Metadata – “data on the data” – is a basic topic in any attempt to standardize statistical data either through uniform names and descriptions or through structured crosswalks and retrieval systems. Statistical “numbers” can only be understood if they are adequately described, especially if these data are as highly aggregated as those found in indicators. Complete and standardized metadata supports the successful selection of appropriate statistical data in urban research and helps ensure that the comparisons of data between two or more locations, or over time, is accurate.

The collaborative development of policy measures and measures of effectiveness and outcome provides an intellectual basis for the sharing and comparison of statistical measures between cities and regions. Metadata is the infrastructure that makes such sharing and comparison possible. A number of organizations are working on the development of metadata standards to support systems for locating, obtaining and working with social science statistics. In addition, they are exploring the type of information that is needed to effectively compare data between different systems and over time.

This workshop explored a number of metadata initiatives, efforts to develop standard definitions, and the use of crosswalks to map between different coding and descriptive structures. The outcome of this workshop was not only a better understanding of how you can contribute to these efforts individually or collectively, and how you can benefit from their work, but also a proposal for an application or relevant findings in the work of SCORUS on standardization of urban and regional indicators.
Importance of metadata in the dissemination of statistics

Metadata in this context refers to closer defining of the figures from which statistics are compiled. The defining may concern matters such as data collection, time of data collection, data content or data source. Users of statistics must have a clear perception of how and for what purposes statistical data can be used.

One of the strategic goals Statistics Finland has set for itself is improvement of statistical literacy and user skills. Whenever possible, it endeavours to attach product presentations to the statistics it disseminates. With population censuses, for example, this is provided in the form of a handbook.

Metadata are particularly important when statistical data are disseminated via the Internet. Ease of access to statistical tables amplifies their utilisation. Attaching some kind of explanatory information to statistical tables becomes almost imperative at times. In Statistics Finland’s Web-based service of Urban and Regional Indicators, for example, metadata are available for each individual table.

To ensure that produced statistics meet high quality standards, up-to-date concept and classification databases are also maintained in Finland. These databases are integrated into statistics production so that the right classification and congruent concepts are used when statistical tables are produced. The classifications and concepts are generally based on international recommendations. One of the domains for which statistical concepts and definitions are vital is Statistics Finland’s Virtual School of Statistics, which has made use of the concept database since 2001.

Plan for the presentation

- Statistics Finland’s meta system
- Content of metadata - Quality Guidelines for Official Statistics
- Product descriptions - Internet publications
- Urban and Regional Indicators
- Statistics production and metadata
Classification Database

- All essential classification standards and their versions in Finnish, Swedish and English
- In different lengths
- Classification metadata
- Indexes for large classifications (search terms)
- Conversion keys between classifications

Concepts Database

- Scientific concepts
- Concepts used in different statistics (statistical publications)
- Sources and links to other sources

At the moment used in Statistics Finland’s Virtual School of Statistics
- potentials for wider use

Checklist

1. Relevance of statistical data
   - Purpose of use
   - Definitions
2. Correctness and accuracy of data
   - Methods
   - Statistical quality of data
3. Timeliness and promptness of published data
4. Accessibility of statistical data and transparency/clarity of the statistical process
   - Data dissemination
   - Interpretability of data
Statistics Finland

Checklist

1. Relevance of statistical data
   - Purpose of use
   - Definitions
2. Correctness and accuracy of data
   - Methods
   - Statistical quality of data
3. Timeliness and promptness of published data
4. Accessibility of statistical data and transparency/clarity of the statistical process
   - Data dissemination
   - Interpretability of data

Statistics Finland

5. Comparability of statistics
6. Clarity and consistency/uniformity
7. Documentation
   - Information system
   - Production process
   - Data sets
   - Metadata related to data administration

Rules of thumb

- Documentation should always fulfil all the requirements the customer has set for the statistics concerned.
- Anything that would cause considerable problems if left undocumented should be documented.
**Product description and separate quality analysis**

- A product description should be attached to every OSF serial publication including press releases and advance reports.
- Product description = a concise assessment of the quality, reliability and applicability of the statistical survey for different purposes.
- A separate quality report should made where customers require such an analysis. It is particularly important in the case of statistics where the number of end-users is very large.

