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)
- **search engine**
  (site-specific, based on keywords)

→ **exploration** of existent information,

→ **identification** of relevant surveys/models

- **metadata documentation**
  (according to DDI standard)
- **NESSTAR Light Explorer**
  (metadata & data)

→ **in-depth examination** of selected information regarding usability,

→ **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 …)

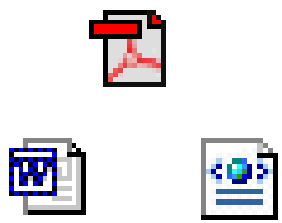
Preparation Process - Data

- 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 categorial 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 **D**ata **D**ocumentation **I**nitiative **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 (docDscr)**	describes the electronic **metadata document** (metadata about metadata)
• **Study Description (stdyDscr)**	describes the underlying **study**
• **Variable Description (dataDscr)**	describes the structure of underlying statistical data
• **Data File Description (fileDscr)**	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 docDscr* (ATT == ID, xml:lang, source)
|     |---- 1.1 citation? (ATT == ID, xml:lang, source, MARCURI)
|     |     |---- 1.1.1 titlStmt (ATT == ID, xml:lang, source)
|     |     |     |---- 1.1.1.1 titl     (ATT == ID, xml:lang, source)
|     |     |     |---- 1.1.1.2 subTitl* (ATT == ID, xml:lang, source)
|     |     |     |---- 1.1.1.3 altTitl* (ATT == ID, xml:lang, source)
|     |     |     |---- 1.1.1.4 parTitl* (ATT == ID, xml:lang, source)
|     |     |     +---- 1.1.1.5 IDNo*    (ATT == ID, xml:lang, source, agency, level)
|     |     |---- 1.1.2 rspStmt? (ATT == ID, xml:lang, source)
|     |     |---- 1.1.3 prodStmt? (ATT == ID, xml:lang, source)
|     |     |---- 1.1.4 distStmt? (ATT == ID, xml:lang, source)
|     |     |---- 1.1.5 serStmt? (ATT == ID, xml:lang, source, URI)
|     |     |---- 1.1.6 verStmt* (ATT == ID, xml:lang, source)
|     |     |---- 1.1.7 biblCit?(ATT == ID, xml:lang, source, format)
|     |     |---- 1.1.8 holdings* (ATT == ID, xml:lang, source, location, callno, URI, media)
|     |     +---- 1.1.9 notes*    (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
|     |---- 1.2 guide?     (ATT == ID, xml:lang, source)
|     |---- 1.3 docStatus? (ATT == ID, xml:lang, source)
|     |---- 1.4 docSrc*    (ATT == ID, xml:lang, source, MARCURI)
|     +---- 1.5 notes*     (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
```

DDI "Codebook" – Study Description (extract)

```
|---- 2.0 stdyDscr+ (ATT == ID, xml:lang, source, access)
