OECD OCDE

Enhancing quality of statistics by applying metadata standards

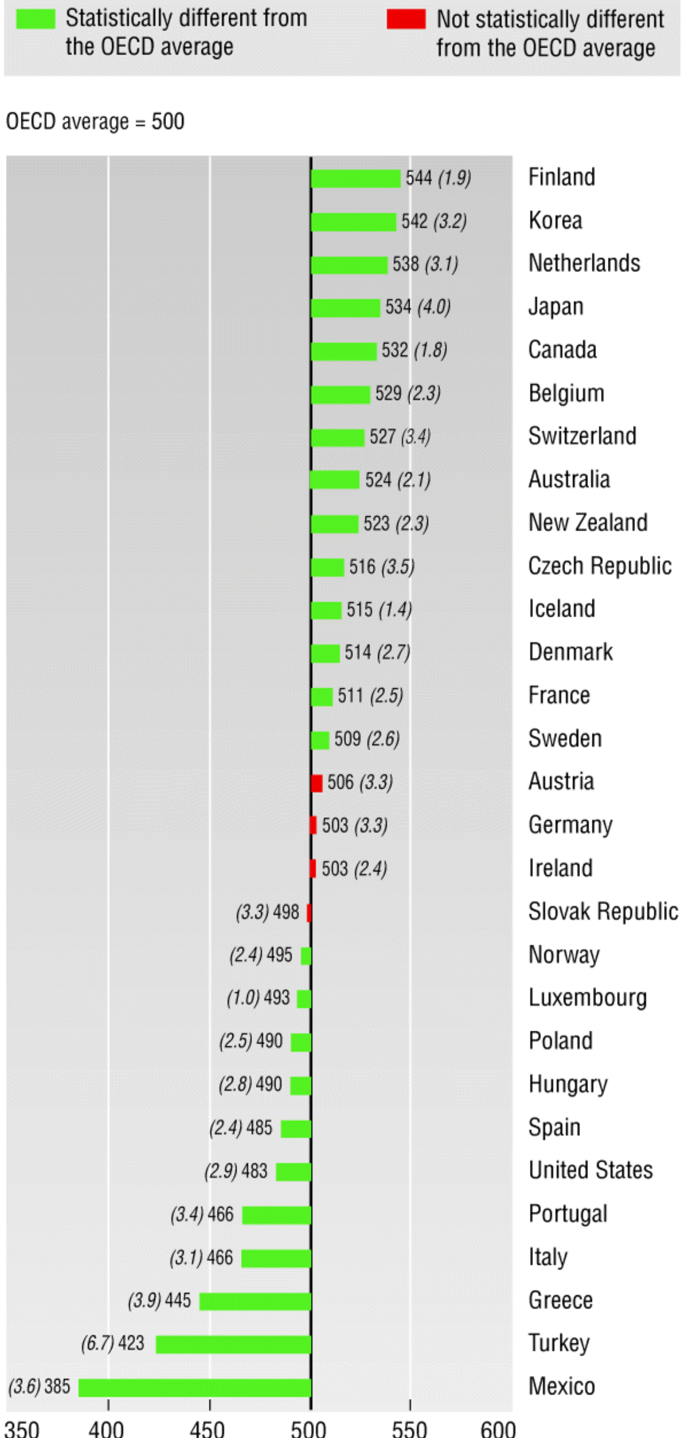
Enrico Giovannini & Lars Thygesen

April 2006

OECD mission

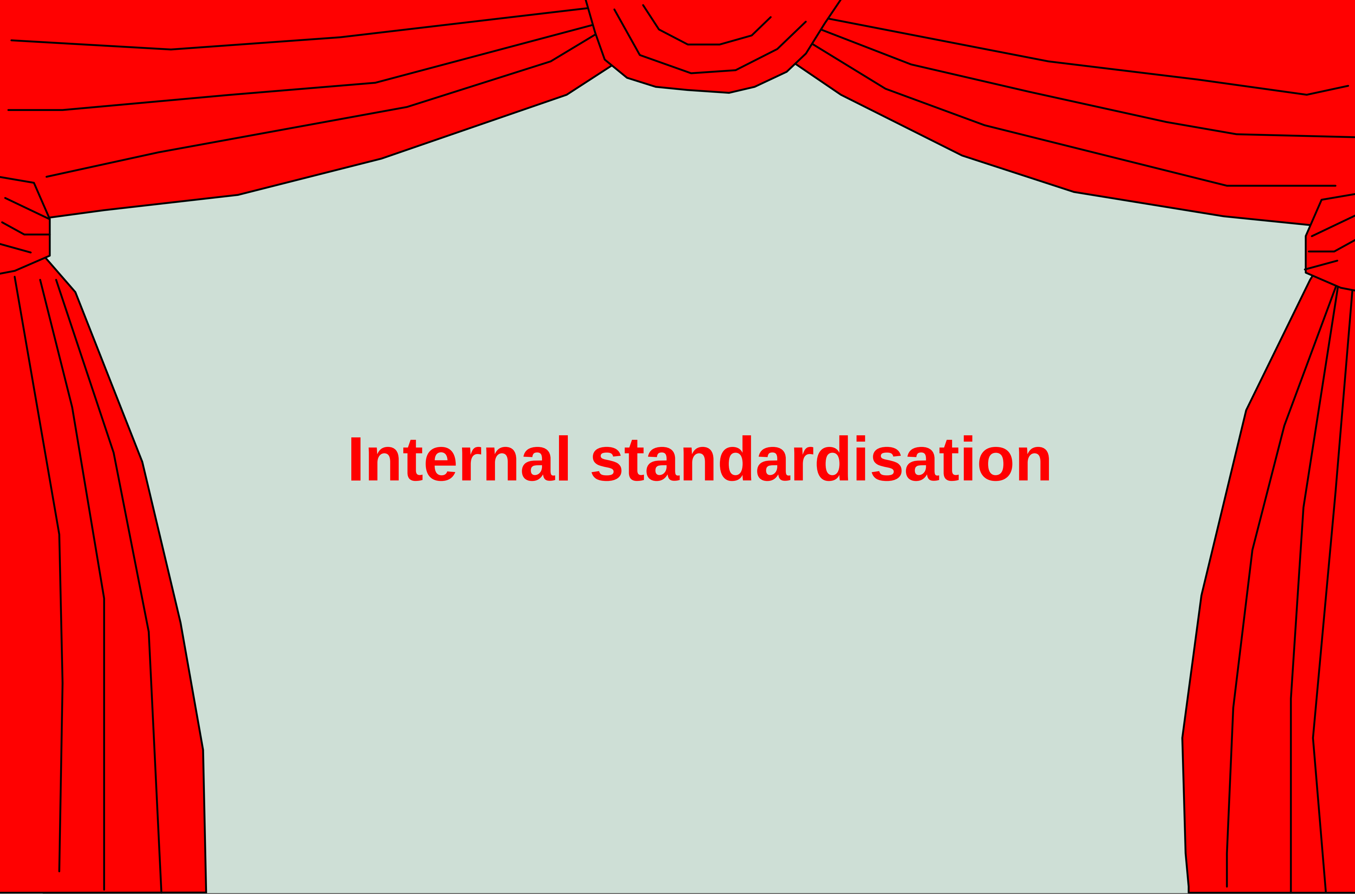
- 30 member countries
- democracy and market economy
- rich
- analyse and compare policies
- identify good outcomes – and less good
- e.g. country economic reviews
- e.g. education

PISA 2003: Performance in mathematics



Statistics in OECD

- Statistical information on society
  - many sectors: economy, labour, health, education, governance...
- tool for policy analysis
- compare countries
- authority & quality



Internal standardisation

Some problems

- how compare? each country has its own way of defining things
- each country *is* different: institutions, etc.
  - e.g. compare immigration?
- we must understand the differences in order to make good judgment

What can we do about it?

- International standardisation work
- concepts
- classifications
- manuals
- good metadata

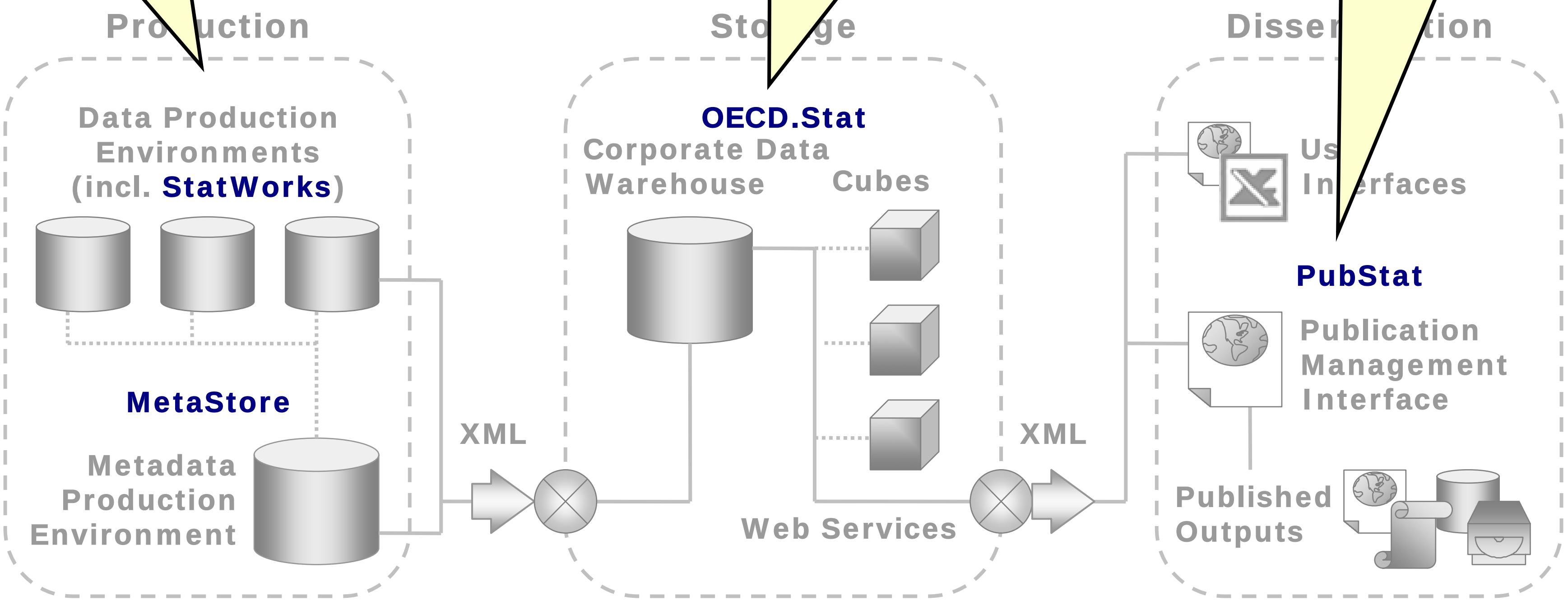
In-house standardisation - OECD

- De-centralised system:
  - Directorates & Committees
  - databases
- Quality framework
  - 7 dimensions
  - best practices
  - quality reviews
- Metadata Guidelines
  - 41 metadata items
  - attachment levels
  - redundancy
- OECD Statistical Information System