---

**Population Census 2000 Handbook**

- Concepts and definitions
- Data items
- Products and services
- 2000 Population Census tables
- Classifications

---

**Internet publications: product description**

- Brief quality descriptions should be attached to databases and tables published in HTML format on the Internet (cf. press releases).
- A link should be provided to the product description together with a reference to the relevant OSF publication.
- If the statistics are published only in electronic format or in the OSF online service, a product description meeting the corresponding requirements for serial publications is required, with references and/or links provided.
Urban and Regional Indicators - data -

- Statistics and indicators on 13 topics, over 50 tables and graphs
- Service contains more than 3 million data cells
- Data on all municipalities in Finland by municipality, urban region, region and sub-regional unit
- Time series, most data starting from 1985
- Data producers: Statistics Finland, Housing Fund of Finland, Ministry of Education, STAKES (National Research and Development Centre for Welfare and Health), Ministry of Labour, University of Jyväskylä
- The latest data available: nearly all tables are updated once a year

Urban and Regional Indicators - metadata -

- Contact information
- Data content
- Classifications
- Data source
- Definition of statistical units
- Reference point of time
- Reliability
- Maintenance
- Key words
- Further information
Main operational process of statistical survey

Metadata
Most important characteristics

- From data provider point of view
  - Integration to production process
  - Easy-to-use tools to reuse the information
  - International standards

- From data user point of view
  - Demands (extent) dependent on purpose of use
  - Easy-to-use (easy to understand)
  - Minimum information - cf. OSF product description - “problems if left undocumented”

Thanks for listening!
SCORUS Workshop 4

Angelika Schulz
German Aerospace Center/Institute of Transport Research, Berlin, angelika.schulz@dlr.de

Metadata for Transport related Data - Preparation & Presentation & Dissemination

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". 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 have 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 kinds of data. The point is, that often virtually nobody - except the data owner himself - knows about their existence and possible availability. Even though a certain survey or modelling approach is known by the community, its third party 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 everyday 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,

1 The Institute of Transport Research at the German Aerospace Center operates the clearing house as a nonprofit 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.
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. Modelling 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 behaviour 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 NESSTAR², 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 may 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.

---

² NESSTAR: Networked Social Science Tool And Resources. Homepage: http://www.nesstar.com
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:

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 Definition4.6.

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 (studDscr): 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.

Contact
Address  
German Aerospace Center
Institute of Transport Research – Clearing House
Rutherfordstrasse 2
12489 Berlin
Germany

Phone  +49 (0) 30 67055 204

Fax  +49 (0) 30 67055 202

Homepage  http://www.clearingstelle-verkehr.de

3 DDI: Data Documentation Initiative (http://www.icpsr.umich.edu/DDI/index.html)
Processing Metadata for Empirical Data
The Example of the Clearing House for Transport Data & Transport Models

German Statistical Week 2003/ SCORUS Satellite Conference
Workshop IV – Metadata for Urban and Regional Statistics
Friday, August 22 – Potsdam, Germany

Angelika Schulz
German Aerospace Center - Institute of Transport Research
Berlin-Adlershof, Germany

Current Situation
- transport research is a highly complex and heterogeneous topic
  - divided in several subdivisions
  - uncountable research projects, respective outcomes, and publications
  - huge amounts of data as basis of scientific work
- insufficient knowledge about already existent information and its availability
- limited usability of data and models due to insufficient documentation
- limited comparability due to varying format and quality of documentation