|    |---- 2.1 citation+ (ATT == ID, xml:lang, source, MARCURI)
|    |---- 2.2 stdyInfo* (ATT == ID, xml:lang, source)
|    |    |---- 2.2.1 subject? (ATT == ID, xml:lang, source)
|    |    |    |---- 2.2.1.1 keyword*  (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |---- 2.2.2 abstract* (ATT == ID, xml:lang, source, date)
|    |    |---- 2.2.3 sumDscr* (ATT == ID, xml:lang, source)
|    |    |    |---- 2.2.3.1 timePrd*   (ATT == ID, xml:lang, source, event, date, cycle)
|    |    |    |---- 2.2.3.4 geogCover* (ATT == ID, xml:lang, source)
|    |    |    |---- 2.2.3.5 geogUnit*  (ATT == ID, xml:lang, source)
|    |    |    |---- 2.2.3.6 anlyUnit*  (ATT == ID, xml:lang, source, unit)
|    |    |    |---- 2.2.3.7 universe*  (ATT == ID, xml:lang, source, level, clusion)
|    |---- 2.3 method* (ATT == ID, xml:lang, source)
|    |    |---- 2.3.1 dataColl* (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.4 sampProc*     (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.7 resInstru*    (ATT == ID, xml:lang, source, type)
|    |    |    |---- 2.3.1.9 collSitu*     (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.10 actMin*      (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.1.12 weight*      (ATT == ID, xml:lang, source)
|    |    |    +---- 2.3.1.13 cleanOps*    (ATT == ID, xml:lang, source, agency)