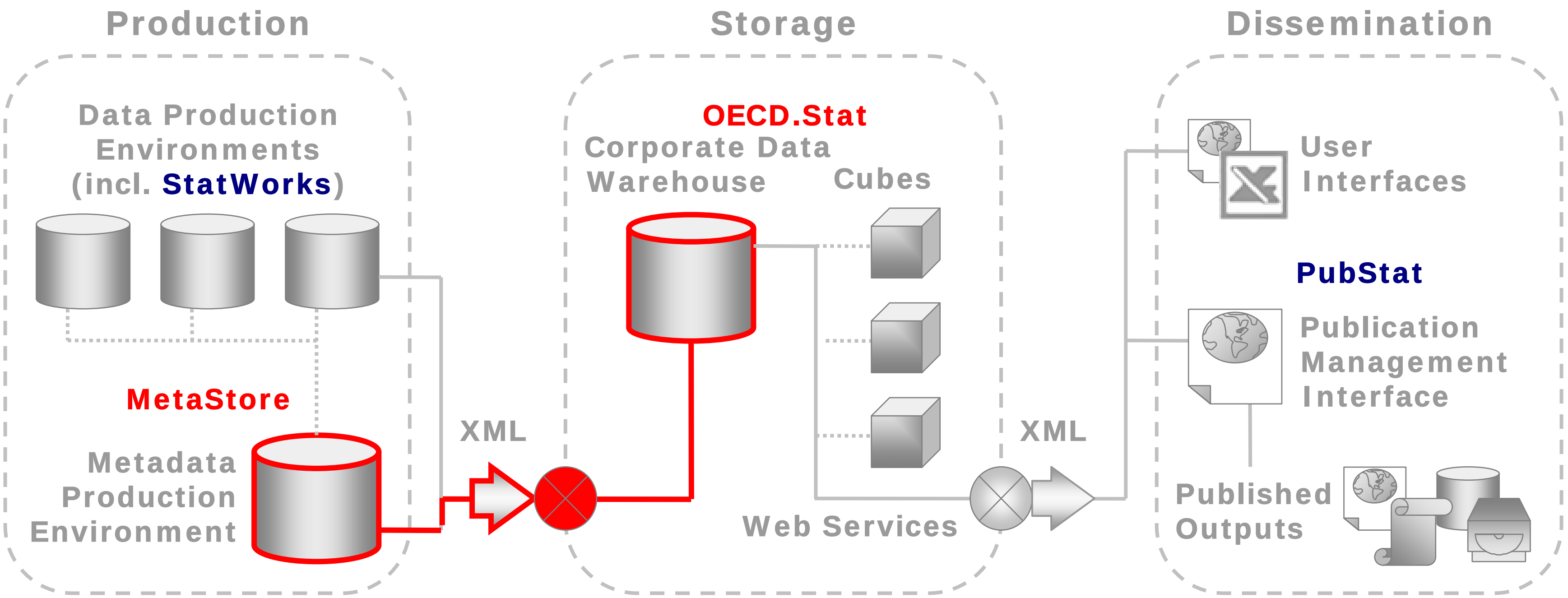
- Data collection & preparation

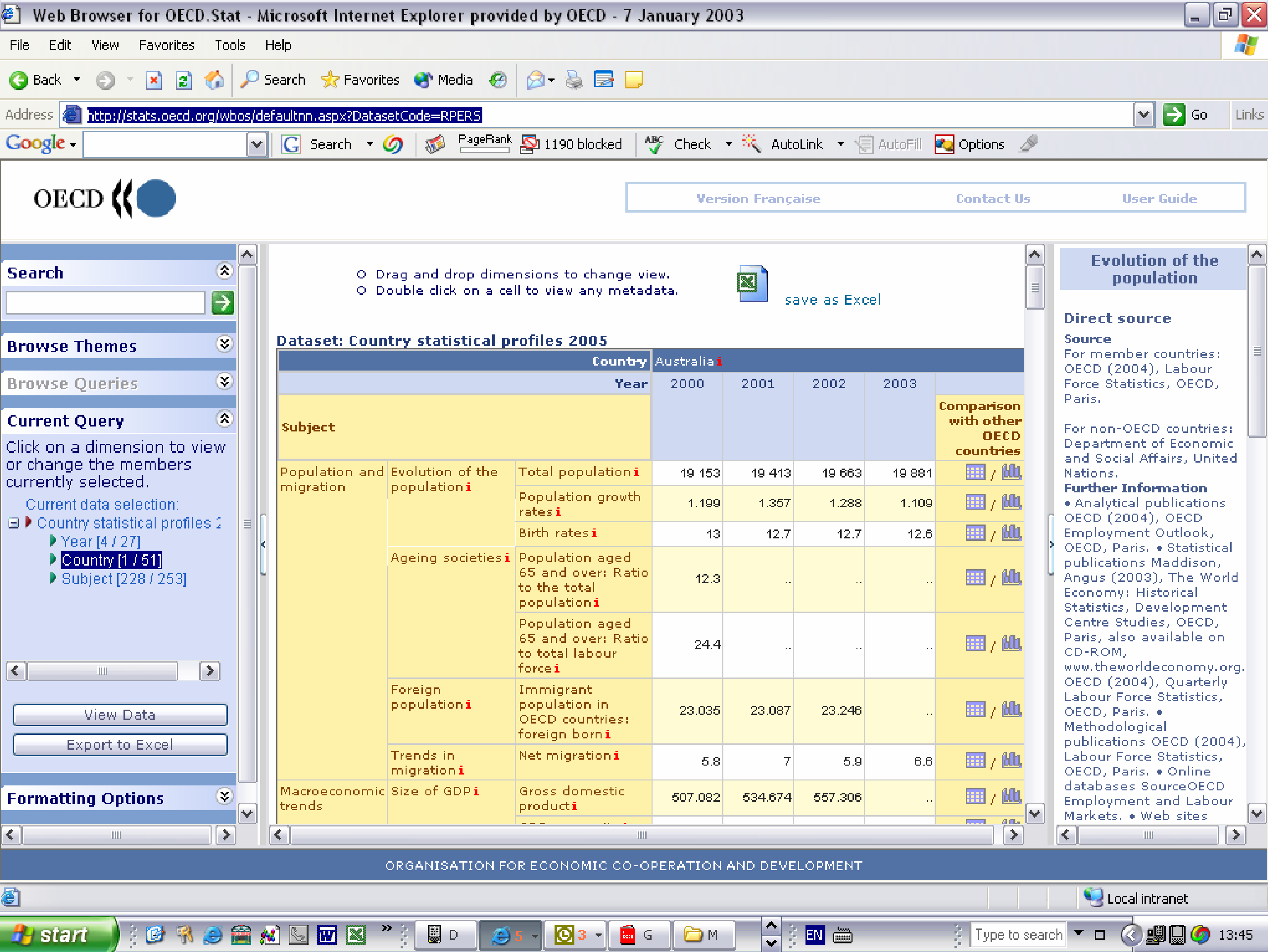
One data warehouse - all statistics and metadata

- One source, many formats & media



The OECD Statistical Information System




Web Browser for OECD.Stat - Microsoft Internet Explorer provided by OECD - 7 January 2003
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://stats.oecd.org/wbos/defaultnn.aspx?DatasetCode=RPERS Go Links
Google Search PageRank 1190 blocked Check AutoLink AutoFill Options
OECD
Version Française Contact Us User Guide
Search
Browse Themes
Browse Queries
Current Query
Formatting Options
View Data
Export to Excel
save as Excel
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
Local intranet
start
EN
Type to search
13:45

Click on a dimension to view or change the members currently selected.

Current data selection:

- Country statistical profiles 2
  - Year [4 / 27]
  - Country [1 / 51]
  - Subject [228 / 253]

○ Drag and drop dimensions to change view.
○ Double click on a cell to view any metadata.

Dataset: Country statistical profiles 2005

		Country	Australia i				
		Year	2000	2001	2002	2003	
Subject							Comparison with other OECD countries
Population and migration	Evolution of the population i	Total population i	19 153	19 413	19 663	19 881	▦ / ▮▮▮
		Population growth rates i	1.199	1.357	1.288	1.109	▦ / ▮▮▮
		Birth rates i	13	12.7	12.7	12.6	▦ / ▮▮▮
	Ageing societies i	Population aged 65 and over: Ratio to the total population i	12.3	..	..	..	▦ / ▮▮▮
		Population aged 65 and over: Ratio to total labour force i	24.4	..	..	..	▦ / ▮▮▮
	Foreign population i	Immigrant population in OECD countries: foreign born i	23.035	23.087	23.246	..	▦ / ▮▮▮
	Trends in migration i	Net migration i	5.8	7	5.9	6.6	▦ / ▮▮▮
Macroeconomic trends	Size of GDP i	Gross domestic product i	507.082	534.674	557.306	..	▦ / ▮▮▮

Evolution of the population

Direct source

Source

For member countries: OECD (2004), Labour Force Statistics, OECD, Paris.

For non-OECD countries: Department of Economic and Social Affairs, United Nations.

Further Information

• Analytical publications OECD (2004), OECD Employment Outlook, OECD, Paris. • Statistical publications Maddison, Angus (2003), The World Economy: Historical Statistics, Development Centre Studies, OECD, Paris, also available on CD-ROM, www.theworldeconomy.org. OECD (2004), Quarterly Labour Force Statistics, OECD, Paris. • Methodological publications OECD (2004), Labour Force Statistics, OECD, Paris. • Online databases SourceOECD Employment and Labour Markets. • Web sites

Governance structure

- Keep local ownership and responsibility
- Carrots rather than sticks
  - e.g. MetaStore not mandatory
- Attractive systems & good results to animate to follow

Different roles of Metadata

inform users about:

- which statistical data are available?
- are they useful to my purpose?
- where to find and how to retrieve? certain statistical data that they need
- how to interpret statistical data, once they are available

