Preferred citation style for this presentation

Chalasani, V.S. (2007) Archiving transport data using NESSTAR tools, IVT-internal workshop, ETH Zürich, March 2007.

Archiving transport data using NESSTAR tools

V.S. Chalasani

IVT
ETH
Zürich

March 2007

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Overview

Need for archiving data

Data archiving: background

Metadata and DDI

Travel data archiving

NESSTAR

ETH Travel Data Archive

Demo – Travel data archiving

Transport DDI

Tools for thought

"Much more time went into finding or obtaining information than into digesting it."

Dr. J.C.R. Licklider

$$\text{Maximize} \left(\frac{\text{time spent on digesting and thinking}}{\text{time spent on finding and accessing}} \right)$$

Data archiving: Is it necessary??

Organizations are under-performing, not because of lack of data and information but because of their inability to find, manage and digest the wealth of data that already exists in their environment.

Data are normally under-exploited. Too much data are locked up in drawers, on private hard-disks, in closed information system and in proprietary formats - only accessible to a small community of users.

The value of data increases through use. Unlocking existing data might therefore be a profitable activity:

- increase the value of investments in data assets
- reduce the need for further spending on data collection

Data archiving: Possible user scenarios

…a user analysing a group of variables in dataset X would like to know if there are similar datasets from other countries that could be used for a comparative study

…he/she would also like to have an overview of knowledge products (papers, articles etc.) based on this study and even to browse these objects if they are available on-line

…moreover he/she would like to contact other researchers that have used the dataset to hear about their experiences

Travel data archiving: user scenarios for future – Possible??

...”The Travel data Web”, the Web that creats a platform to share all the digital resources in such a way that the main objective of the Information Socienty Technologies (IST) be fulfilled.

…finding a problem with one of the variables, he/she writes a note and appends it to the ”user experience-section” of the metadata to allert future users (she also leaves her e-mail address to allow them to contact him/her

...and when the research paper is ready and published in an on-line journal, links to the dataset is added to allow future users to revisit his/her analysis

Data observatory requirements

- Data
  - Survey
  - Indicators
  - Administrative
  - Geographical
- Tools
  - Finding and sorting
  - Browsing
  - Analysing
  - Publishing
- Text
  - Journal articles
  - User guides
  - Methodology instructions
- People
  - Email
  - Conferences
  - Experts
  - Discussion lists

Data archiving: Presenting and preserving the data

Data – A set of numbers with hindered information
Information – Knowledge hidden in the data or numbers
Metadata – Data about data

Data Archiving:

- Presents and preserves the data
- Meets the data observatory requirements

Data presentation:

- Better understanding with less effort
- Easy in identifying the mistakes

Data preservation:

- Secondary use
- Unchanged data in quality

Data Archiving:Standards

Consortiums / Counsels :

- Inter-university Consortium for Political and Social science Research (ICPSR, 1995)
- Council of European Social Science Data Archives (CESSDA, 1995)
- Language Independent Metadata Browsing for European Resources (LIMBER)

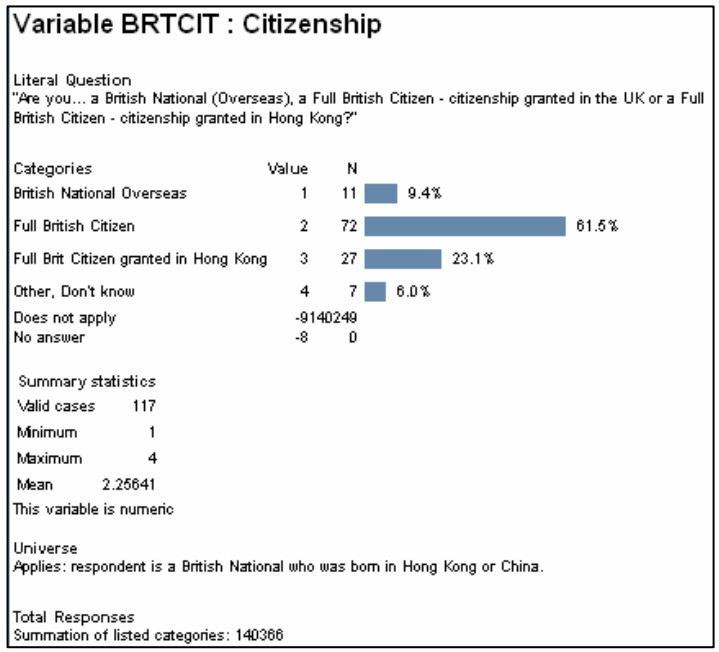
Metadata standards:

- Structure
    - Data Documentation Initiative (DDI)
    - Dublin core
    - Resource Descriptive Formation (RDF)
- Syntax (SGML, XML)
    - Standard Generalised Markup Language (SGML)
    - eXtensible Markup Language (XML)

Metadata – data about.....

1	1	4	5	13
1	1	4	5	7
1	1	4	5	4
1	1	4	5	21
1	1	4	2	7
1	1	3	4	4
1	1	4	5	6
1	1	1	5	4
1	1	2	5	1
3	1	1	3	1
3	1	9	3	16
3	1	9	2	4
3	1	9	9	19
3	3	2	9	4
3	1	9	3	99

Unlabeled stuff



Variable BRTCIT : Citizenship

Literal Question
"Are you... a British National (Overseas), a Full British Citizen - citizenship granted in the UK or a Full British Citizen - citizenship granted in Hong Kong?"

Categories	Value	N	
British National Overseas	1	11	9.4%
Full British Citizen	2	72	61.5%
Full Brit Citizen granted in Hong Kong	3	27	23.1%
Other, Don't know	4	7	6.0%
Does not apply	-9	140249	
No answer	-8	0	

Summary statistics

Valid cases	117
Minimum	1
Maximum	4
Mean	2.25641

This variable is numeric

Universe
Applies: respondent is a British National who was born in Hong Kong or China.

Total Responses
Summation of listed categories: 140366

Labeled stuff

The bean example is taken from: A Manager's Introduction to Adobe eXtensible Metadata Platform, http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf

Users of m

Metadata for

- .... any
  asses
  them i

Metadata for

- ....any
  softwa
  objects
  of a human

An example from the world of statistics

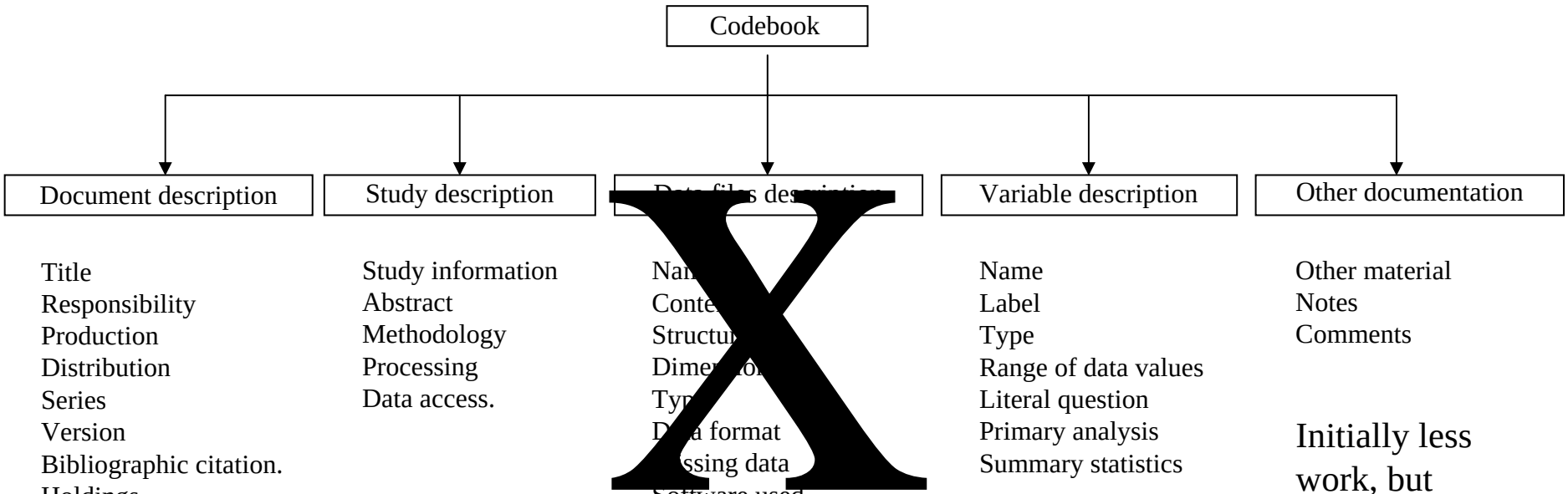
Human readable information about data quality:
information about sampling methods and procedures allowing a human reader to assess whether a statistical resource has been created according to a certain standard