|    |    |---- 2.3.3 anlyInfo? (ATT == ID, xml:lang, source)
|    |    |    |---- 2.3.3.1 respRate*  (ATT == ID, xml:lang, source)
|    |    |---- 2.4.1 setAvail* (ATT == ID, xml:lang, source, media, callno, label, type)
|    |    |-----2.4.2 useStmt* (ATT == ID, xml:lang, source)
|    |---- 2.5 othrStdyMat* (ATT == ID, xml:lang, source)
```

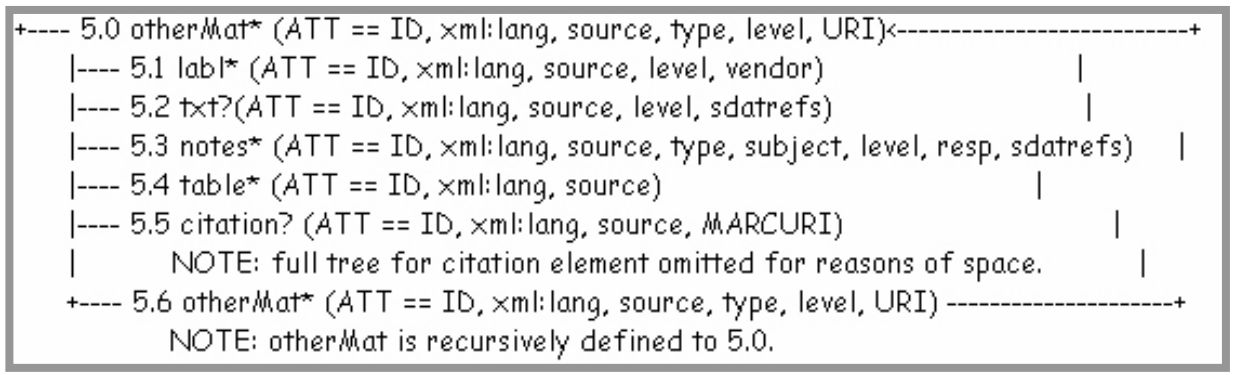
DDI "Codebook" – File Description (extract)

```
|---- 3.0 fileDscr* (ATT == ID, xml:lang, source, URI, sdatrefs, methrefs, pubrefs, access)
|    |---- 3.1 fileTxt? (ATT == ID, xml:lang, source)
|    |    |---- 3.1.1 fileName? (ATT ==ID, xml:lang, source)
|    |    |---- 3.1.2 fileCont? (ATT == ID, xml:lang, source)
|    |    |---- 3.1.3 fileStrc? (ATT == ID, xml:lang, source, type)
|    |    |    |---- 3.1.3.1 recGrp* (ATT == ID, xml:lang, source, recGrp, rectype, keyvar, recidvar, rtypeloc, rtypewidth, rtypev
|    |    |    |    |---- 3.1.3.1.1 labl*     (ATT == ID, xml:lang, source, level, vendor)
|    |    |    |    +---- 3.1.3.1.2 recDimnsn? (ATT == ID, xml:lang, source, level)
|    |    |    |         |---- 3.1.3.1.2.1 varQnty?  (ATT == ID, xml:lang, source)
|    |    |    |         |---- 3.1.3.1.2.2 caseQnty? (ATT == ID, xml:lang, source)