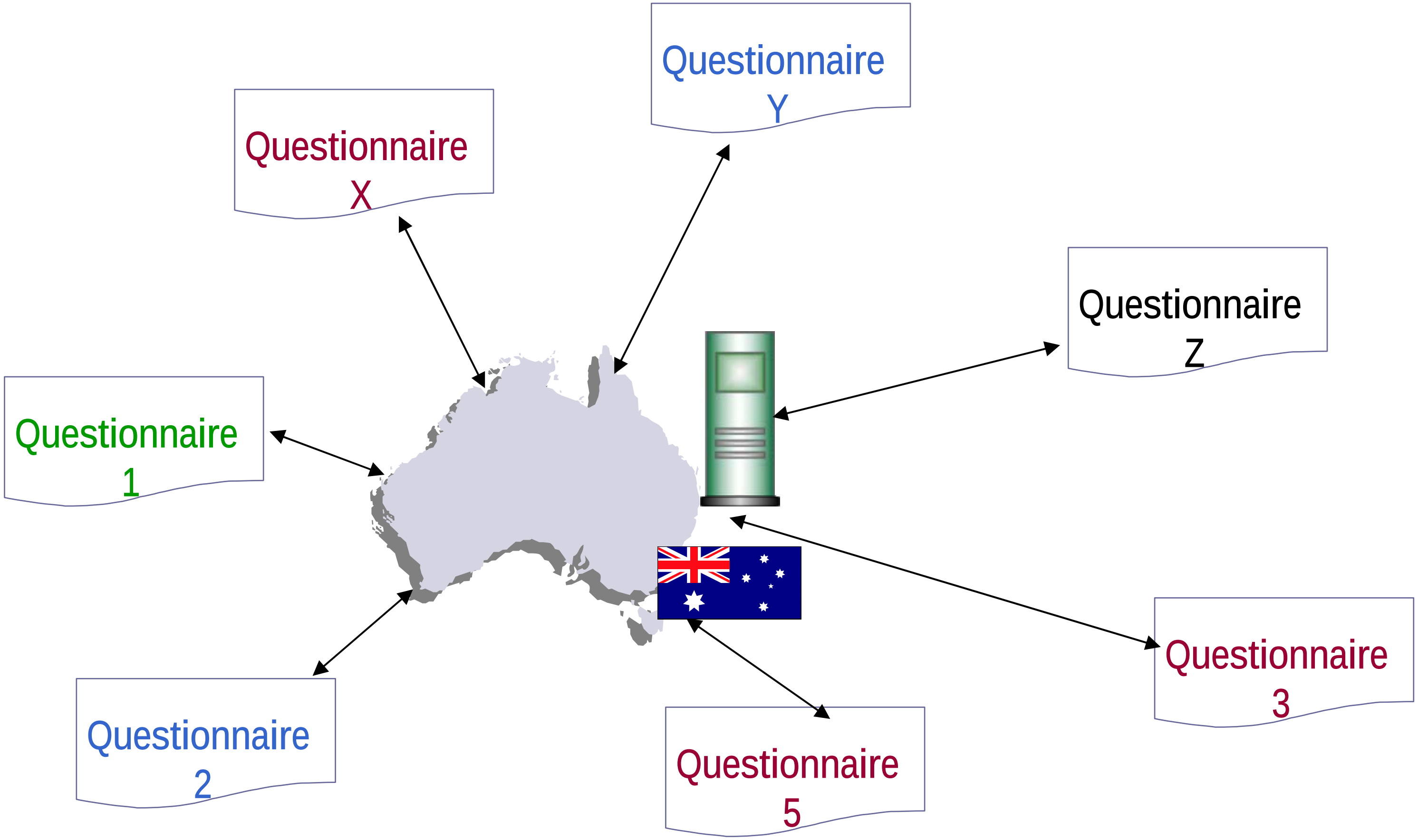
- infrastructure

External standardisation

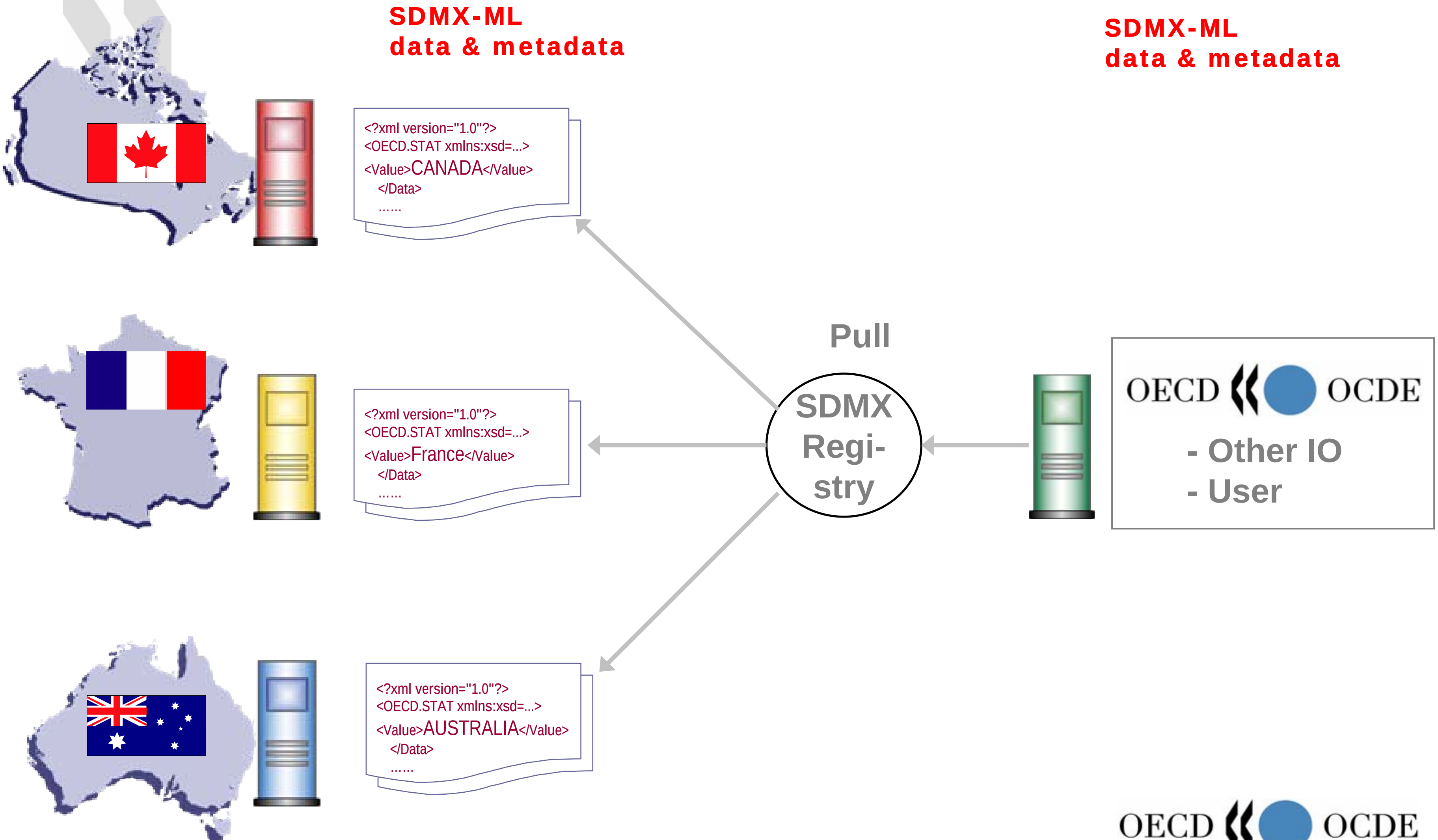
OECD's place in an international system

- reporting from 30 Member countries
  - burden on countries
- share data with other organisations
  - a wider community
- duplication of information
- is it really the same?

The National agency under bombardment



SDMX and data sharing



SDMX-ML
data & metadata

SDMX-ML
data & metadata

```
<?xml version="1.0"?>
<OECD.STAT xmlns:xsd=...>
<Value>CANADA</Value>
   </Data>
   ……
```

```
<?xml version="1.0"?>
<OECD.STAT xmlns:xsd=...>
<Value>France</Value>
   </Data>
   ……
```

```
<?xml version="1.0"?>
<OECD.STAT xmlns:xsd=...>
<Value>AUSTRALIA</Value>
   </Data>
   ……
```

Pull

SDMX Regi- stry

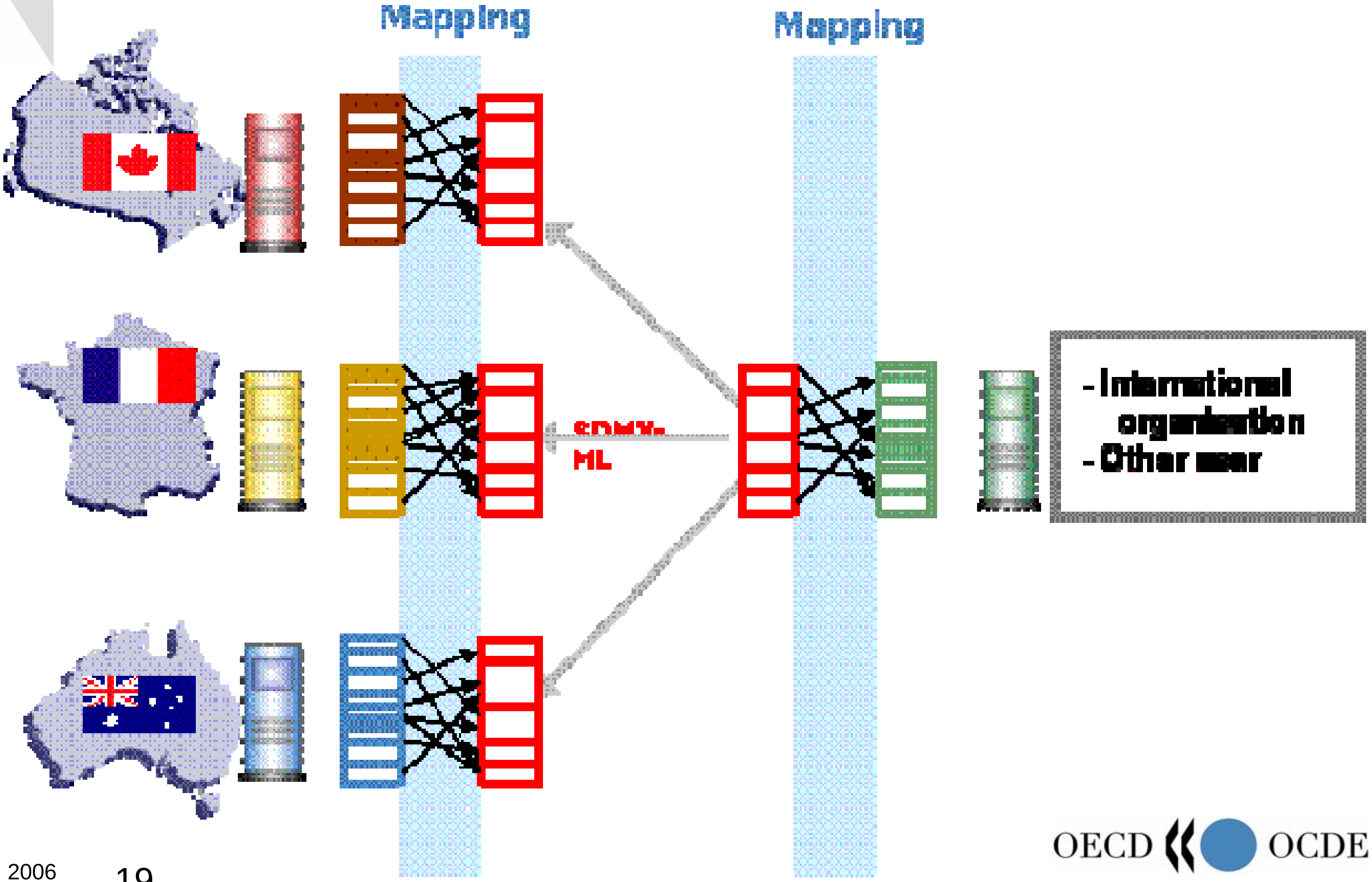
OECD OCDE

- Other IO
- User

SDMX standards

- [SDMX-ML formats for data]
- SDMX Metadata Message
- Cross-domain Concepts
- Metadata Common Vocabulary (MCV)
- ISO 17369

SDMX and metadata mapping



Future scenario

- No metadata collection necessary
- No “central” repository
- Sharing in real time
- Anyone could access
  - if authorised

Terminology problems

IT WOULD BE A PERFECT WORLD IF ONLY PEOPLE ALL OVER THE WORLD WOULD TRY TO UNDERSTAND EACH OTHER!
YOU'RE RIGHT!
THE PROBLEM IS TO GET EVERYONE TO SPEAK NORWEGIAN LIKE WE DO!!
CHRIS BROWNE
7-17
©2001 by King Features Syndicate, Inc. World rights reserved.

Terminology & interoperability

- MCV: Metadata Common Vocabulary
- Defines basic exchange terms
- Draws on most authoritative sources
- Agreed among 7 international organisations

Reference Metadata

Reference metadata

Definition: Reference metadata describe statistical concepts, methodologies for the generation of data and information on data quality.

Source: Statistical Data and Metadata Exchange (SDMX) - BIS, ECB, Eurostat, IBRD, IMF and OECD, “Framework for SDMX standards”, Version 1.0, First revision December 2004

Hyperlinks: www.sdmx.org, www.sdmx.info

Context: Reference metadata, sometimes generated, collected or disseminated separately from the data to which they refer can be relevant to all instances of data described: entire collections of data, data sets from a given country, or for a data item concerning one country and one year. Preferably, reference metadata should include all of the following: a) "conceptual" metadata, describing the concepts used and their practical implementation, allowing users to understand what the statistics are measuring and, thus, their fitness for use; b) "methodological" metadata, describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes); c) "quality" metadata, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy).

Related term:

Metadata, statistical

Subject-matter standardisation

– OECD Glossary of Statistical Terms
– 6000 terms

Metadata sharing - Challenges

- Connectivity
  - can concepts be mapped?
  - revision and transformation?
- Study metadata systems of NSOs
- Compare with systems of IOs
- What can go in between?
- Illustrate with real world examples

The treasure is worthless if you cannot find it

- use of metadata for making your data more searchable
- Key words
- Link density
- Google
- fortunately the same criteria apply as for making good metadata

Conclusions

- Metadata for
  - discovery
  - understanding
- internal standards
  - formats
  - contents
  - terminology
  - tools
- external agreements
  - formats
  - contents
  - terminology
  - tools
- mapping necessary

Conclusions #2

- Capture metadata in the process
- Avoid redundancy or conflict
  - reuse same information
  - record once, establish ownership
- Must be a top management priority
  - or else nothing happens
- Systems must prove their value
- Quality assurance framework
  - review and assess
  - enforce standards
- Interoperability does not necessarily mean *real comparability*

Advances in Survey Lifecycle Documentation – Towards a Survey Information System

Peter Ph. Mohler
ZUMA

Beth-Ellen Pennell, Karl Dinkelmann
Institute for Social Research

Paper prepared for Q2006

European Conference on Quality in Survey Statistics

ZUMA ▪ *Centre for Survey Research and Methodology*

ISR ▪ *Survey Research Center* ▪ *Survey Research Operations*

Overview

- To know or not to know
- A conceptual revolution: from documentation to knowledge management
- New tools
- An invitation to survey information

To know or not to know

- To know or not to know

  Wittgenstein (on the question what did you mean by saying x?)

  I have forgotten

Pappi (on the question about the category ‚working class' in a SES scale)

I know

What is documentation?

Preserving

Organising

&

Accumulating knowledge

How not to take note

- No research without notes
- It is not about why to take note (documentation)
- It is about how to ….

Systematically

Preserve,

Organise

&

Accumulate knowledge

Why, What, Who, When, & How to

- **Why** documentation should be paramount and standard practice?
- **What** concepts and materials are needed to adequately document a study?
- **Who** is the audience?
- **When** should documentation be built into the survey process or survey lifecycle?

- **How to document efficiently?**

Interlude: Information overload

- *Industry analysts estimate that workers*
  - *"information overload" is becoming a serious drag on productivity -- the typical worker in North America gets 10 times as much e-mail as in 1997, and that number continues to increase.*
  - *spend up to 30 percent of their working day just looking for data they need.*

Executive E-mail from Bill Gates, The New World of Work May 19, 2005

This leads us to GIGO & computers

- Complexity & Irrelevance
  - Introduced through CAI: CATI, CAPI, CASI, etc.
  - Increased by specialists (DDI)
  - Multiplied by time and/or space
  - Thread of irrelevance (GIGO)

Which in turn is a major challenge Relevance

- GIGO
- Knowledge management as the art of forgetting and preservation

Why doesn't proper documentation get done?

