA Denmark
- EKKE Greece
- NIWI The Netherlands
- UGOT Sweden
- SIDOS Swiss
- NSD Norway
- UKDA Great Britain

Internet: www.metadater.org

The members of the MetaDater consortium started the EC funded project in Cologne under participation of Dr. *Peter Fisch* (EC - DG Research K.4), Dr. *André Schlochtermeyer* and *Angelika Schindler-Daniels* (Bundesministerium für Bildung und Forschung). Opening the official project launch, the co-ordinator *E. Mochmann* addressed the relevance of the project in the context of “Improving the evidence base for comparative socio-economic research”. Peter Fisch underlined in his speech on "Citizens and Governance in a knowledge-based Society" (FP6 Priority 7) the challenges and opportunities of the recent EC Framework Programme for the Social Sciences. The project partners presented work packages and organisation

of the MetaDater Project, including a summary of the work packages to be accomplished over the next three years. The MetaDater¹ project is linked with the MADIERA² Project (Multilingual Access to Data Infrastructures of the European Research Area). *Bjarne Oymyr*, co-ordinator of the MADIERA Project, gave an outline of the project plan for MADIERA.

The meeting proceeded with a detailed discussion and planning of the work for the different MetaDater working packages.

The following text gives an overview of the MetaDater project aims and its main products. The project's website www.metadater.org will inform the community about major milestones, which will be realised in the project period between January 2003 and December 2005.

1 Objectives of the MetaDater Project

Modern democracies produce a growing database for empirical social research. This increase is only manageable with data and metadata management instruments that make the preparation of data files for access and further analysis more efficient. So far, there exists no comprehensive system that integrates the functionalities required for metadata standardisation, storage and output into the workflow.

Overarching objectives of the MetaDater project are to develop standards for the description of large scale comparative surveys over space and time and to provide tools for metadata creation and management for such surveys. In order to achieve these objectives, the MetaDater project has to develop a comprehensive data model for comparative surveys and develop tools for metadata management in survey research.

The aim of the project is to help **Primary Data Producers** (principal investigators, research and fieldwork institutes) and **Data Providers** (large research institutes, data services and data archives) to manage small to large collections of metadata and especially to make metadata capture at the source more efficient. In the long term, this should help to enhance the general quality of metadata, making trend data much more reliable and strengthening the role of social sciences as providers of data for sound information of society.

1 <http://www.metadata.org>

2 <http://www.madiera.net>

It will **bridge the gap** between minimalist documentation that is just sufficient for one-time analysis of the data by the principal investigator and an extensive self-explaining documentation that is required for a further analysis to make optimal use of the data.

The challenge will be to overcome the heterogeneity of data set structures and documentation to feed more easily, with a harmonised input, the front ends presently under development.

2 Definition of Metadata for Comparative Social Research

The **metadata** referred to in this project is the physical representation of meta-information including all elements of information which effectively guide and support the process of identification and extraction of relevant survey data and those which are needed for their valid interpretation. A special focus will be on information needed in secondary and comparative analyses. These "human oriented" metadata range from basic information on the meaning of coded numbers to external background information. They are complemented by machine-readable "process" meta-data for statistical systems or workflow processes.

Data providers have been the first to insist on the importance of good metadata. This is due to their intermediate position between primary data producers and data users. They have also initiated and supported the development of the now broadly acknowledged DDI specifications for survey metadata.

The emerging standard for single cross-section survey data description has been established by the Data Documentation Initiative (DDI) and agreed upon by the major data providers as the data archives represented by IFDO and CESSDA and used in the development of NESSTAR, SDA and ILSES.

*'It is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of "metadata" about data sets in the social and behavioural sciences. Metadata (data about data) constitute the information that enables the effective, efficient, and accurate use of those data sets.'*³

Since the DDI still lacks the adequate description of time series survey data and international comparative studies the metadata model of this project will have to be more elaborate. Therefore the MetaDater project will also extend the data definition

3 cited from: 'DDI Purpose and Goals' at: <http://www.icpsr.umich.edu/DDI/ORG/index.html>

model to include elements for comparative surveys and time series in close co-operation with and as members in the recently formed DDI Alliance.

Survey data, which are in the focus of this Metadata Management and Production System, can only be analysed properly if the knowledge of and about the data is maintained permanently and available to the end-users. The description of the content, the diversity of question formulations and translation problems as well as the context of individual data sets and of data collections are at the core of what metadata means in this project. This includes the comprehensive technical and methodological description of the collected survey data that are necessary to gain empirical evidence through further statistical analysis.

2.1 Management of Metadata Data and the Data Model of Metadata

The general management of metadata includes two types of tasks:

- The 'Management' must handle comprehensive data holdings and collections of single data sets in an integrated manner. This is achieved in a relational database through normalisation, each bit of information being recorded only once, in a specific field of a specific table. That aspect distinguishes most clearly this approach from the document approach of XML files.
- Secondly, a database application has to include fields and functionalities appropriate for managing processes in which data and metadata are involved.