|    |    |    |         +---- 3.1.3.1.2.3 logRecL?  (ATT == ID, xml:lang, source)
|    |    |    +---- 3.1.3.2 notes* (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
|    |    |---- 3.1.4 dimensns? (ATT == ID, xml:lang, source)
|    |    |    |---- 3.1.4.1 caseQnty*  (ATT == ID, xml:lang, source)
|    |    |    |---- 3.1.4.2 varQnty*   (ATT == ID, xml:lang, source)
|    |    |    |---- 3.1.4.3 logRecL*   (ATT == ID, xml:lang, source)
|    |    |    |---- 3.1.4.4 recPrCas*  (ATT == ID, xml:lang, source)
|    |    |    +---- 3.1.4.5 recNumTot* (ATT == ID, xml:lang, source)
|    |    |---- 3.1.5 fileType?  (ATT == ID, xml:lang, source, charset)
|    |    |---- 3.1.6 format?    (ATT == ID, xml:lang, source)
|    |    |---- 3.1.7 filePlac?  (ATT == ID, xml:lang, source)
|    |    |---- 3.1.8 dataChck*  (ATT == ID, xml:lang, source)
|    |    |---- 3.1.9 ProcStat?  (ATT == ID, xml:lang, source)
|    |    |---- 3.1.10 dataMsng? (ATT == ID, xml:lang, source)
```

DDI "Codebook" – Data Description (extract)

```
|---- 4.0 dataDscr* (ATT == ID, xml:lang, source)
|    |---- 4.2 var* (ATT == ID, xml:lang, source, name, wgt, wgt-var, weight, qstn, files, vendor, dcml, intrvl, rectype, sdatrefs
|    |    |---- 4.2.1 location*   (ATT == ID, xml:lang, source, StartPos, EndPos, width, RecSegNo, fileid, locMap)