Machine understandable information about data quality:
• sample size
• response rate
• sampling methodology (according to a specific classification of methodologies – a controlled vocabulary identified by a namespace where the meaning of the different terms are well defined)
• the number of publications produced on the basis of the resource
• ratings (given by persons and organizations according to a well defined rating system – allowing a user to instruct a software agent to skip resources where the quality score given by a defined set of trusted organizations are below a certain limit)

Metadata meant for humans versus machines are often just different representations of the same type of information.

Data Documentation Initiative (DDI)

Currently “DDIalliance” is working on DDI 3.0



Nesstar - background

Originally funded by EC under the Electronic Publishing Program

- NESSTAR (1998-1999), 4th Framework Program
- FASTER (2000-2001), 5th Framework Program

Nesstar Limited established in June 2001

- Owned jointly by The Norwegian Social Science data Services (NSD) and the UK Data Archive (through the University of Essex)
- Offices in Colchester (UK), Bergen (Norway) and Ottawa (Canada)

NESSTAR software suite – data archiving tools

- Nesstar Server 3.50
- Nesstar Publisher 3.50
- Nesstar Hierarchy Builder 3.50
- Nesstar Cube Builder 3.50

Nesstar - a data sharing technology

Designed to meet the requirements of organization with a need to share data internally among its researchers and analysts, or externally to a wider audience

A fully Web-based technology – it facilitates sharing of statistical resources more or less in the same way as the “standard” Web enables sharing of documents and pictures.

It allows you to locate relevant statistical data across several locations using standard search methods.

It allows you to browse information (metadata) about these data.

It allows you to analyse and visualise data in a standard web-browser.

….and to download data in the format of your favourite statistical package.

You can even bookmark resources or the result of a statistical analysis (a table, graph or map) in the same way as you are used to with normal Web resources.

NESSTAR Data Archiving Tools

NESSTAR Server:

- Presents and preserves the data.
- Restricted access is possible

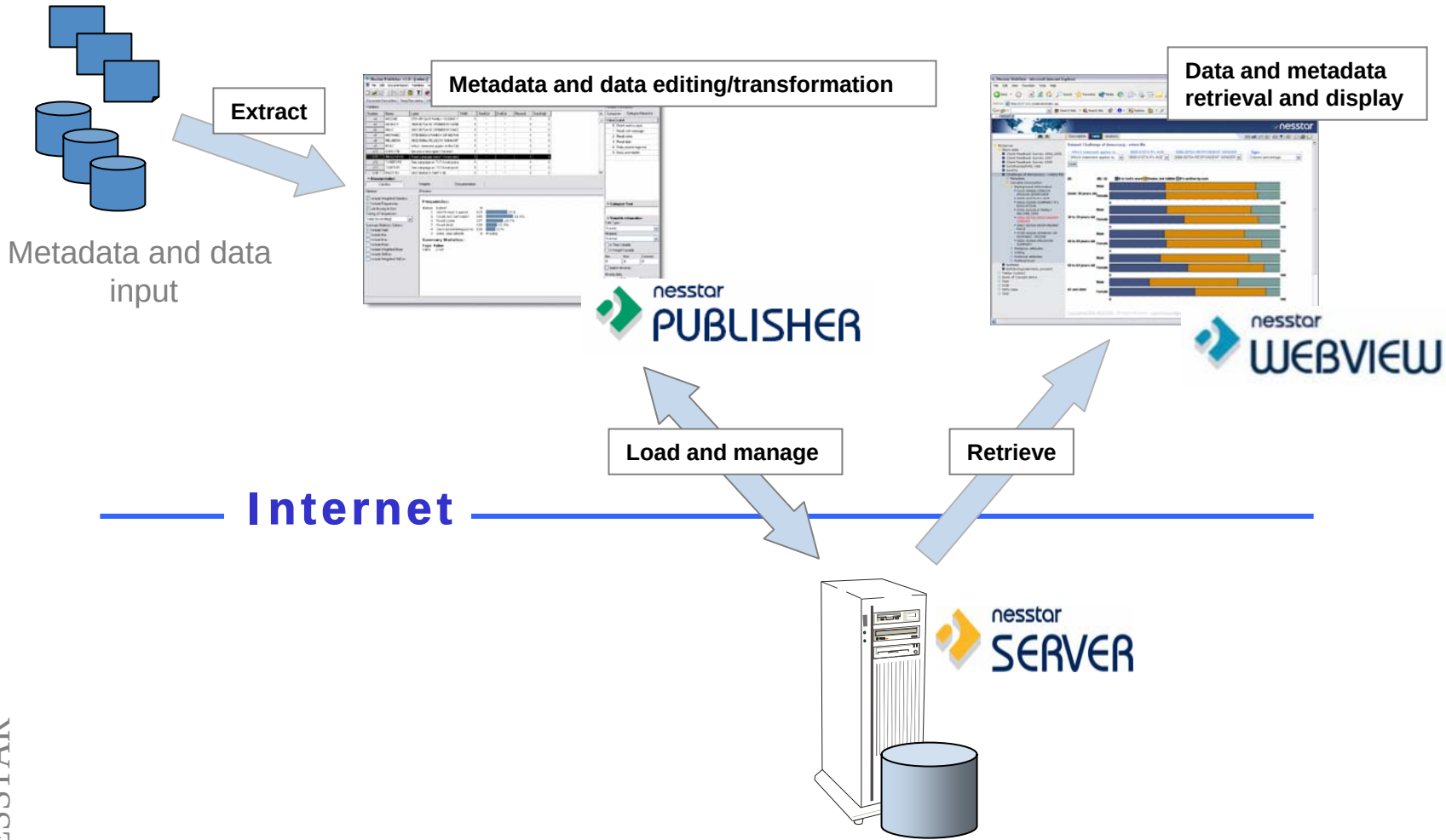
NESSTAR Publisher:

- Integrates the data and associated information with the help of XML syntax and DTD based DDI metadata structure
- Allows to publishing the archived data