To organise these requirements the first product of the MetaDater project will be a comprehensive data model of metadata for the outlined scope of surveys. The model will cover the whole life cycle of such surveys within the logic of metadata. It gives the frame for different tasks like managing, preserving, standardising and exporting metadata and the respective applications covering

- the survey design and implementation phase,
- the data collection and data processing phase, as well as including
- generic management and dissemination tasks with different quality protection aspects.

The model will be distributed as a product itself, i.e. independently from the implementations and as a resource for any survey metadata system development within different institutional frameworks and in accordance with their specific requirements. The feasibility of the data model will be exemplified and tested by two prototype implementations. The data provider application (MD-PRO) and the data col-

lector application (MD-COLL) are the second and third main products to be realised in the project.

2.2 Data Provider's metadata and the application MD-PRO

MetaDater will support the different stages of the comprehensive data provider workflow for a survey. The target groups for this tool are social science data archives, but also large research institutes or data service centres as far as they are carrying out comparable tasks such as extensive data documentation and publication for secondary analysis.

Construction and management of metadata concerning this workflow will also apply for the primary data producers. The main routines of the procedure are:

- Data acquisition, cataloguing and archiving and quality management;
- Data control and processing (including harmonisation across time and space);
- Extensive documentation on study and variable level and Metadata standardisation;
- User, data dissemination and metadata publication management.

The complexity of the system follows from the variety of objects and levels to be managed in the database. While data acquisition is made on study level and negotiated with persons or institutions, variable standardisation is made at the lowest level in the database: variables and values.

The respective data provider system (**MD-PRO**) will contain instruments for the different management tasks and supply access to resources like question databases, codebooks, and metadata files in various formats. The system will include import and export routines for metadata in various formats, and will supply the access to its content via end-user interfaces or portals like NESSTAR or MADIERA.

2.3 Data Collector's metadata and the application MD-COLL

One source of difficulties with an adequate documentation of a survey and data is the lack of an integrated instrument to capture related metadata when they first appear and in a standardised format. Collection of metadata starts with the survey design and the development of suitable instruments (questionnaire; collection instrument), continues with its implementation in the field and may find a preliminary end when data are validated or analysed for the first time or if new variables are constructed in this phase.

The target groups for the data collector tool (**MD-COLL**) are primary investigators and small or middle-sized research services. To make their work as efficient as possible the application will support the workflow facilitating the structured entry and storage of information when a survey is born and launched (e.g. survey design; fieldwork documentation; social and cultural survey context).

Basic management facilities are considered as well as routines to support the editing of multilingual questionnaires and to set up a data documentation (codebooks, question repositories). This should include extended possibilities to capture comments on the data to facilitate the management of or information on problems with not sufficiently specified variables or codes following international standards for data documentation.

While supporting the data producers' internal workflow, standard formats will make exchange with data providers (archives) more efficient by avoiding double work and guaranteeing appropriateness of documentation.

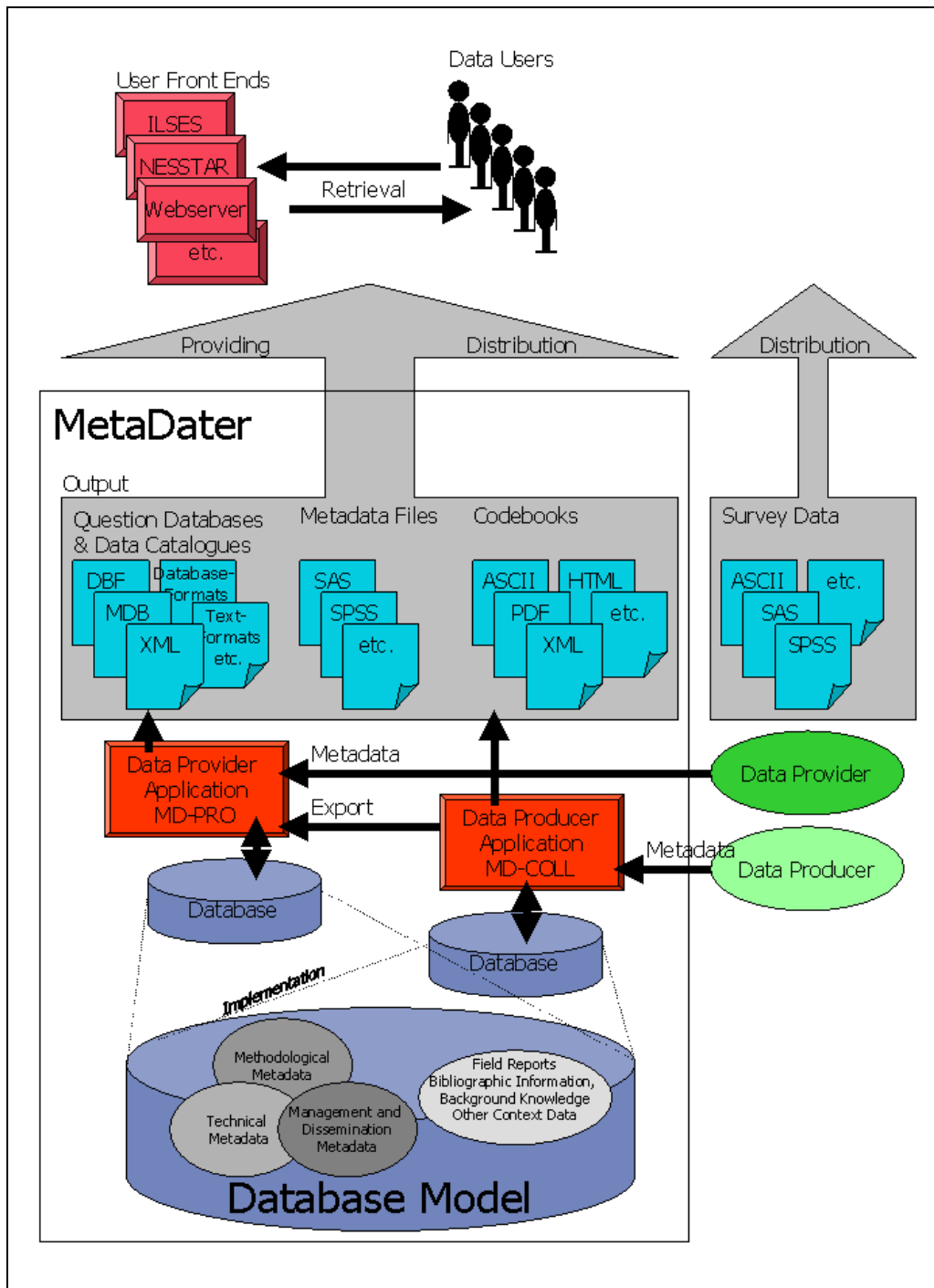
3 MetaDater components – The data model of metadata and tools MD-PRO and MD-COLL with its output