Objectives of the Clearing House
- to encourage publication and dissemination of relevant research outcomes
- to provide metadata as well as data (in a standardized format)
- to emphasize transport issues in order to supplement other archives (e.g. federal statistical authorities)
- to facilitate easy access to information

all in all:
- to provide better information on transport data and fundamentals of transport models
Thematic Focus I – Transport Related Data

- data as **basic input** for transport research and related disciplines, e.g.
  - empirical survey data on everyday mobility, individual time-use, infrastructure, ...
  - measurement data such as traffic flow data, floating car data (FCD), ...
- various **components** have to be documented, e.g.
  - large data sets (statistical software formats, flat files)
  - extensive supplementary material (codebooks, digitized questionnaires)
  - availability status, holdings information
  - related publications and project websites

Thematic Focus II – Transport Models

- modeling as **basic method** used in transport research, e.g.
  - models of traffic flow, traffic demand, ...
- various **components** have to be documented, e.g.
  - formulas, source code
  - information about required input data
  - related software applications (simulation tools)
  - availability status (open source vs. commercial software product)
  - related publications and project websites

Components

- **data base for quantitative transport data**
  - empirical and traffic flow data resulting from surveys or measuring campaigns
  - explanatory metadata and supplementary material
- **data base for transport models**
  - mainly explanatory metadata
- **test suite for transport models**
- **cross references** (hyperlinks) to additional information sources
  - other archives
  - research institutions
Retrieval and Information Access

- catalogues (thematic, spatial, chronological)
  exploration of existent information, identification of relevant surveys/models
- search engine (site-specific, based on keywords)
- metadata documentation (according to DDI standard)
  in-depth examination of selected information regarding usability,
- NESSTAR Light Explorer (metadata & data)
  preliminary insight into data set or modeling approach

Documentation - Preparation Process

- Metadata are data about data.
- Well structured metadata allows for the effective, efficient, and accurate use of empirical datasets.
- The process of providing both data and metadata includes several steps:
  1) preparation of data
  2) preparation of metadata
  3) publication
  4) dissemination

Preparation Process – Input

- As input we get …
  - data as output from empirical surveys or measurement campaigns
  - original data documentation (including methodological and project reports, codeplan)
  - related publications
- One have to deal with …
  - various data formats (statistical software formats, flat files)
  - differing quality of documentation (more or less extensive, well structured or not ...)
**Preparation Process - Data**

- The complexity of *data preparation* depends on quality of incoming data files. The process normally includes several steps:
  - import of original files (i.e. ASCII flat files, SAS, SPSS) and transformation in the preferred working format (SPSS)
  - adding variable and categorical labels (flat files don’t include this information)
  - formal data checking such as simple frequencies or cross-tabs (search for missing/uncoded values, logical inconsistencies, inclusion of additional variables for processing reasons)
  - recoding (harmonization, comparability, NESSTAR requirement)
  - export to other formats (especially for internal use, NESSTAR NSD format, for dissemination SPSS portable files, EXCEL or ASCII flat files)

**Preparation Process - Metadata**

- The extensive process of *metadata preparation* includes the following steps:
  - collecting information about survey and dataset (explanatory material such as methodological and project reports, codeplans, questionnaires, related publications, hyperlinks to project websites)
  - extracting and summarizing the relevant information
  - preparing the standard “codebook” according to DDI Document Type Definition (XML-file)
  - modification of standard codebook file (adding elements and attributes if necessary)
  - producing metadata documents in various formats for publication and dissemination purposes

**Metadata Standard – DDI "Codebook"**

- specification to *describe* quantitative social science *data*
  - developed since 1995 by the *Data Documentation Initiative DDI*, which could be traced back to the IASSIST conference (International Association of Social Science Information Service and Technology)
  - *co-operation* of ICPSR (Inter-university Consortium for Political and Social Research) and CESSDA (Council of European Social Science Data Archives) and others
  - a way of *structuring* and *formatting* the metadata documentation (separation of format and content)
  - tagged structure enables *computer aided processing* of the information
DDI – "Codebook" Elements

- The whole set of DDI elements is defined within the Document Type Definition "codebook.dtd" (base of XML documents).