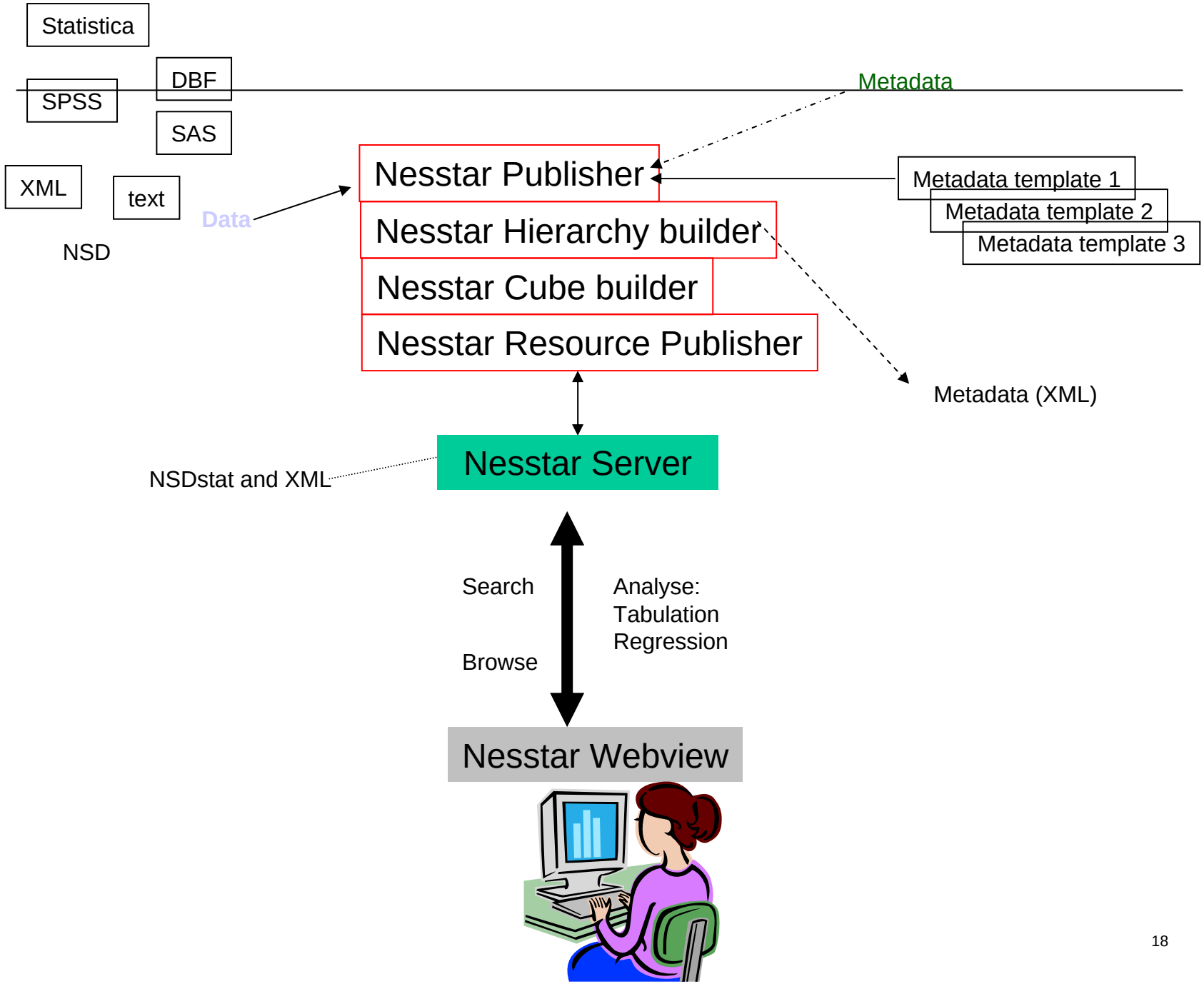
NESSTAR WebView:

- Allows for searching and browsing
- With a valid user identity secondary analysis and downloading the data is possible

Nesstar – a bird's eye view



NESSTAR


Statistica
DBF
SPSS
SAS
XML
text
NSD
Data
Metadata
Nesstar Publisher
Nesstar Hierarchy builder
Nesstar Cube builder
Nesstar Resource Publisher
Metadata template 1
Metadata template 2
Metadata template 3
Metadata (XML)
Nesstar Server
NSDstat and XML
Search
Browse
Analyse:
Tabulation
Regression
Nesstar Webview