|    |    |---- 4.2.2 labl*       (ATT == ID, xml:lang, source, level, vendor)
|    |    |---- 4.2.6 respUnit?   (ATT == ID, xml:lang, source)
|    |    |---- 4.2.7 anlysUnit?  (ATT == ID, xml:lang, source)
|    |    |---- 4.2.8 qstn*       (ATT == ID, xml:lang, source, qstn, var, seqNo, sdatrefs)
|    |    |    |---- 4.2.8.1 preQTxt*  (ATT == ID, xml:lang, source)
|    |    |    |---- 4.2.8.2 qstnLit*  (ATT == ID, xml:lang, source, sdatrefs)
|    |    |    |---- 4.2.8.3 postQTxt* (ATT == ID, xml:lang, source)
|    |    |    |---- 4.2.8.4 forward*  (ATT == ID, xml:lang, source, qstn)
|    |    |    |---- 4.2.8.5 backward* (ATT == ID, xml:lang, source, qstn)
|    |    |    +---- 4.2.8.6 ivuInstr* (ATT == ID, xml:lang, source)
|    |    |---- 4.2.11 undocCod*  (ATT == ID, xml:lang, source)
|    |    |---- 4.2.12 universe*  (ATT == ID, xml:lang, source, level, clusion)
|    |    |---- 4.2.18 catgry*    (ATT == ID, xml:lang, source, missing, missType, country, sdatrefs, other, total)
|    |    |    |---- 4.2.18.1 catValu? (ATT == ID, xml:lang, source)
|    |    |    |---- 4.2.18.2 labl*    (ATT == ID, xml:lang, source, level, vendor)
|    |    |    |---- 4.2.18.3 txt*     (ATT == ID, xml:lang, source, level, sdatrefs)
|    |    |---- 4.2.19 codInstr*  (ATT == ID, xml:lang, source)