- No documentation paradigm
  - No documentation culture
  - No common concept
  - No standard provisions
  - No regular budget
  - No time
  - **No professional training**

  - **No professional user tools**

Essential tool elements

- Human intelligence
  - Descriptions
  - Explanations
  - Data links

- Data streams
  - Response records
  - Digitised audio/video records

Data streams

- Clock, Dialer records, Digitised audio record, Keystroke record, Time stamps
- Data editing records

New tools

- Tools which support human intelligence:
  - SMDS Survey Metadata Documentation System
  - MQDS Michigan Questionnaire Documentation System

- Tools which handle data streaming

Human Intelligence
A breakthrough in survey documentation:
Introducing forms

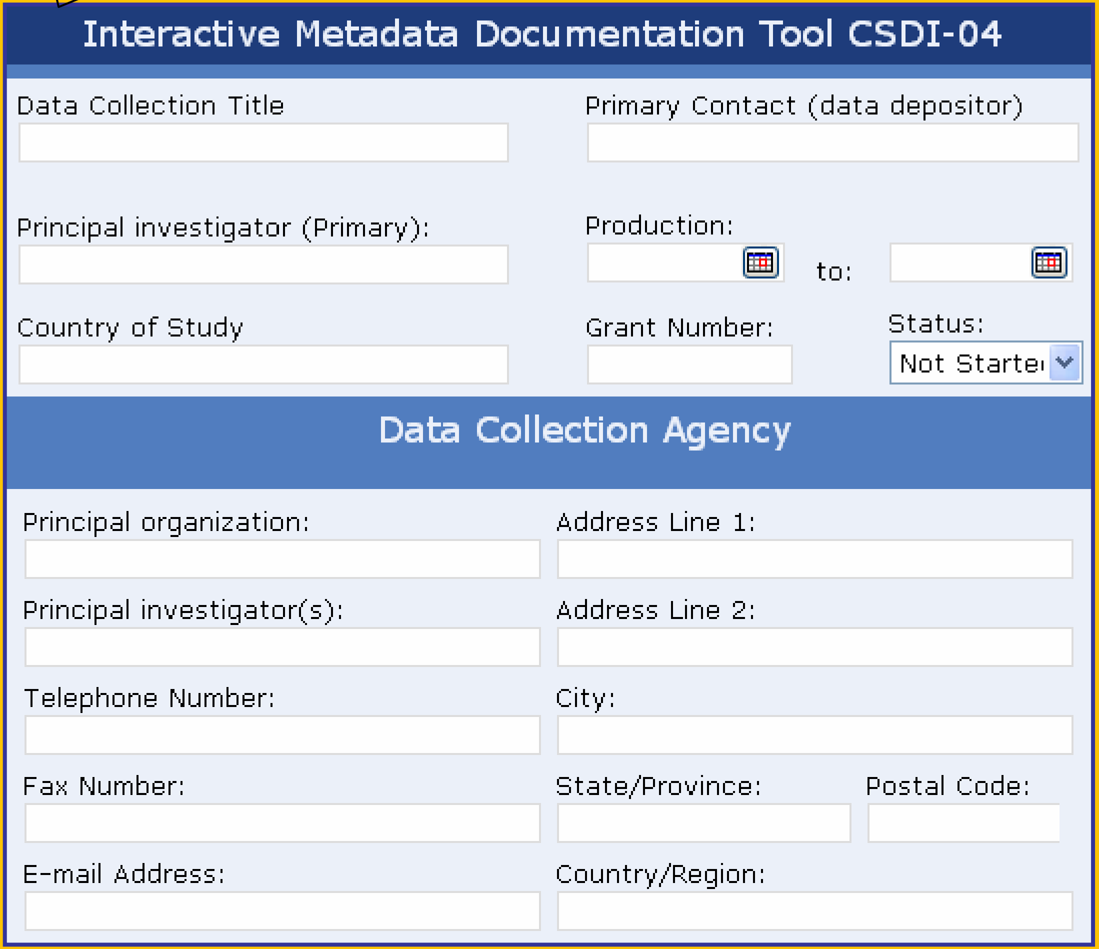
- Asking questions (yourself and others)

- Creating forms (standardised questions)

- *Invented as PAPI for the ISSP by Jowell & Park*
- *Enhanced into web forms by Harkness*
- *Developed into a Survey Information System by ISR/ZUMA*

Introducing forms (documentation surveys)

ISR - Conceptual~Microsoft Office 2003 InfoPath (storing directly in XML)



Interactive Metadata Documentation Tool CSDI-04

Data Collection Title

Primary Contact (data depositor)

Principal investigator (Primary):

Production: to:

Country of Study

Grant Number:

Status: Not Starte

Data Collection Agency

Principal organization:

Address Line 1:

Principal investigator(s):

Address Line 2:

Telephone Number:

City:

Fax Number:

State/Province:

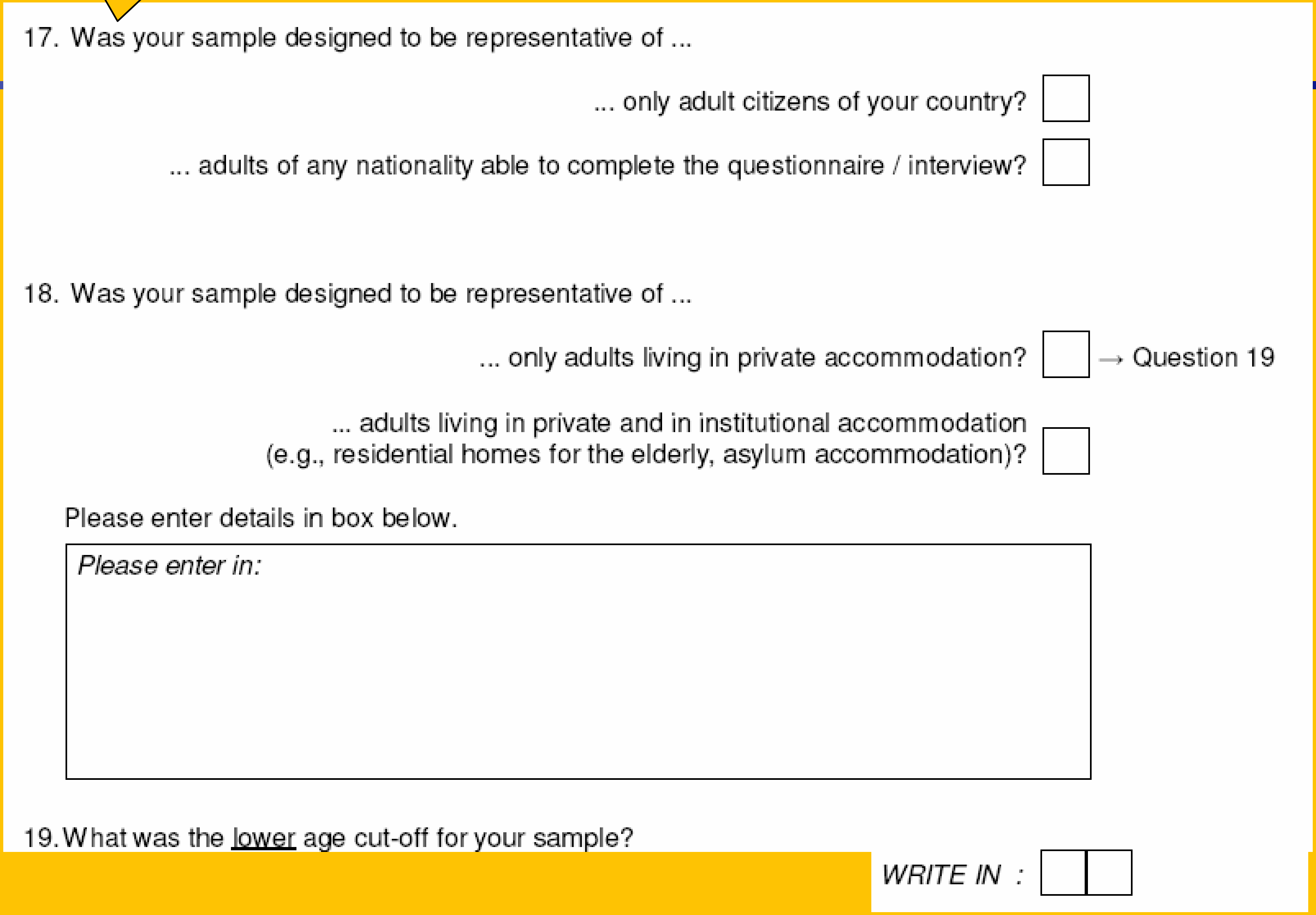
Postal Code:

E-mail Address:

Country/Region:

ISSP
Web & Paper

web forms



17. Was your sample designed to be representative of ...

... only adult citizens of your country? □

... adults of any nationality able to complete the questionnaire / interview? □

18. Was your sample designed to be representative of ...

... only adults living in private accommodation? □ → Question 19

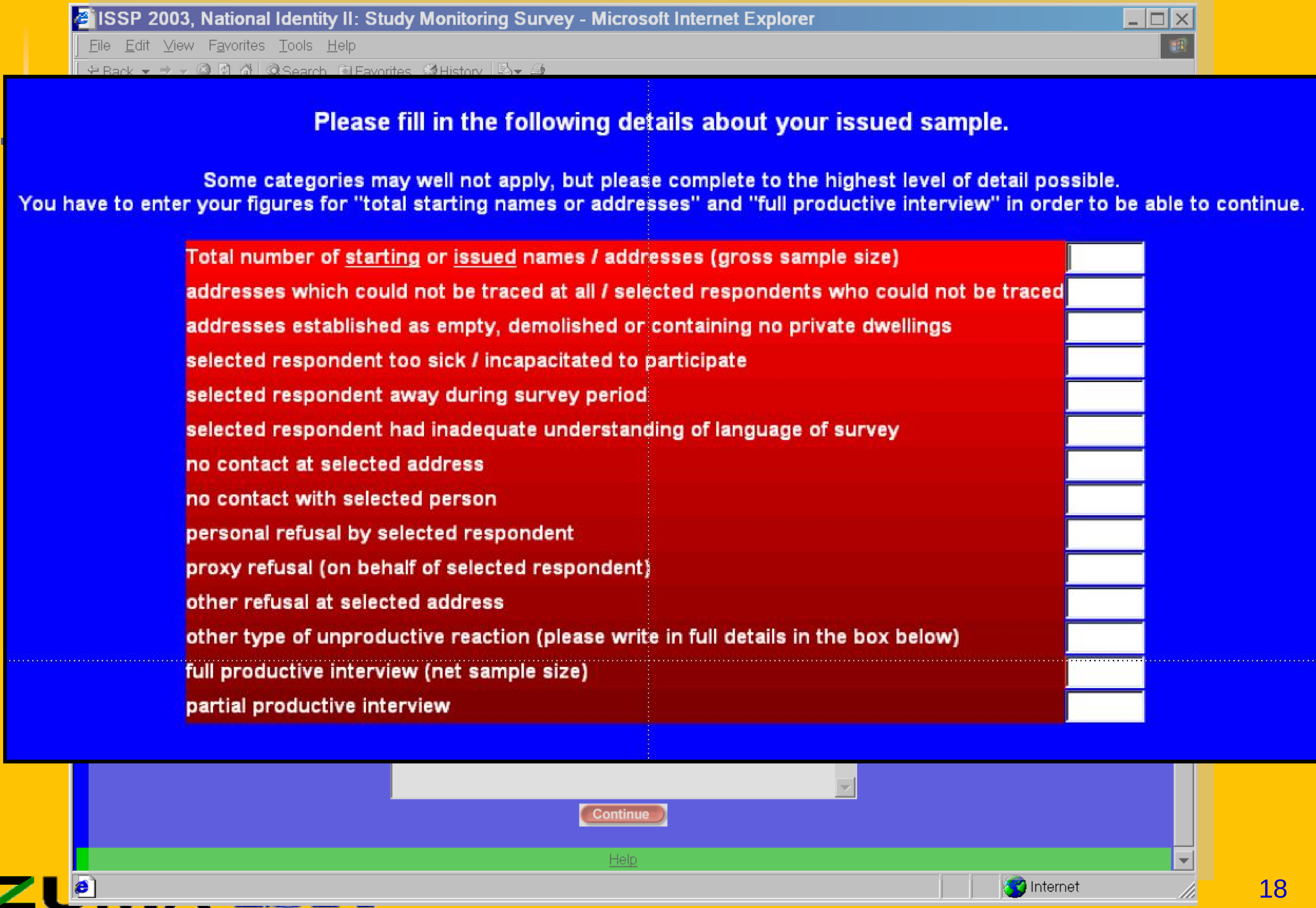
... adults living in private and in institutional accommodation
(e.g., residential homes for the elderly, asylum accommodation)? □

Please enter details in box below.

Please enter in:

19. What was the lower age cut-off for your sample?

WRITE IN : □□


ISSP 2003, National Identity II: Study Monitoring Survey - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Continue
Help
Internet

Please fill in the following details about your issued sample.