- The codebook is structured in 5 parts, each of it including numerous elements:
  - **Document Description** (docDescr) describes the electronic metadata document (metadata about metadata)
  - **Study Description** (studyDescr) describes the underlying study
  - **Variable Description** (dataDescr) describes the structure of underlying statistical data
  - **Data File Description** (fileDescr) describes the respective data file, i.e. a SPSS file (*.sav)
  - **Study-Related Material** (otherMat) includes references to supplementary material

DDI "Codebook" – Document Description (extract)

```
0.0 codeBook (ATT == ID, xml:lang, source, version)
| 1.0 docDescr (ATT == ID, xml:lang, source)
|   | 1.1 creation (ATT == ID, xml:lang, source, MARCURL)
|   |   | 1.1.1 title (ATT == ID, xml:lang, source)
|   |   | 1.1.2 subTitle (ATT == ID, xml:lang, source)
|   |   | 1.1.3 altTitle (ATT == ID, xml:lang, source)
|   |   | 1.1.4 patronTitle (ATT == ID, xml:lang, source)
|   |   |   | 1.1.5 IDNo. (ATT == ID, xml:lang, source, agency, level)
|   |   | 1.2 resType (ATT == ID, xml:lang, source)
|   |   | 1.3 prodStmt (ATT == ID, xml:lang, source)
|   |   | 1.4 distStmt (ATT == ID, xml:lang, source)
|   |   | 1.5 serStmt (ATT == ID, xml:lang, source, URI)
|   |   | 1.6 verStmt (ATT == ID, xml:lang, source)
|   |   | 1.7 biblio (ATT == ID, xml:lang, source, format)
|   |   | 1.8 holdings (ATT == ID, xml:lang, source, location, online, URI, media)
|   |   | 1.9 notes (ATT == ID, xml:lang, source, type, subject, level, resp, sidetext)
|   |   | 1.2 guid (ATT == ID, xml:lang, source)
|   | 1.3 docStatus (ATT == ID, xml:lang, source)
|   | 1.4 docSet (ATT == ID, xml:lang, source, MARCURL)
|   | 1.5 notes (ATT == ID, xml:lang, source, type, subject, level, resp, sidetext)
```
**DDI "Codebook" – Data Description (extract)**

```xml
<datadescription>
  <dataset name="DATASET_NAME">
    <variable name="VAR_NAME" type="VAR_TYPE">
      <attribute name="ATT_NAME" value="ATT_VALUE"/>
    </variable>
    ...
  </dataset>
</datadescription>
```

**5.0 otherMat** (ATT == ID, xlink:long, source, type, level, URI)

```xml
<otherMat>...
  <note name="NOTE_NAME" type="NOTE_TYPE" level="LEVEL">
    ...
  </note>
</otherMat>
```

**NOTE:** otherMat is recursively defined to 5.0.

**DDI "Codebook" – Other Material (extract)**

**(extract of content)**
Preparation Process – Publication

- In order to make all information available to the public, data and metadata have to be published:
  - updating the website (HTML files, catalogues)
  - uploading new data and metadata files to the web server and the NESSTAR server (including reboot for updating purposes)
  - configuring document access conditions (NESSTAR access control unit)

Dissemination of Information

- online access to information is provided via the website [http://www.clearingstelle-verkehr.de](http://www.clearingstelle-verkehr.de)
  - downloadable metadata documents in different formats
  - list of related publications (if available)
  - hyperlink to NESSTAR server (access to data, if available and not restricted)
  - hyperlinks to related websites (if available)

- postal distribution of confidential data on CD ROM via snail mail (subject to owner’s permission)

Dissemination - Search Result

*link* to available metadata *link* to available data (NESSTAR)
Dissemination: List of Metadata Documents

Dissemination: NESSTAR - Light Explorer

Networked
Social
Science
Tools
And
Resources
**NESSTAR Light Explorer – Metadata I**