|    |    |---- 4.2.20 verStmt*   (ATT == ID, xml:lang, source)
|    |    |---- 4.2.21 concept*   (ATT == ID, xml:lang, source, vocab, vocabURI)
|    |    |---- 4.2.23 varFormat? (ATT == ID, xml:lang, source, type, formatname, schema, category, URI)
|    |---- 4.3 notes*     (ATT == ID, xml:lang, source, type, subject, level, resp, sdatrefs)
```

DDI "Codebook" – Other Material (extract)



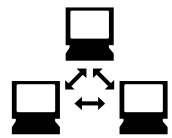
Preparation Process – Publication

- 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

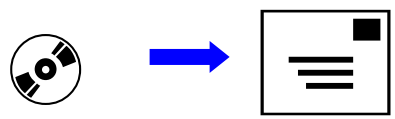
- **online access** to information is provided via the website

  http://www.clearingstelle-verkehr.de

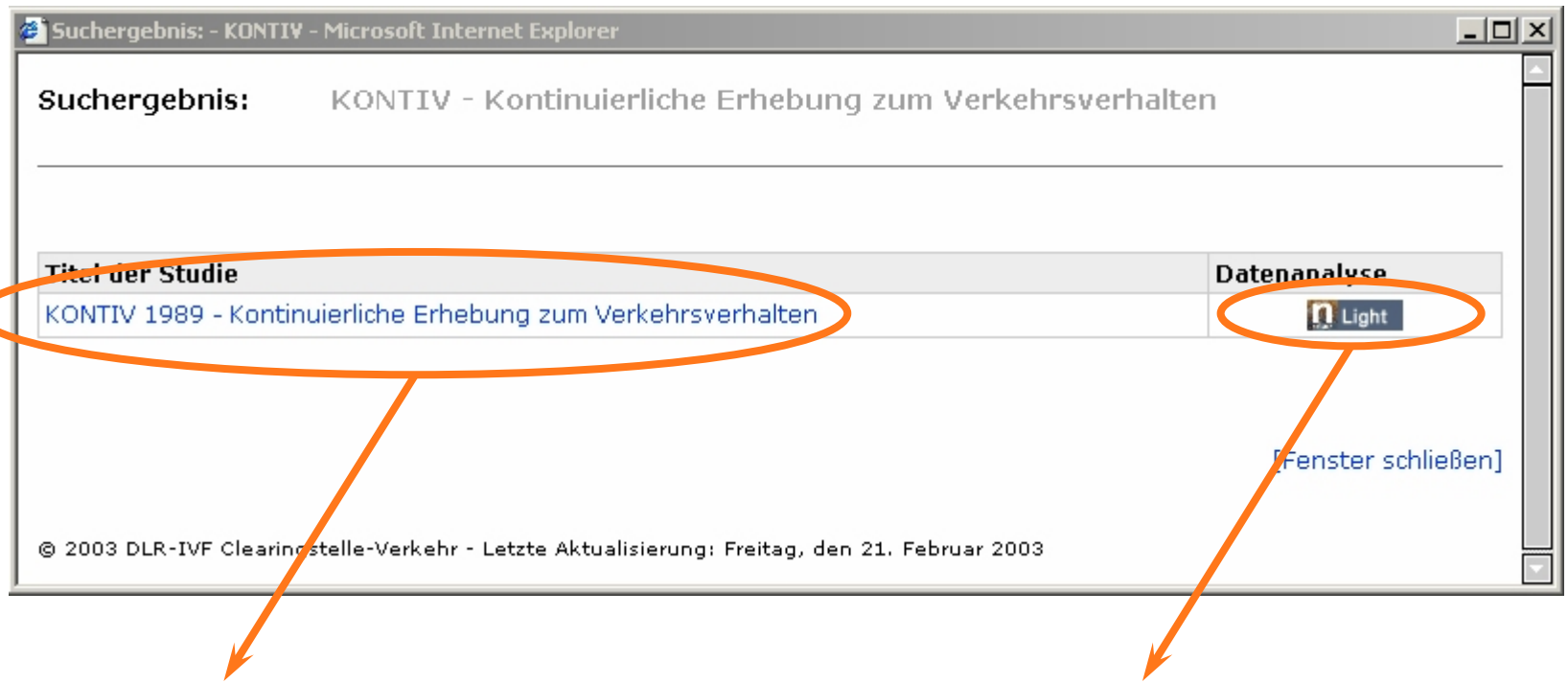
  

  - downloadable metadata documents in different formats
  - list of related publications (if available)
  - hyperlink to NESSTAR server (access to data, if available and not restricted)
  - hyperlinks to related websites (if available)

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

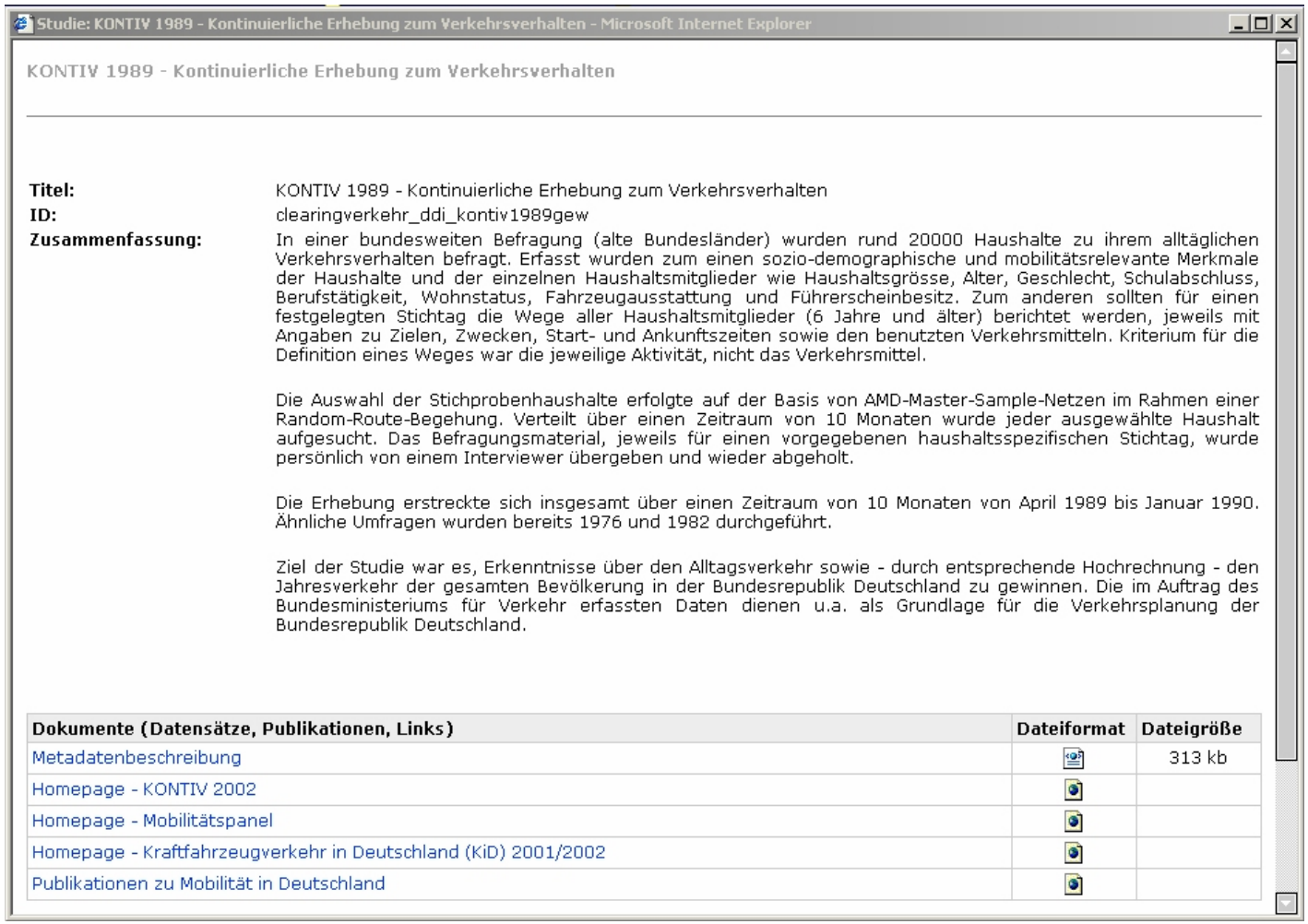
Dissemination - Search Result



link to available **metadata**

link to available **data** (NESSTAR)

Dissemination: List of Metadata Documents



Dissemination: NESSTAR - Light Explorer

Networked

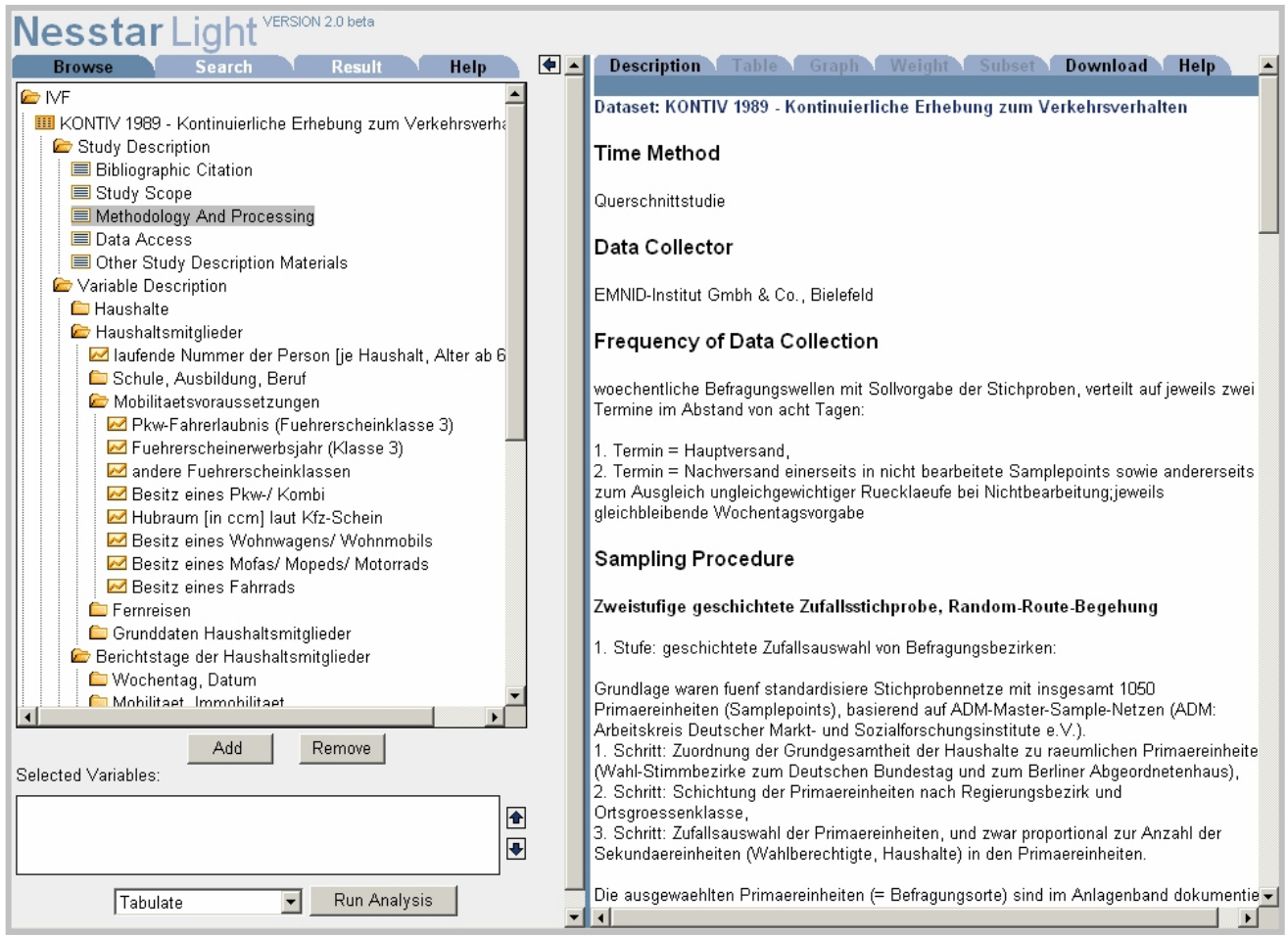