Some categories may well not apply, but please complete to the highest level of detail possible.
You have to enter your figures for "total starting names or addresses" and "full productive interview" in order to be able to continue.

Total number of starting or issued names / addresses (gross sample size) □

addresses which could not be traced at all / selected respondents who could not be traced □

addresses established as empty, demolished or containing no private dwellings □

selected respondent too sick / incapacitated to participate □

selected respondent away during survey period □

selected respondent had inadequate understanding of language of survey □

no contact at selected address □

no contact with selected person □

personal refusal by selected respondent □

proxy refusal (on behalf of selected respondent) □

other refusal at selected address □

other type of unproductive reaction (please write in full details in the box below) □

full productive interview (net sample size) □

partial productive interview □

Survey Metadata Documentation System

<SMDS>

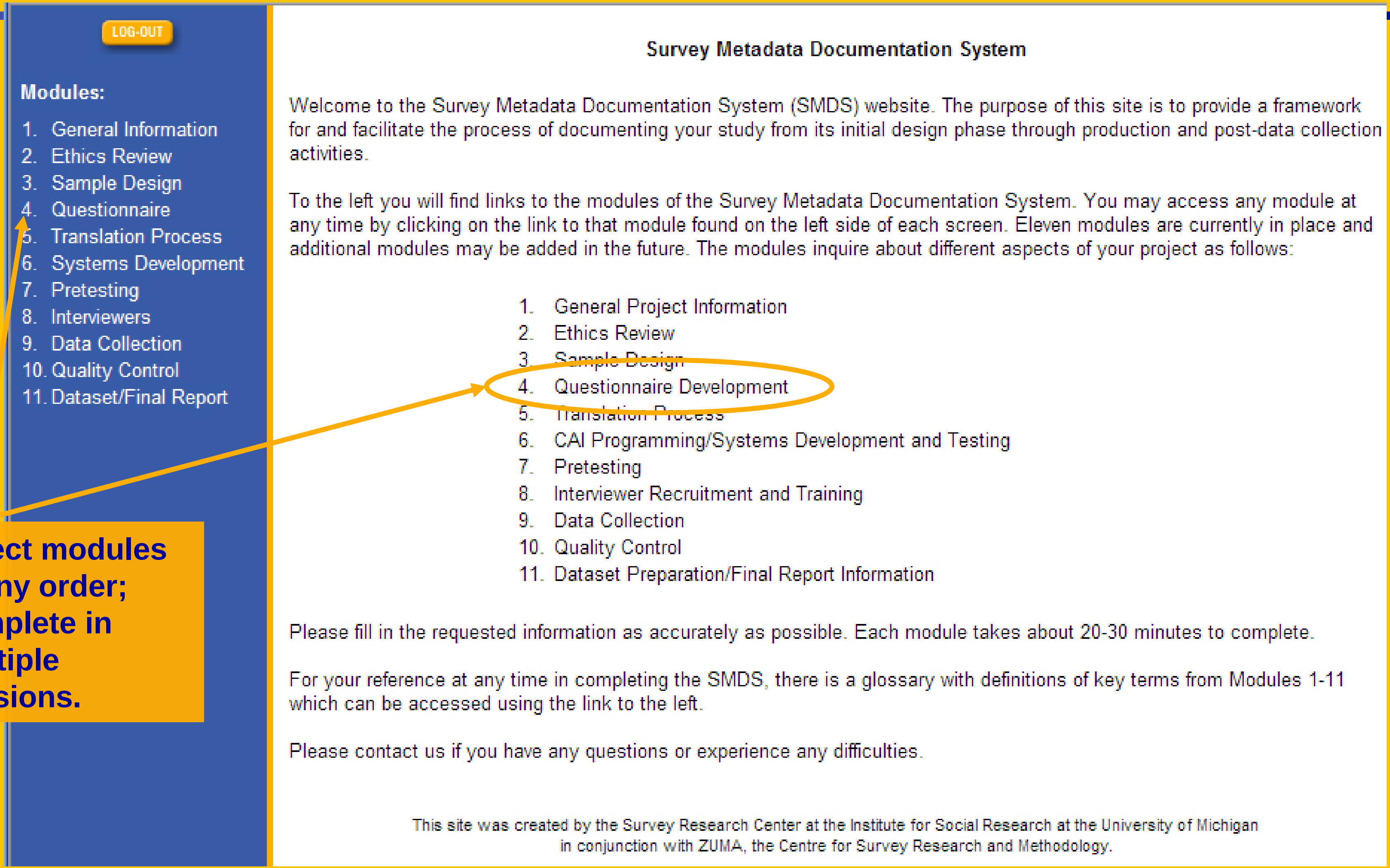
Survey Metadata Documentation System

- ISR (ICPSR) and ZUMA collaborative development
- Tool designed to facilitate documentation of survey lifecycle:
  - from initial design
  - through data collection
  - to post-survey processing and archiving

SMDS-Features

- Supports multiple users simultaneously
- Modularized
- Web-based
- Easy navigation
- Built in skip logic
- Data reporting options by country, module, or question
- Data extraction to third party software package

SMDS Modules



Modules:

1. General Information
2. Ethics Review
3. Sample Design
4. Questionnaire
5. Translation Process
6. Systems Development
7. Pretesting
8. Interviewers
9. Data Collection
10. Quality Control
11. Dataset/Final Report

Survey Metadata Documentation System

Welcome to the Survey Metadata Documentation System (SMDS) website. The purpose of this site is to provide a framework for and facilitate the process of documenting your study from its initial design phase through production and post-data collection activities.

To the left you will find links to the modules of the Survey Metadata Documentation System. You may access any module at any time by clicking on the link to that module found on the left side of each screen. Eleven modules are currently in place and additional modules may be added in the future. The modules inquire about different aspects of your project as follows:

1. General Project Information
2. Ethics Review
3. Sample Design
4. Questionnaire Development
5. Translation Process
6. CAI Programming/Systems Development and Testing
7. Pretesting
8. Interviewer Recruitment and Training
9. Data Collection
10. Quality Control
11. Dataset Preparation/Final Report Information

Please fill in the requested information as accurately as possible. Each module takes about 20-30 minutes to complete.

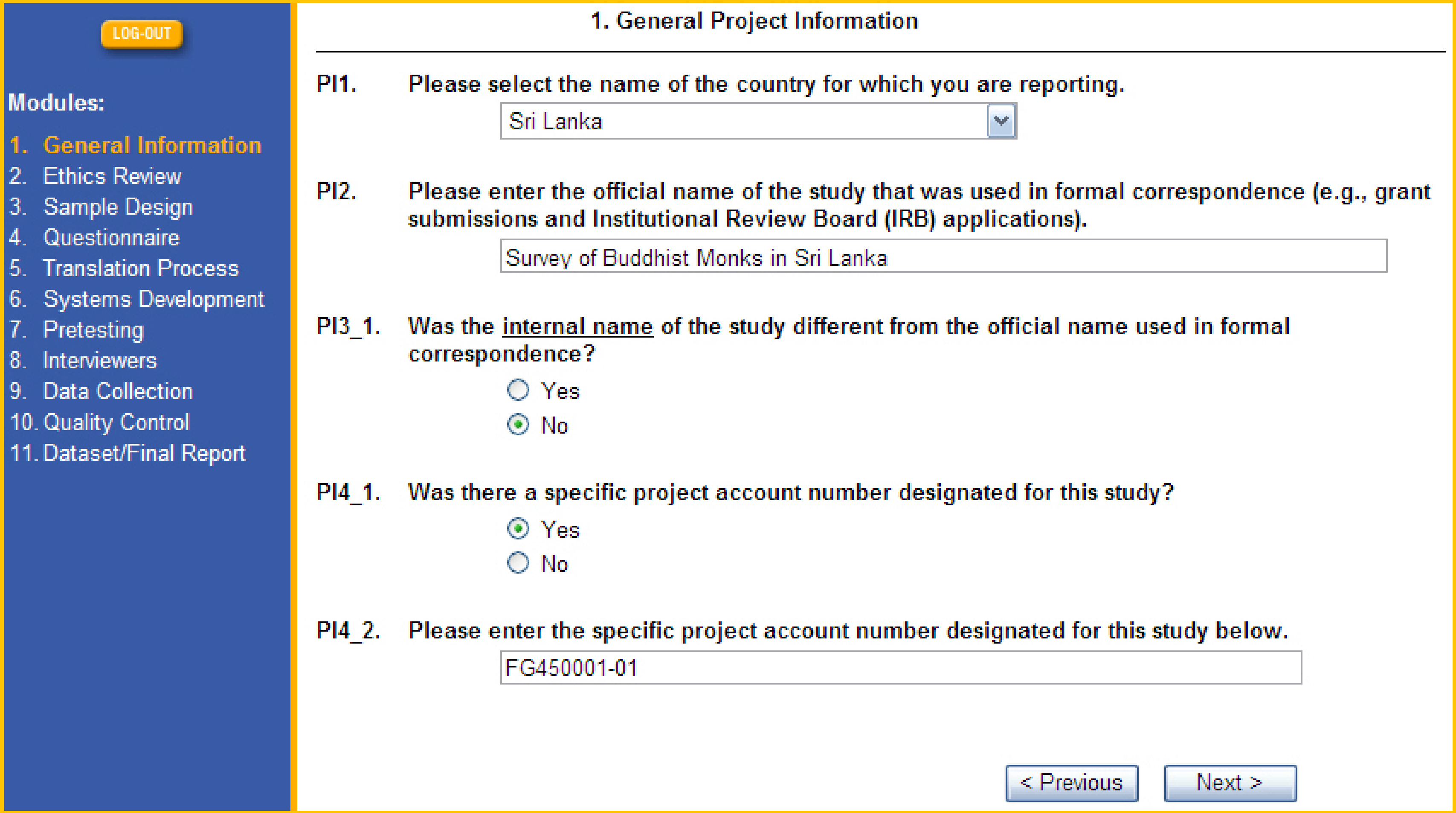
For your reference at any time in completing the SMDS, there is a glossary with definitions of key terms from Modules 1-11 which can be accessed using the link to the left.

Please contact us if you have any questions or experience any difficulties.

This site was created by the Survey Research Center at the Institute for Social Research at the University of Michigan
in conjunction with ZUMA, the Centre for Survey Research and Methodology.

Select modules in any order; complete in multiple sessions.

General Project Information Module



Modules:

1. **General Information**
2. Ethics Review
3. Sample Design
4. Questionnaire
5. Translation Process
6. Systems Development
7. Pretesting
8. Interviewers
9. Data Collection
10. Quality Control
11. Dataset/Final Report

1. General Project Information

PI1. Please select the name of the country for which you are reporting.

Sri Lanka

PI2. Please enter the official name of the study that was used in formal correspondence (e.g., grant submissions and Institutional Review Board (IRB) applications).

Survey of Buddhist Monks in Sri Lanka

PI3_1. Was the internal name of the study different from the official name used in formal correspondence?

○ Yes

● No

PI4_1. Was there a specific project account number designated for this study?