- access to metadata via explorer tree
- structured representation of underlying DDI Document Type Definition (XML-metadata file)

**NESSTAR Light Explorer – Metadata II**

- metadata associated with a particular node are displayed in the description area

Dataset: KONTIV 1989 - Kontinuierliche Erhebung zum Verkehrsverhalten

- **Times Method**:
  - Guerschmultitude

- **Data Collector**:
  - EIMN-Institut GmbH & Co., Bielefeld

- **Frequency of Data Collection**:
  - Wocheweise\n
- **Sampling Procedure**:
  - Zweckmäßige geschätzte Zufallsstichprobe, Random-Nein-Überleitung
  - 1. Stich: geschätzte Zufallsauswahl von Befragungsbereichen
  - Grundgesamtheit enthält standardisierte Stichprobenermittlung mit insgesamt 1500 Personen
NESSTAR Light Explorer – Data Access

- results of basic analyses, e.g. cross-tabs, are displayed in the table area

NESSTAR – Access Control

- survey specific access control
- analysis and download of data only for authorized users
- access to metadata without any restriction

Prospects – Documentation: What's left to do?

- to develop overall templates for survey data as well as for other data formats such as traffic flow data (DDI standard)
- to develop an internal workflow in terms of data and metadata preparation
DUVA --the Metadata-based Information Management System

The discussion about data and their descriptions - the metadata - is an old one. Basic concepts go back to the seventies. For DUVA development the paper Statistical Meta-Information and Meta-Information Systems by B. Sundgren (Working Paper No. 4 of the First Working Session on Statistical Metadata (METIS), Geneva, Oct., 8.-10. 1991) as well as the ideas and concepts of Prof. Günther Appel (former president of the Statistisches Landesamt Berlin) have to be mentioned as fundamental impulses. So the realisation of the principle that 'meta information is systematically collected at its source and is subsequently reused and augmented repeatedly throughout the statistical production process' makes DUVA a special tool developed by and for the KOSIS union of German and other European municipalities and public institutions.

DUVA is a metadata-based information management system capable of performing standardized processes for generating and distributing information derived from any data sources. More than 10 years of practical experience and the knowledge acquired during that time by more than 45 users in Germany resulted in DUVA98, a powerful 32 bit application which can be used with WINDOWS operating systems. Its major strengths are the features of a conceptually and technically open system (Paradox, MS-SQL, Oracle, DB2…) in addition to the consistent implementation of the metadata concept. The open systems approach was also used for the further processing of information obtained from the system as well as information reusability. This makes DUVA an open, metadata-based data warehouse system.

The basic idea underlying DUVA system development was to provide a management system for meta- and physical information without splitting between different media and to provide meta-information interfaces for all program elements forming part of the statistical production process (data bases, aggregation and analysis programs, tabulation, graphics, plotting programs, etc.). The content-related flexibility of the system is extended to include technical flexibility. Furthermore, the system should offer exhaustive search functions (e.g. search for individual data, linked information, analytic results, etc.).

The DUVA referral system (NWS) collects and maintains meta-information and uses this set of information for system control purposes. Although meta- and physical data are held by separate media, the NWS referral system provides access to physical information. Throughout the statistical production process, communication between users and the NWS system is in natural language. This applies to both process control instructions and all search processes.

All the steps - from data collection (e.g. by using DUVA data entry masks) via standardized data editing (macro file concept) to the presentation of results (e.g. dynamic generation of tables of results by the DUVA Internet Assistant, dynamic generation of population pyramids by GIZEH, etc.) - are controlled by meta-information interfaces.

Descriptive information obtained from each step flows into the meta database of the NWS referral system and contributes to a complete metadata documentation. Import and export functions enable a superior division of work principle and result in an effective transfer of knowledge within an open system of DUVA users.
**General demands for today's statistical work:**

- Speed and flexibility in providing high-quality information and knowledge
- Development of information and knowledge by using various departments and data sources
- To meet the requests of a multitude of clients

**DuvA**

Is the system that helps you to manage today's statistical work

The system provides access, e.g., to all fields or databases.