Social

Science

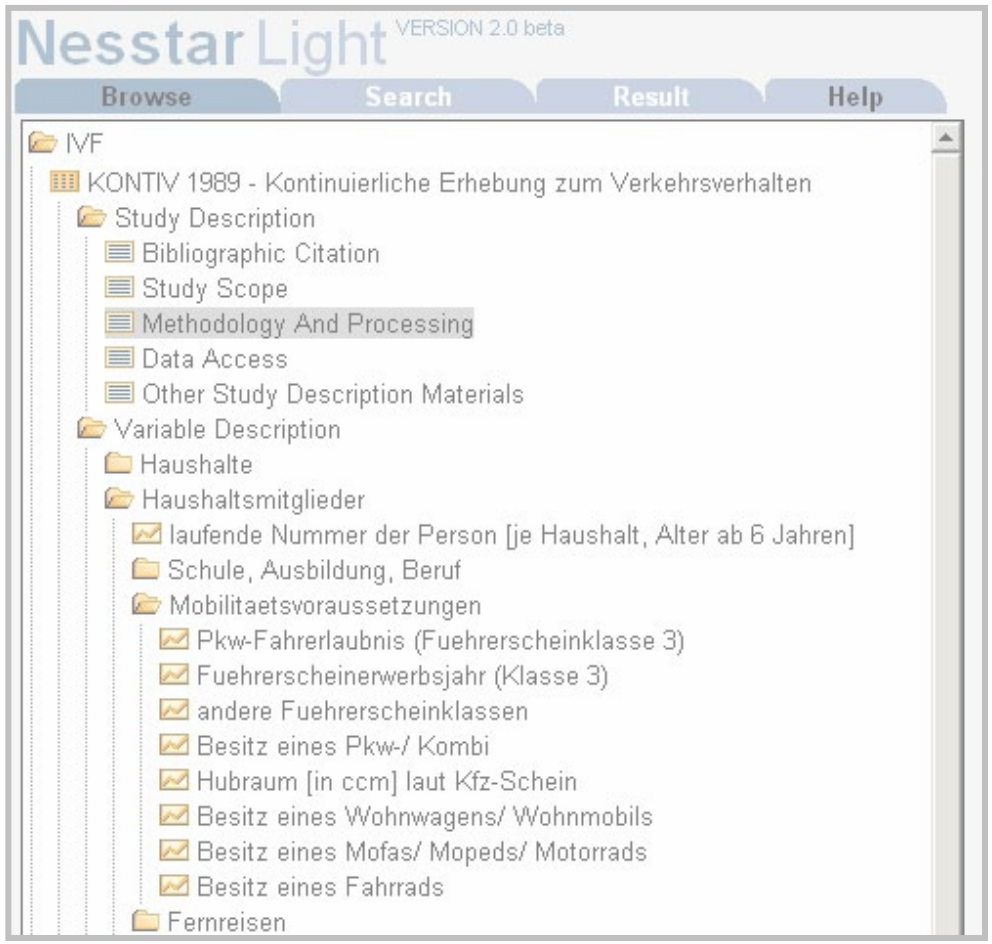
Tools

And

Resources

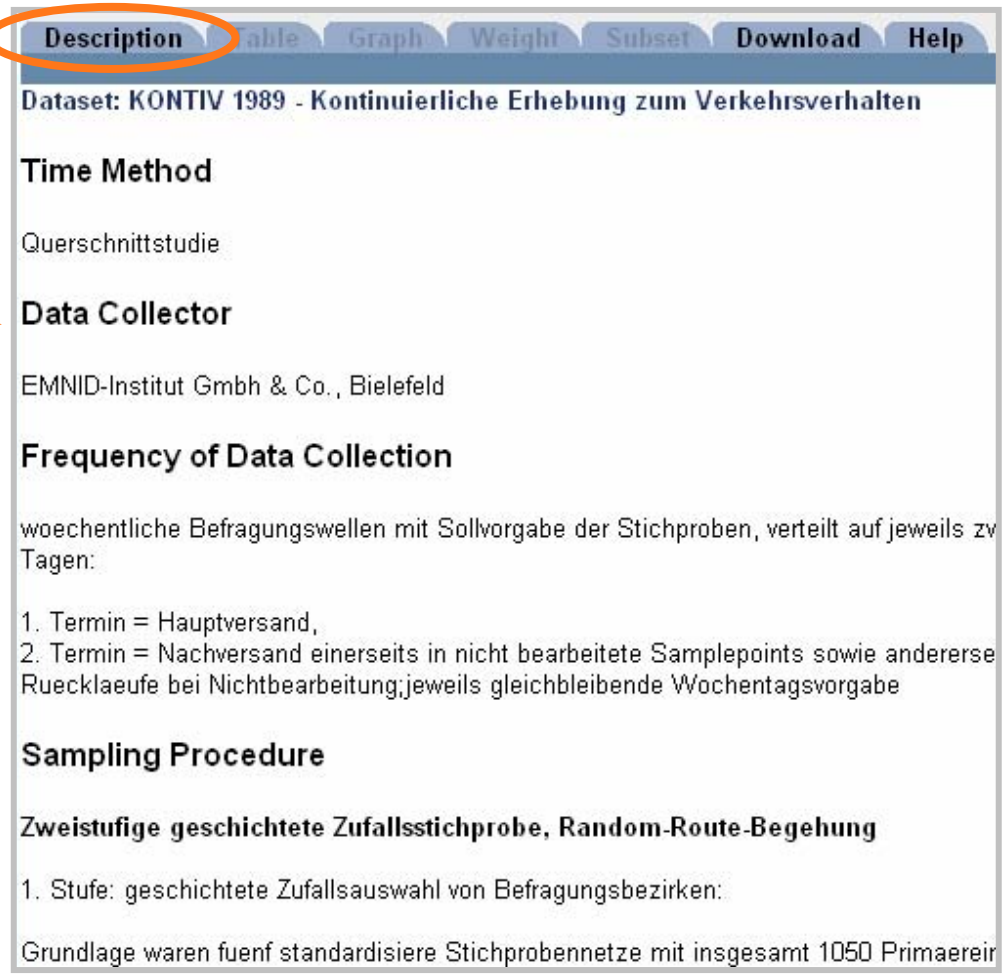


NESSTAR Light Explorer – Metadata I



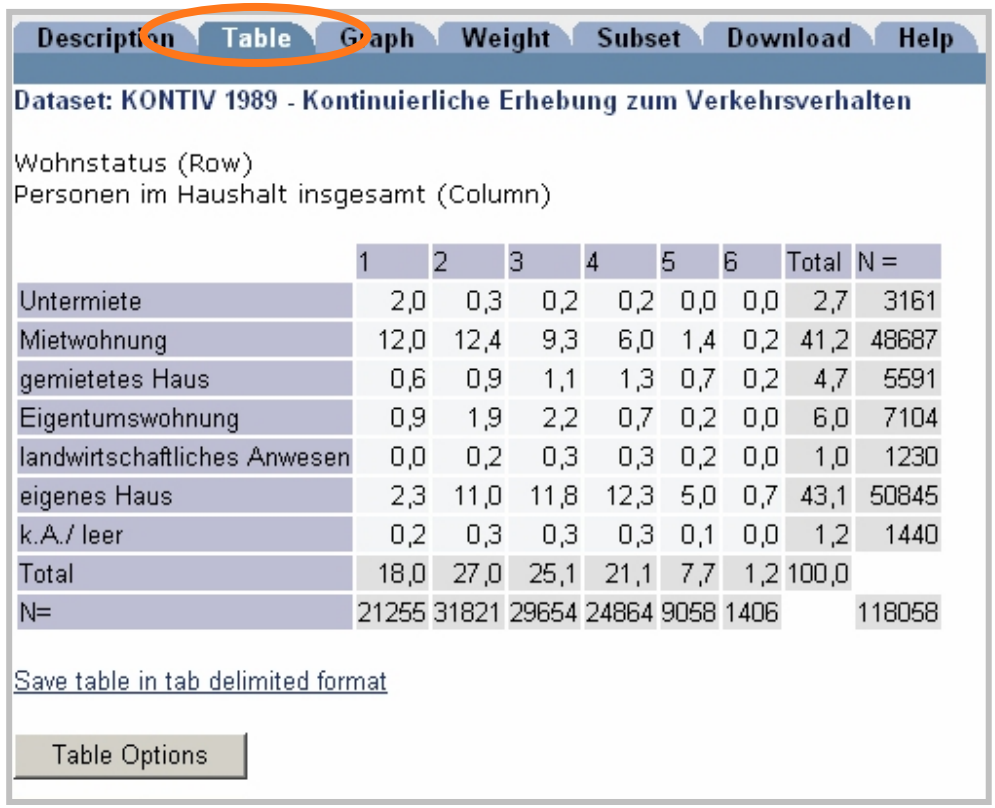
- access to metadata via **explorer tree**
- structured **representation** of underlying **DDI Document Type Definition** (XML-metadata file)

NESSTAR Light Explorer – Metadata II



- **metadata** associated with a particular node are displayed in the d*escription* area

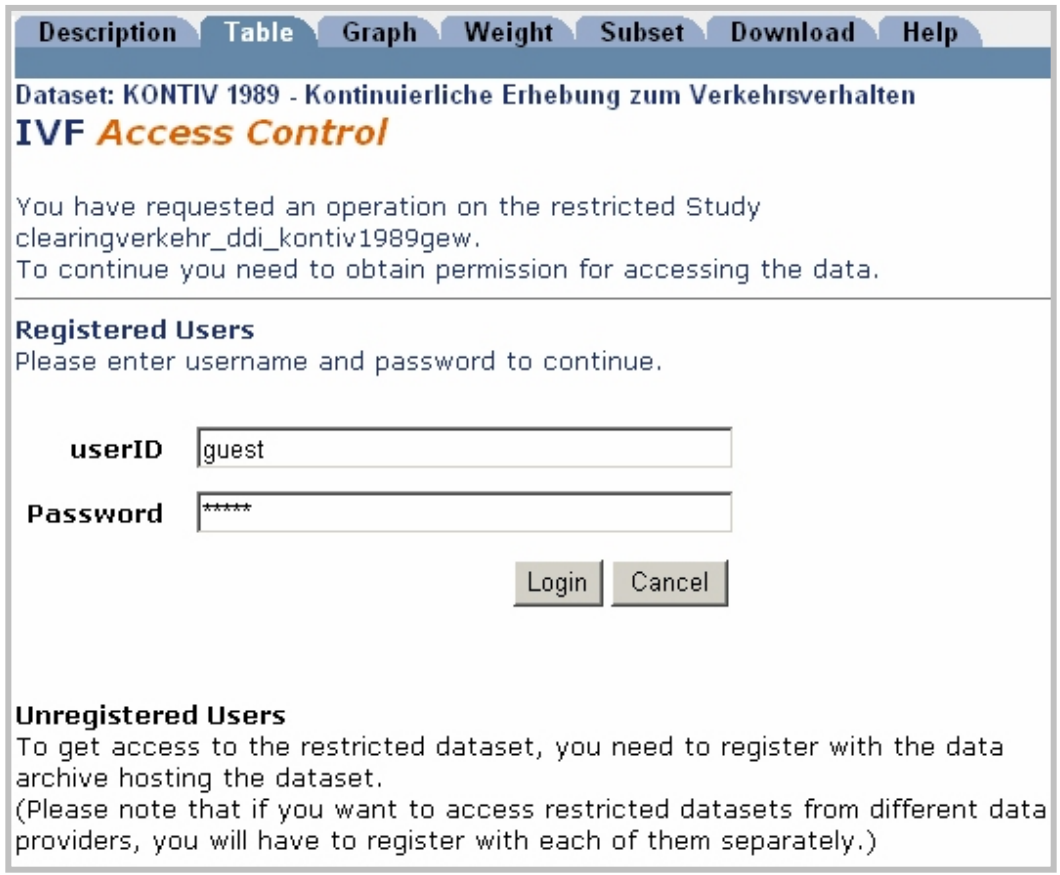
NESSTAR Light Explorer – Data Access



	1	2	3	4	5	6	Total	N =
Untermiete	2,0	0,3	0,2	0,2	0,0	0,0	2,7	3161
Mietwohnung	12,0	12,4	9,3	6,0	1,4	0,2	41,2	48687
gemietetes Haus	0,6	0,9	1,1	1,3	0,7	0,2	4,7	5591
Eigentumswohnung	0,9	1,9	2,2	0,7	0,2	0,0	6,0	7104
landwirtschaftliches Anwesen	0,0	0,2	0,3	0,3	0,2	0,0	1,0	1230
eigenes Haus	2,3	11,0	11,8	12,3	5,0	0,7	43,1	50845
k.A./ leer	0,2	0,3	0,3	0,3	0,1	0,0	1,2	1440
Total	18,0	27,0	25,1	21,1	7,7	1,2	100,0	
N=	21255	31821	29654	24864	9058	1406		118058

- results of basic **analyses**, e.g. cross-tabs, are displayed in the t*able* 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

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
E-Mail	-	info@clearingstelle-verkehr.de

Thanks for your attention !

End