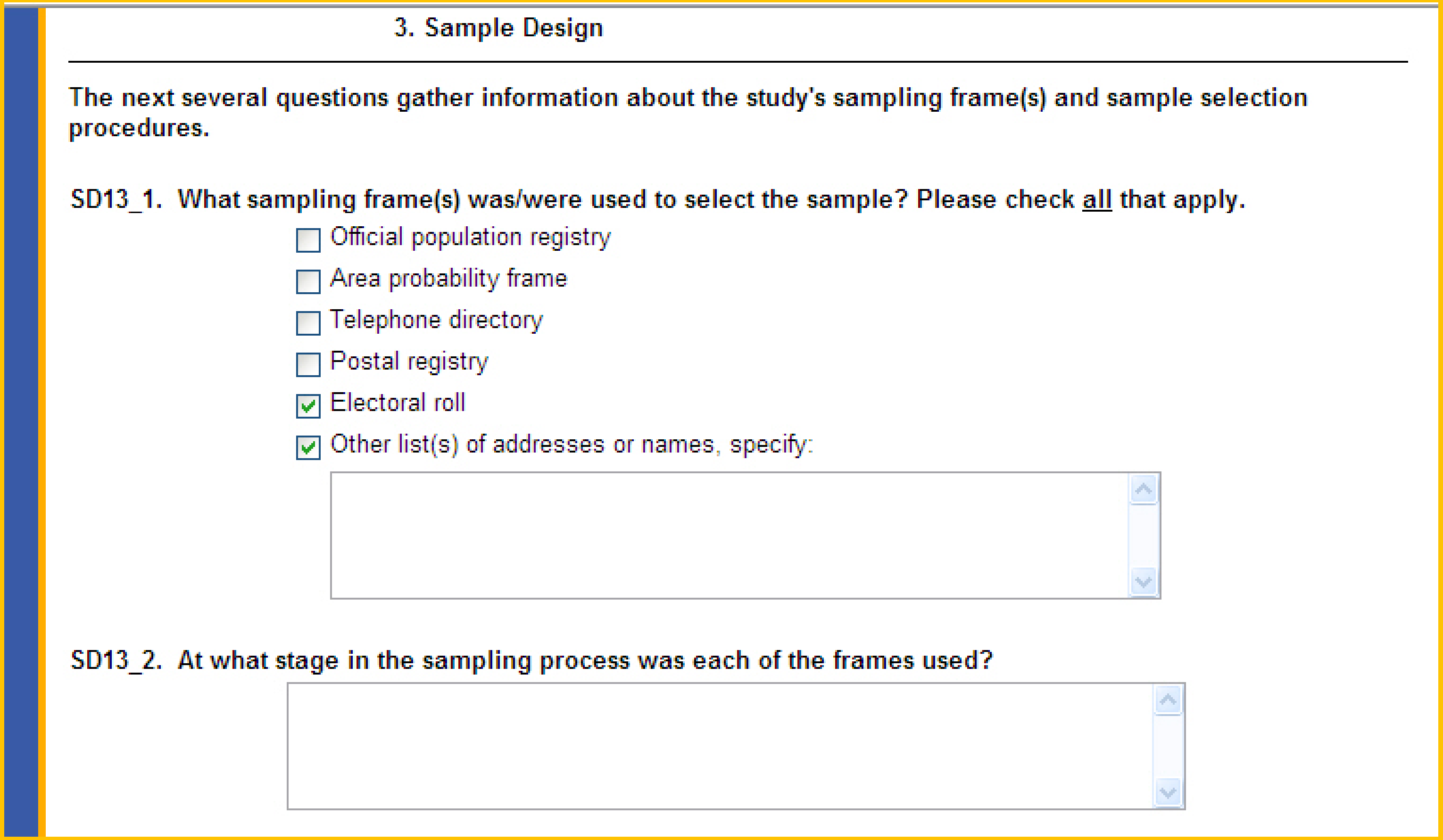
● Yes

○ No

PI4_2. Please enter the specific project account number designated for this study below.

FG450001-01

Sample Design Module



3. Sample Design

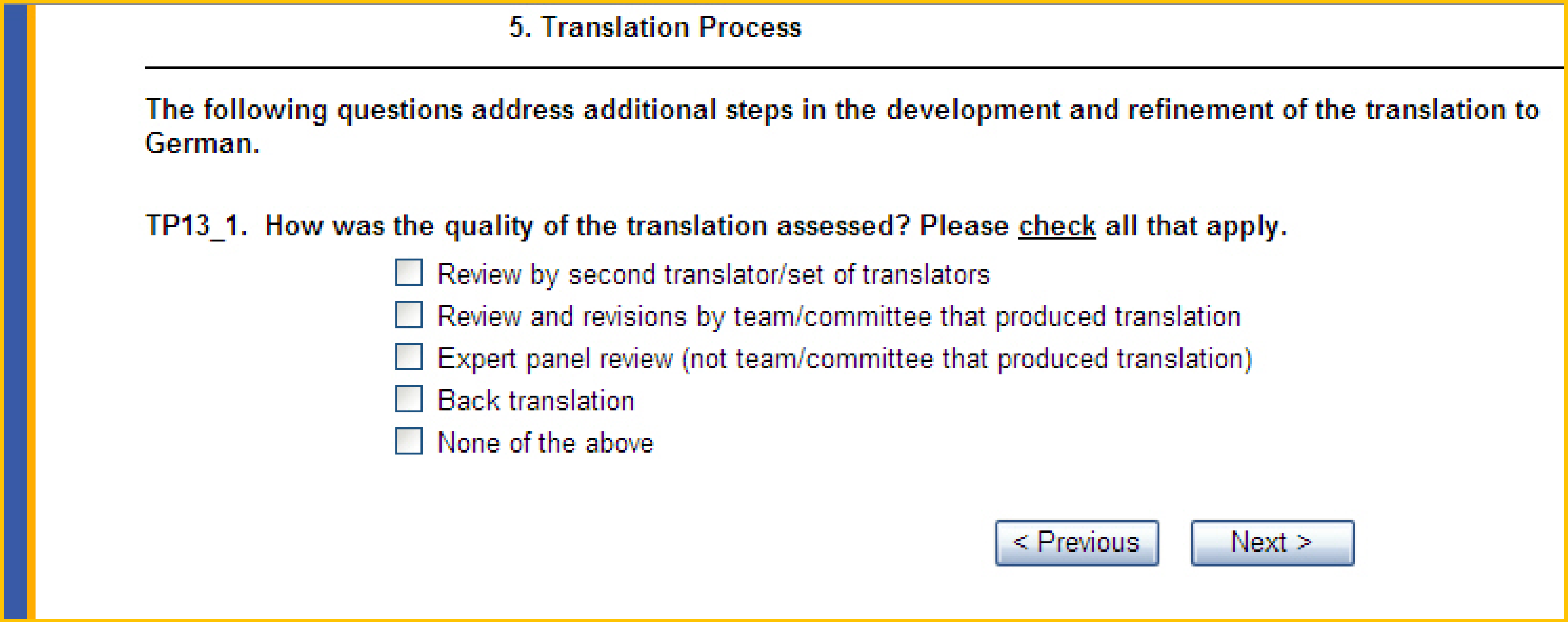
The next several questions gather information about the study's sampling frame(s) and sample selection procedures.

SD13_1. What sampling frame(s) was/were used to select the sample? Please check all that apply.

- □ Official population registry
- □ Area probability frame
- □ Telephone directory
- □ Postal registry
- ☑ Electoral roll
- ☑ Other list(s) of addresses or names, specify:

SD13_2. At what stage in the sampling process was each of the frames used?

Translation Module



5. Translation Process

The following questions address additional steps in the development and refinement of the translation to German.

TP13_1. How was the quality of the translation assessed? Please check all that apply.

□ Review by second translator/set of translators
□ Review and revisions by team/committee that produced translation
□ Expert panel review (not team/committee that produced translation)
□ Back translation
□ None of the above

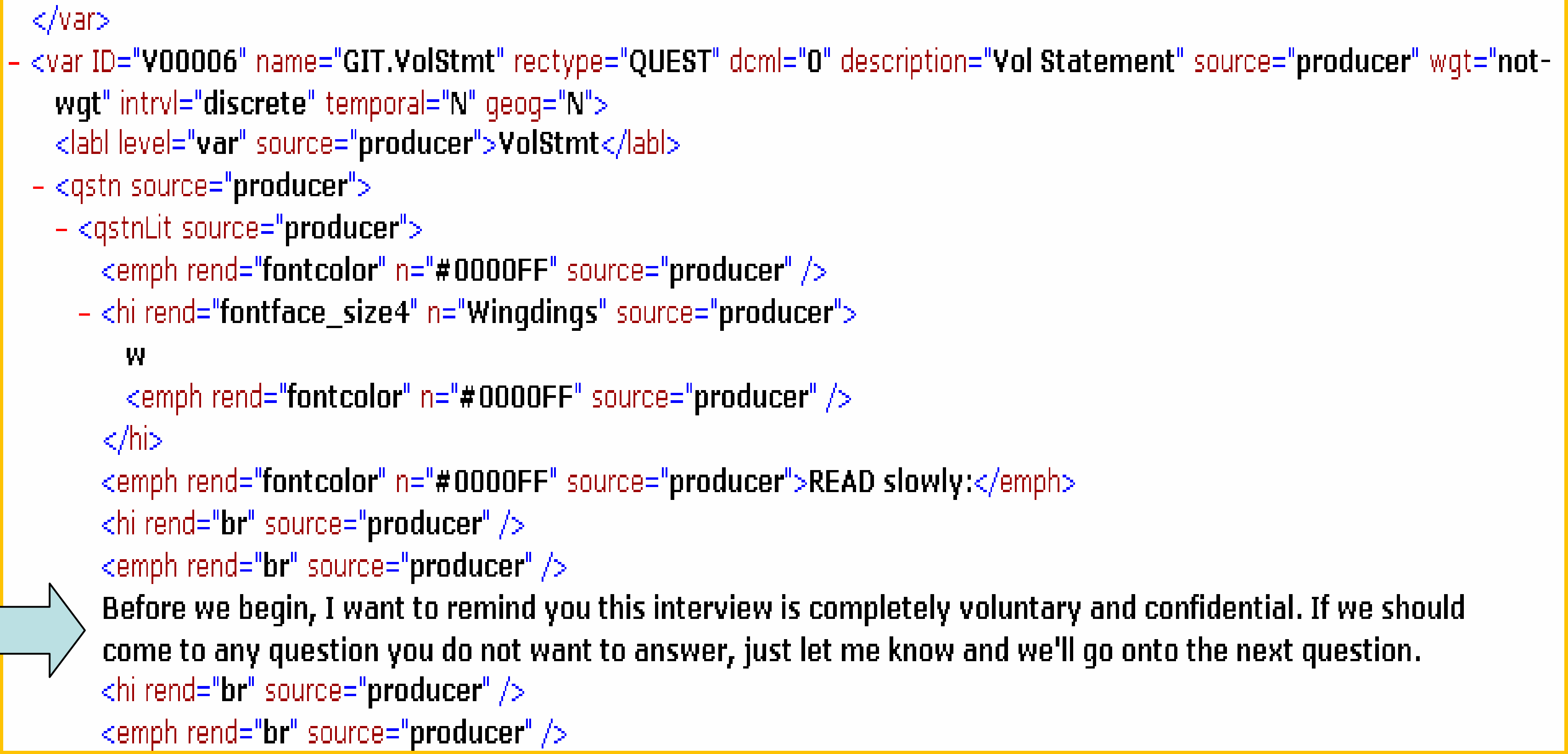
Data Reporting by Module

Module	Question	Question Text	Brazil
3	**SD01**	**Were persons in institutionalized settings (e.g., persons in hospitals, nursing homes, jails, prisons) eligible for this study?**	**No**
…	…	…	…
3	**SD25_7**	**What was the total sample size released to the field for this study, (including all replicates released)?**	**8000**
3	**SD26_1**	**During data collection, was there any subsampling to reduce the number of active cases in the field?**	**Yes**
3	**SD26_2**	**Please describe the cases that were eligible for subsampling.**	**All active cases as of March 1, 2004**
3	**SD26_3**	**What method was used to subsample cases?**	**Random selection**

Michigan Questionnaire Documentation System (MQDS)

- Goals:
  - To facilitate:
    - Testing
    - Human subjects/ethics review
    - Version control/translation documentation
      - Comparison of instruments used in data collection
      - Comparison of data collection instrument against newest version
    - Codebook generation/archiving
    - Public release data files (with appropriate links)

BLAISE or ...

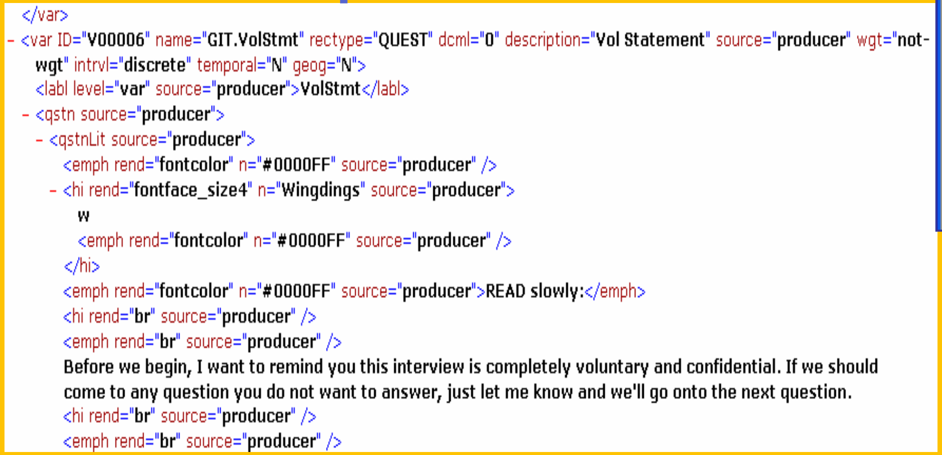


```
  </var>
- <var ID="V00006" name="GIT.VolStmt" rectype="QUEST" dcml="0" description="Vol Statement" source="producer" wgt="not-
    wgt" intrvl="discrete" temporal="N" geog="N">
    <labl level="var" source="producer">VolStmt</labl>
  - <qstn source="producer">
    - <qstnLit source="producer">
        <emph rend="fontcolor" n="#0000FF" source="producer" />
      - <hi rend="fontface_size4" n="Wingdings" source="producer">
          w
          <emph rend="fontcolor" n="#0000FF" source="producer" />
        </hi>
        <emph rend="fontcolor" n="#0000FF" source="producer">READ slowly:</emph>
        <hi rend="br" source="producer" />
        <emph rend="br" source="producer" />
        Before we begin, I want to remind you this interview is completely voluntary and confidential. If we should
        come to any question you do not want to answer, just let me know and we'll go onto the next question.
        <hi rend="br" source="producer" />
        <emph rend="br" source="producer" />
```