For periodic analysis you can automate projects.
is the system that helps you to manage today's statistical work

DuVA provides qualified tools for professional data presentation (three clicks to produce a table (including labels etc.), tables with animated population charts)

People in Exempleton according to sex and age (time series 1995-1997)

Zeitraum: 31.12.95 bis 31.12.97
People 1995, People 1996, People 1997 each Age, Sex.

<table>
<thead>
<tr>
<th>Age</th>
<th>Male</th>
<th>Female</th>
<th>People 1995</th>
<th>People 1996</th>
<th>People 1997</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;15 years</td>
<td>1200</td>
<td>1200</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
</tr>
<tr>
<td>15 - 65 years</td>
<td>11500</td>
<td>11500</td>
<td>23000</td>
<td>23000</td>
<td>23000</td>
</tr>
<tr>
<td>65 years and older</td>
<td>1200</td>
<td>1200</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
</tr>
</tbody>
</table>
**is the system that helps you to manage today's statistical work**

**Introduction**

DUVA is the system that helps you to manage today's statistical work. DUVA provides a tool to design HTML-data entry masks for your online surveys.

DUVA tools make it absolutely easy to organize a documentation-system without programming (e.g. case documentation in a social welfare project). Through the 'three-click' method the end users can start to work right away.

DUVA integrates your system environment (e.g. standard software and databases).
What is the background and what makes DUVA\textsuperscript{98} special?

Metadata management:
Process controlling through metadata

DUVA- user admin.

DUVA98 referral system NWS

Metadata database

DUVA98 data-entry module DUMA

Flow of metainformation

Flow of physical information

DUVA98 - modules
- ASEXCEL (Excel-interface)
- IASS (internet-assistant)
- DABANK-assistant
- MDIREKT (tables, charts, maps)
- SPSS-interface
- GIZEH (pop.-pyramid charts)

Phys.-data

DUVA98 target file generator ZDG

Phys. Data

DUVA- macro file

Flow of metadata and physical data

Metadata: total amount of information describing and explaining physical data

Physical data:
- codes, keys
- frequencies
- values...

Metadata:
- Descriptions of the physical data; e.g.
- labels, units,
- record structures.....
The increase of metadata during the flow of statistical work

The amount of metadata grows with every step of the statistical process chain

Definition of research objective
- Theoretical reflection
- Concept specification, operationalization
- Data collection concept
- Data collection
- Data editing
- Data analysis
- Data communication/presentation

Meta data

The compression of physical data

Result of data collection: physical data

Definition of research objective
- Theoretical reflection
- Concept specification, operationalization
- Data collection concept
- Data collection
- Data editing
- Data analysis
- Data communication/presentation

Meta data

Physical data: steps of compression

35.7 km
Knowledge based data analysis and flexible supply of information to develop and evaluate policy measures require...

... that the whole amount of meta information is completely recorded, maintained and used for control purposes throughout the process flow.

Metadata management requires a systematic metadata-organization.
**Metadata management requires a systematic metadata-organization**

The coding system ensures the transformation from observations to symbols.

- **Code tables**
- **Range of values**
- **Identifying codes**

---

**DUVA provides 100% metadata management**

Metadata are systematically entered at their source and subsequently reused.
Metadata and physical data must be permanently linked without media splitting.

DUVA provides 100% metadata management.

Process controlling and the whole statistical workflow is organized by metadata.

The core system.

The communication between user and system is organized in natural language.
**The core system**

Retrieval starts at least with one keyword or syllable but makes you to find anything in the metadata database.

---

**Effects:**
- unnecessary and expensive double-work will be prevented
- through common metadata structures, comparable statistics (national and international) can be produced

---

**System development**

DUVA is a cooperative project of the KO SIS city network. The DUVA steering group under the custodianship of the city of Freiburg determines its editorial and organizational structure.