This figure visualizes the major MetaDater components. At the bottom, the general database model including three types of metadata:

- **Methodological** metadata concerning the measurement instruments (survey design; questionnaire; fieldwork items)
- **Technical** metadata which give concrete meaning to the (raw) data values in form of coded numbers and which define them for statistical systems
- **Management and dissemination** metadata, which enable provider and end-user to identify relevant data under different tasks and quality perspectives (from classifications to release dates).

In addition, links to various types of relevant context information will be organized, like:

- Bibliographic information on the research context, i.e. publication-based interpretation and validation of the data
- Background knowledge such as social or cultural background information. They are of major importance for the adequate interpretation in the context of comparative research. The access to such well structured "metadata knowledge systems" will supply context information on the origin of data in a specific society at a specific time point (educational systems, religion, party systems etc.)



Relevant subsets of this model will be implemented in the two applications, MD-PRO and MD-COLL. The data themselves (to which these metadata refer) are administrated in the system but kept outside.

Possible outputs of the tools are SPSS data definition files, codebook type documentation on study and variable levels, data catalogues or question databases. These products are the expected standard input for public access via existing and emerging data service portals.

4 Expectations and Benefits of the MetaDater

The MetaDater will be an infrastructure instrument which facilitates metadata transfer from primary data producers to data providers and supports primary data producers and data providers in supplying the end-user efficiently and continuously with reliable, high quality, standardized and durable information about survey data. The MetaDater also supports long-term preservation of data and related metadata, to keep the "digital heritage" in the field of social research. The benefits will not just be relevant for the existing institutes but also for the emerging infrastructures in Middle and Eastern European countries.

It is expected that the greater integration of processes over time and various kinds of data services will bring additional incentives to preserve data and increase their usability, so the whole data and metadata production process becomes more efficient and economical. As a result, there should also be more well documented data available for secondary analysis and social indicator research.

The resulting standards and the tools will also be applied to test technical harmonization and integration of survey data held by data providers in the participating countries. By making the standards and tools available to academic and commercial research and through publishing the standards, the MetaDater project will contribute to best practice in survey data resource sharing and data distribution. It facilitates next generation processing and analysis of huge amounts of data in order to increase empirical evidence and knowledge about European and global socio-economic developments.

MetaDater is expected to strengthen and develop the European technological infrastructure for the social sciences and to facilitate access to well documented data for its users. It considers the increasing request for access to comparative data across the European Union and world-wide. In particular, it has the potential to improve the data basis for the analysis of social change in the context of the European integration and globalization processes and thereby significantly contribute to the growth of the socio-economic knowledge base. With increasing visibility, MetaDater should contribute to implement best practice in social survey documentation and distribution.

References: DDI: <http://www.icpsr.umich.edu/DDI/>
MADIERA: <http://www.madiera.net>
MetaDater: <http://metadater.org>

We gratefully acknowledge the conceptual contributions to the MetaDater project based on the rich experience of documenting large-scale comparative surveys over decades, in particular by *Meinhard Moschner, Rolf Uher, Oliver Watteler, Wolfgang Zenk- Möltgen* and other ZA staff members.