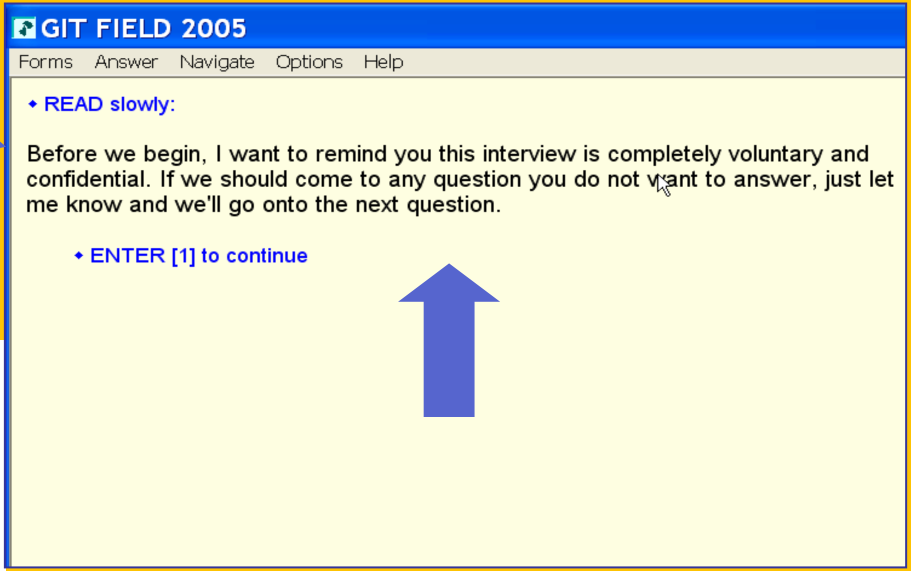
Adapted from Gina-Qian Cheung, Sue Ellen Hansen, Beth-Ellen Pennell, Peter Sparks, Megan Turf, Karl Dinkelmann ©ISR 2005

…not to BLAISE …

MQDS captures Blaise screen formatting characteristics, e.g., bold, blue, special characters, etc.



```
 </var>
- <var ID="V00006" name="GIT.VolStmt" rectype="QUEST" dcml="0" description="Vol Statement" source="producer" wgt="not-
   wgt" intrvl="discrete" temporal="N" geog="N">
   <labl level="var" source="producer">VolStmt</labl>
 - <qstn source="producer">
   - <qstnLit source="producer">
       <emph rend="fontcolor" n="#0000FF" source="producer" />
     - <hi rend="fontface_size4" n="Wingdings" source="producer">
         w
         <emph rend="fontcolor" n="#0000FF" source="producer" />
       </hi>
       <emph rend="fontcolor" n="#0000FF" source="producer">READ slowly:</emph>
       <hi rend="br" source="producer" />
       <emph rend="br" source="producer" />
       Before we begin, I want to remind you this interview is completely voluntary and confidential. If we should
       come to any question you do not want to answer, just let me know and we'll go onto the next question.
       <hi rend="br" source="producer" />
       <emph rend="br" source="producer" />
```



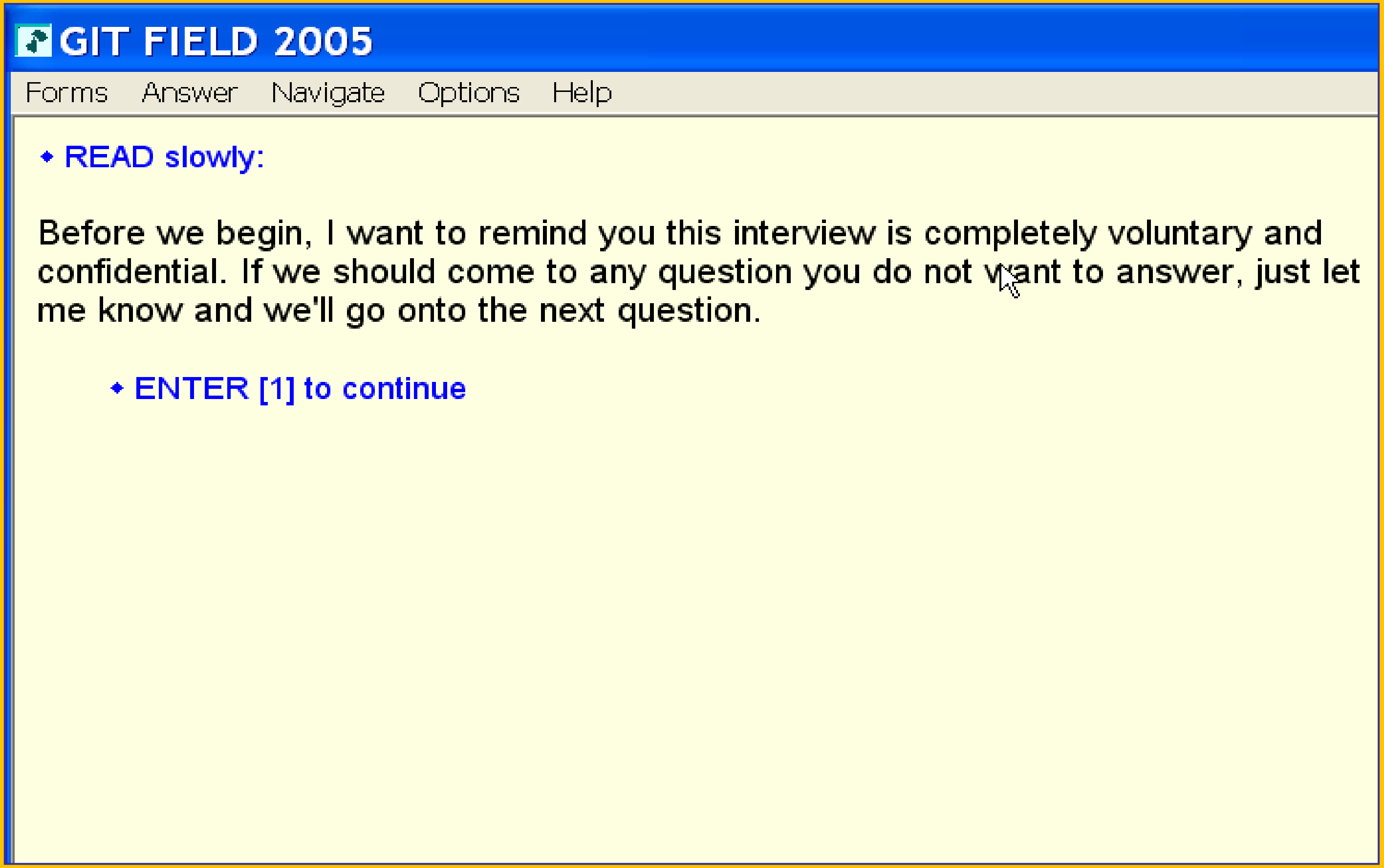
◆ READ slowly:

Before we begin, I want to remind you this interview is completely voluntary and confidential. If we should come to any question you do not want to answer, just let me know and we'll go onto the next question.

◆ ENTER [1] to continue

Adapted from Gina-Qian Cheung, Sue Ellen Hansen, Beth-Ellen Pennell, Peter Sparks, Megan Turf, Karl Dinkelmann ©ISR 2005

…but working with real text



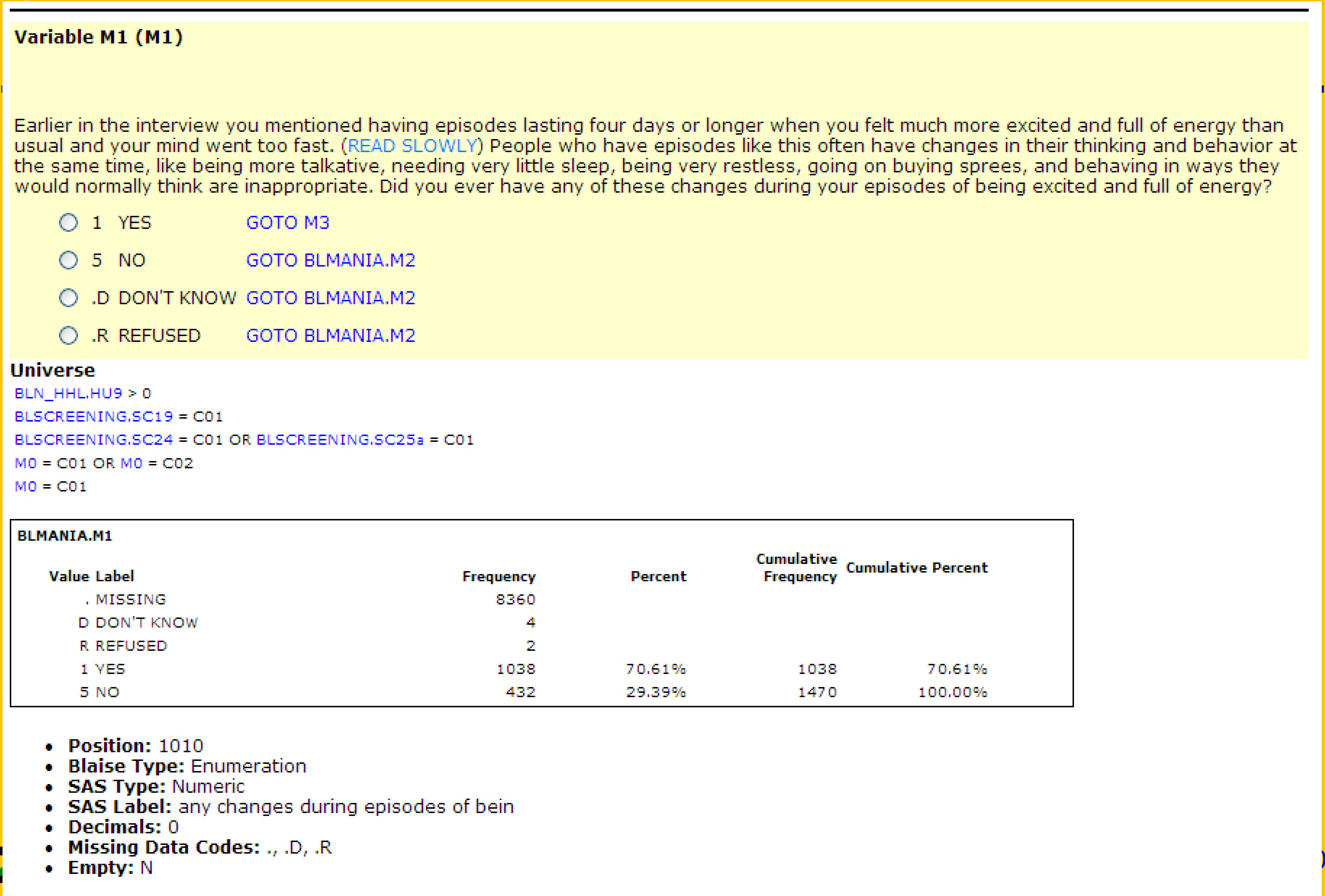
- READ slowly:

Before we begin, I want to remind you this interview is completely voluntary and confidential. If we should come to any question you do not want to answer, just let me know and we'll go onto the next question.

- ENTER [1] to continue

Adapted from Gina-Qian Cheung, Sue Ellen Hansen, Beth-Ellen Pennell, Peter Sparks, Megan Turf, Karl Dinkelmann ©ISR 2005

Codebook From MQDS



Variable M1 (M1)

Earlier in the interview you mentioned having episodes lasting four days or longer when you felt much more excited and full of energy than usual and your mind went too fast. (READ SLOWLY) People who have episodes like this often have changes in their thinking and behavior at the same time, like being more talkative, needing very little sleep, being very restless, going on buying sprees, and behaving in ways they would normally think are inappropriate. Did you ever have any of these changes during your episodes of being excited and full of energy?