– a distributed system

Nesstar allows you
to build an integrated
data portal / archive even in situations
where data are distributed
and maintained on several distributed
servers

nesstar WEBVIEW

Internet

nesstar SERVER

nesstar SERVER

nesstar SERVER

nesstar SERVER

nesstar SERVER

ETH Travel Data Archive

Established at IVT, ETH Zurich in May 2002

Open for public and available @http://tda.ethz.ch/webview

Datasets archived so far:

- Mobidrive 1999 (Travel dairy)
- Microcensus 2000 (Daily mobility)
- DATELINE 2001 (Long distance travel)
- Travel market Switzerland 2001 (Leisure panel survey)
- Travel module of household income and consumption (EVE) 1998
- Continuous survey about the travel behaviour of Swiss population (KEP) 2001 (Stated preference and revealed preference)
- Traffic counts [Railway and road network] (Observational data)
- 12 week leisure travel (Leisure activity survey)

Experiences in travel data archiving

Difficultes in defining the hierarchy among various data files resulted from a single survey

Cases - Microcensus

Experiences in travel data archiving

Metadata standards developed for social science data do not fully support travel data

Basic metadata structure needs additional features such as survey protocol, response rates, etc., to suit the travel data archiving

Development of an exclusive travel metadata standard "Transport DDI" is necessary

NESSTAR archiving tools are not compatible with the travel data

Transport DDI

An exclusive Metadata standard for transport data

Extension to the existing DDI

Extensions follow the DDI structure and standards

Important extension elements:

- Survey protocol
- Response rates
- Category grouping
- Metadata for non-survey data (work in progress)
- Reference searching (planned)

Transport DDI – Survey protocol

S.P. – an important element that explains the data collection methodology

DDI 3.0 completely ignored the survey protocol

Most of the transport surveys implements a complex protocol

Proposed structure for the survey protocol element

2.3 Methodology and processing

2.3.1 Survey protocol

2.3.1.1 Invitation/information*

2.3.1.2 Survey instrument*

2.3.1.3 Reminder*

2.3.1.4 Thanks*

2.3.1.5 Waves

Transport DDI – Response rate

Response rate in DDI 3.0 – Single attribute

Does not fully explain the microlevel effects of sampling procedures

Proposed response rate element :

2.3.4 Data appraisal information

2.3.4.1 Responses

2.3.4.1.1 Complete interview
2.3.4.1.2 Partial interview
2.3.4.1.3 Refusal and break-off
2.3.4.1.4 Non-contact
2.3.4.1.5 Other
2.3.4.1.6 Response rate
2.3.4.1.7 Cooperation rate
2.3.4.1.8 Refusal rate
2.3.4.1.9 Contact rate

Transport DDI – Category grouping

Category grouping - often used to reduce the number of categories to publish the data for better understanding

The “DDI alliance” technical committee is reviewing the category grouping to include in the forthcoming DDI (i.e. DDI 3.0)

The additional element that represents the category grouping

4.3.17 Category group

- 4.3.17.1 Category group label
- 4.3.17.2 Category group statistic
- 4.3.17.3 Category group text

Transport DDI – Metadata for non-survey data

To develop a metadata standard to document the non-survey data such as results from statistical models, input data for different models, etc.

Nesstar resource publisher – generates metadata for resource data

Currently reviewing various reports on metadata for statistical models (e.g.: OPUS methodology, Metanet end report, etc.)

Plans:

- Add a mechanism to refer / search references
- A platform for end-user feedback (e.g.: blog..)

Vision

A worldwide federated transport database to cover the transport data (survey and non-survey) and minimise the information loss

Conclusions

Data archiving helps in:

- Harmonising the data (individual and historical)
- Minimising the data errors
- Documenting with a standard metadata
- Presenting the data on common platform
- Preserving the data and metadata for future
- Minimising the information loss during transfers

Travel data archiving is growing up rapidly and the existing DDI is not fully supporting transport metadata

Transport DDI – an exclusive metadata standard for transport data

- An extension to the existing DDI
- Need more diverse data to extend further
- Two metadata extensions (survey protocol and response rates) are done

Need for a federated transport database to share the expertise and data exchange