More than 50 municipal bureaus and other institutes that work with statistics form the DUVA-user group.

Contact us via internet: http://www.duva.de
Metadata -- data about data -- is integral to the work that we do. Numbers without definition are meaningless. Structured metadata allows us to store information about data and the process of creating data for our own use, for presenting to data users of all types, and for sharing metadata and data in a uniform way. SCORUS has a vested interest in both the content and structure of metadata. The members of the workshop focused on identifying areas of activity that would help individual members as well as areas where we could constructively contribute to the international development of metadata standards (structural and content standards).

The two metadata structures that were discussed were the Data Documentation Initiative (DDI http://www.icpsr.umich.edu/ddi/) and DUVA which is used by a growing number of state and local statistical offices in Germany. Conceptually we discussed metadata and metadata activities in the following framework:

- We are concerned with the development of consistent metadata for ourselves AND for data users of all types
  - This means we need to view metadata more broadly than our current internal needs may dictate and encourage our statistical offices to consider developing international standards. To do this, we need to integration concepts of metadata into our work processes. These include:
    - creation
    - updating to accommodate new versions of a metadata standard
    - integrating new data sets
    - version management

- Rather than trying to dictate adherence to a specific structure, SCORUS should assist in comparing different structures, assist in mapping between structures currently in use, and assisting individuals with finding additional information and expert resources for various structures. The intent here is that as a particular system emerges as a transport structure, locally used systems would be in position to export their locally held metadata into a transport structure for sharing and broader access. An example of this and perhaps a prototype project would be mapping between DUVA and DDI or other widely used standards. Specific issues to consider:
  - do we cover the same fields
  - are required elements in DUVA available in DDI
  - comparison of standard contents – controlled vocabulary
  - the ability to transfer relationships
  - Is there a structure in DDI to record derivation information including source, filters, recoding etc.

The purpose of this workshop was to develop a work plan or work possibilities that would provide the opportunity for members to share their expertise and to contribute to the international development of metadata standards. The following list of work possibilities were divided into structural development issues and content development issues.

**Structural**

- Compare ones own system with DDI
  - Those of us involved in working with structured metadata need to compare what we are doing with the DDI. This includes content differences (elements and attributes identified within each system), coverage differences (type of data or data structures covered), and mapping between the elements of the local system and the DDI. This may require the development of a template for comparison so that these comparisons can be shared on the SCORUS web site.

- Provide resources to assist with questions on DDI and other systems
  - Developing a section of the SCORUS web site that provides
    - links to resources for different systems
    - comparative information developed in the process noted above
    - Contact information for experts within SCORUS that can assist members with specific questions or relay information and concerns to the developers of specific structures.
  - Identify areas where enhancements can be made to DDI or other structures
    - Develop a process for developing and relaying recommendations for enhancement (this will vary based on the organization developing the structure)
  - Sharing concept models – resource for identifying and describing models
  - Define the set of specifications for DDI entry tools for metadata creation – processing tools for updating, version control etc.

**Contents**
Content here means the intellectual content of what is entered into a metadata structure. It covers the development of harmonizing terms and descriptions across borders and over time, and identifying those pieces of information needed to make reasonable comparisons across time and space.

- Criteria checklist for content –
  - recommendations on variables and categories to improve the possibilities for comparison
    - structured terminology where possible
    - common core of categories for specific variables to improve comparison of aggregated data
  - Identify a basic list of recommended metadata elements (e.g. time coverage, geographic coverage, sampling, weighting, etc.)

- Create cross-walks between different descriptive systems; identifying those that are comparable and those that are not; perhaps expanding to details about how they differ...and what has to be done to harmonize

We expect a thorough discussion of this wide range of questions relating to metadata in order to try and identify the topics to be addressed in SCORUS, the International Network for Urban and Regional Statistics.