○	1	YES	GOTO M3
○	5	NO	GOTO BLMANIA.M2
○	.D	DON'T KNOW	GOTO BLMANIA.M2
○	.R	REFUSED	GOTO BLMANIA.M2

Universe

BLN_HHL.HU9 > 0

BLSCREENING.SC19 = C01

BLSCREENING.SC24 = C01 OR BLSCREENING.SC25a = C01

M0 = C01 OR M0 = C02

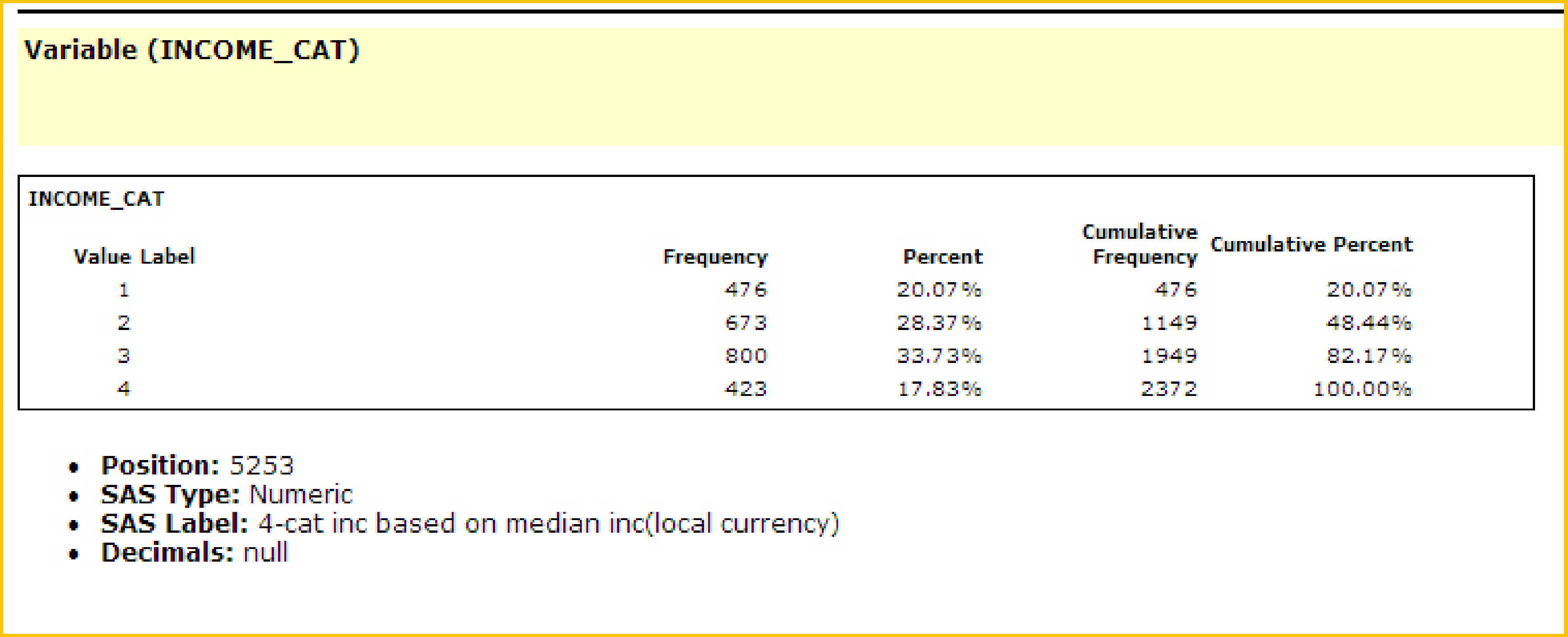
M0 = C01

BLMANIA.M1

Value Label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
. MISSING	8360			
D DON'T KNOW	4			
R REFUSED	2			
1 YES	1038	70.61%	1038	70.61%
5 NO	432	29.39%	1470	100.00%

- **Position:** 1010
- **Blaise Type:** Enumeration
- **SAS Type:** Numeric
- **SAS Label:** any changes during episodes of bein
- **Decimals:** 0
- **Missing Data Codes:** ., .D, .R
- **Empty:** N

CAPI Created Variable Output



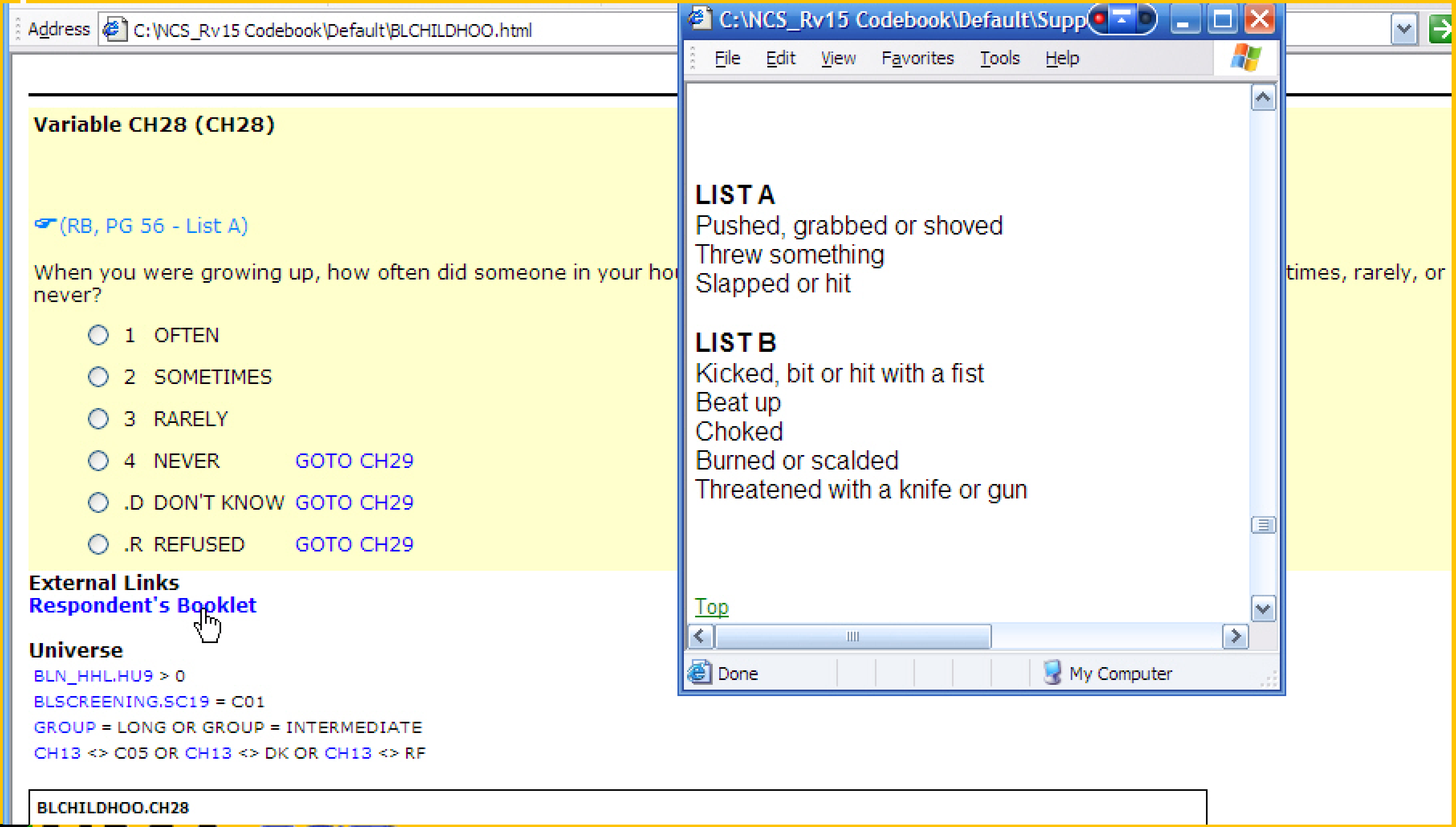
Variable (INCOME_CAT)

INCOME_CAT

Value Label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	476	20.07%	476	20.07%
2	673	28.37%	1149	48.44%
3	800	33.73%	1949	82.17%
4	423	17.83%	2372	100.00%

- **Position:** 5253
- **SAS Type:** Numeric
- **SAS Label:** 4-cat inc based on median inc(local currency)
- **Decimals:** null

Features: Link to Respondent Booklet


Address
C:\NCS_Rv15 Codebook\Default\BLCHILDHOO.html
C:\NCS_Rv15 Codebook\Default\Supp
File
Edit
View
Favorites
Tools
Help
Done
My Computer

Variable CH28 (CH28)

☞(RB, PG 56 - List A)

When you were growing up, how often did someone in your ho… times, rarely, or never?

○ 1 OFTEN

○ 2 SOMETIMES

○ 3 RARELY

○ 4 NEVER GOTO CH29

○ .D DON'T KNOW GOTO CH29

○ .R REFUSED GOTO CH29

External Links

Respondent's Booklet

Universe

BLN_HHL.HU9 > 0

BLSCREENING.SC19 = C01

GROUP = LONG OR GROUP = INTERMEDIATE

CH13 <> C05 OR CH13 <> DK OR CH13 <> RF

BLCHILDHOO.CH28

LIST A

Pushed, grabbed or shoved

Threw something

Slapped or hit

LIST B

Kicked, bit or hit with a fist

Beat up

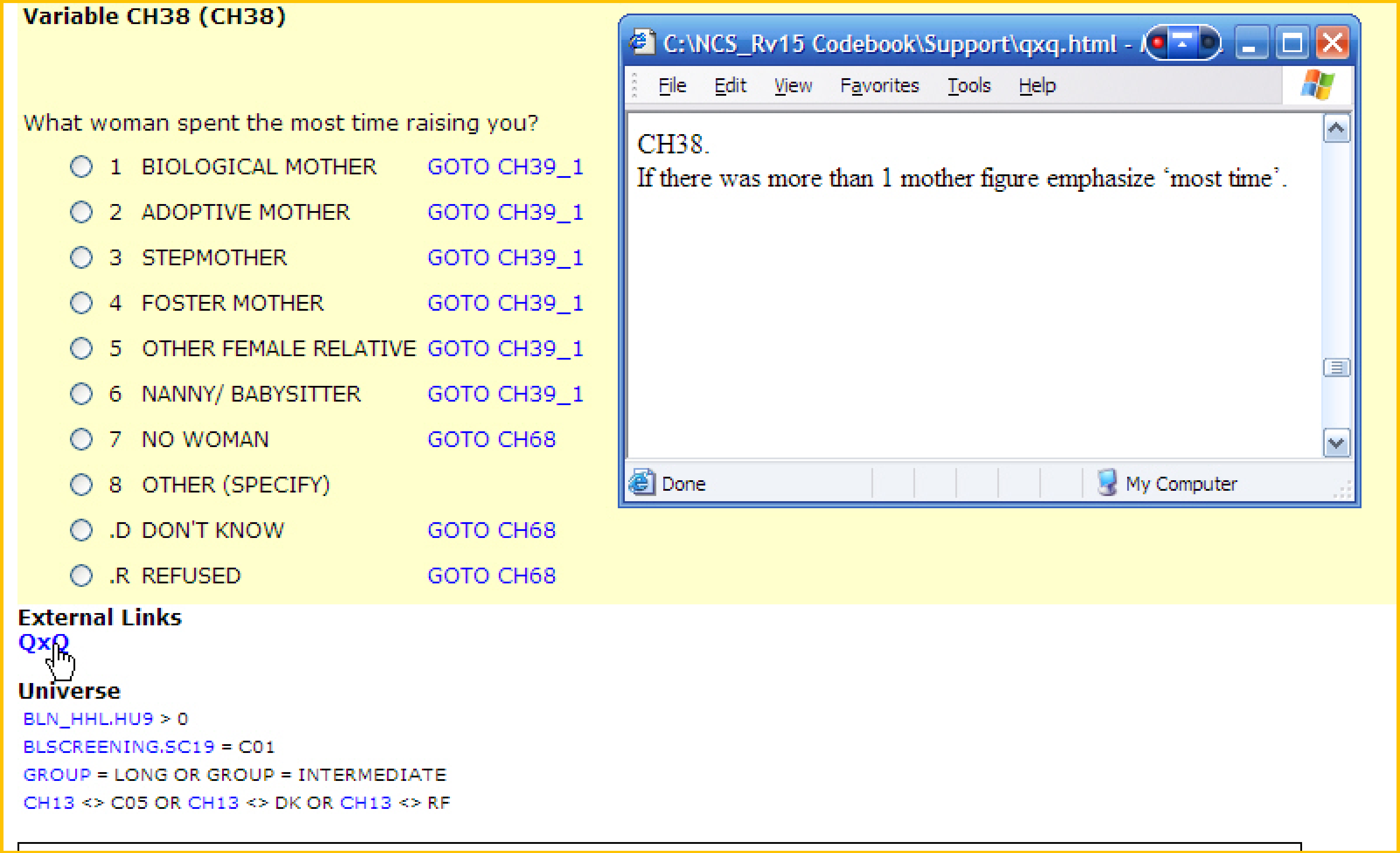
Choked

Burned or scalded

Threatened with a knife or gun

Top

Features: Link to Online Q by Q's



Variable CH38 (CH38)

What woman spent the most time raising you?

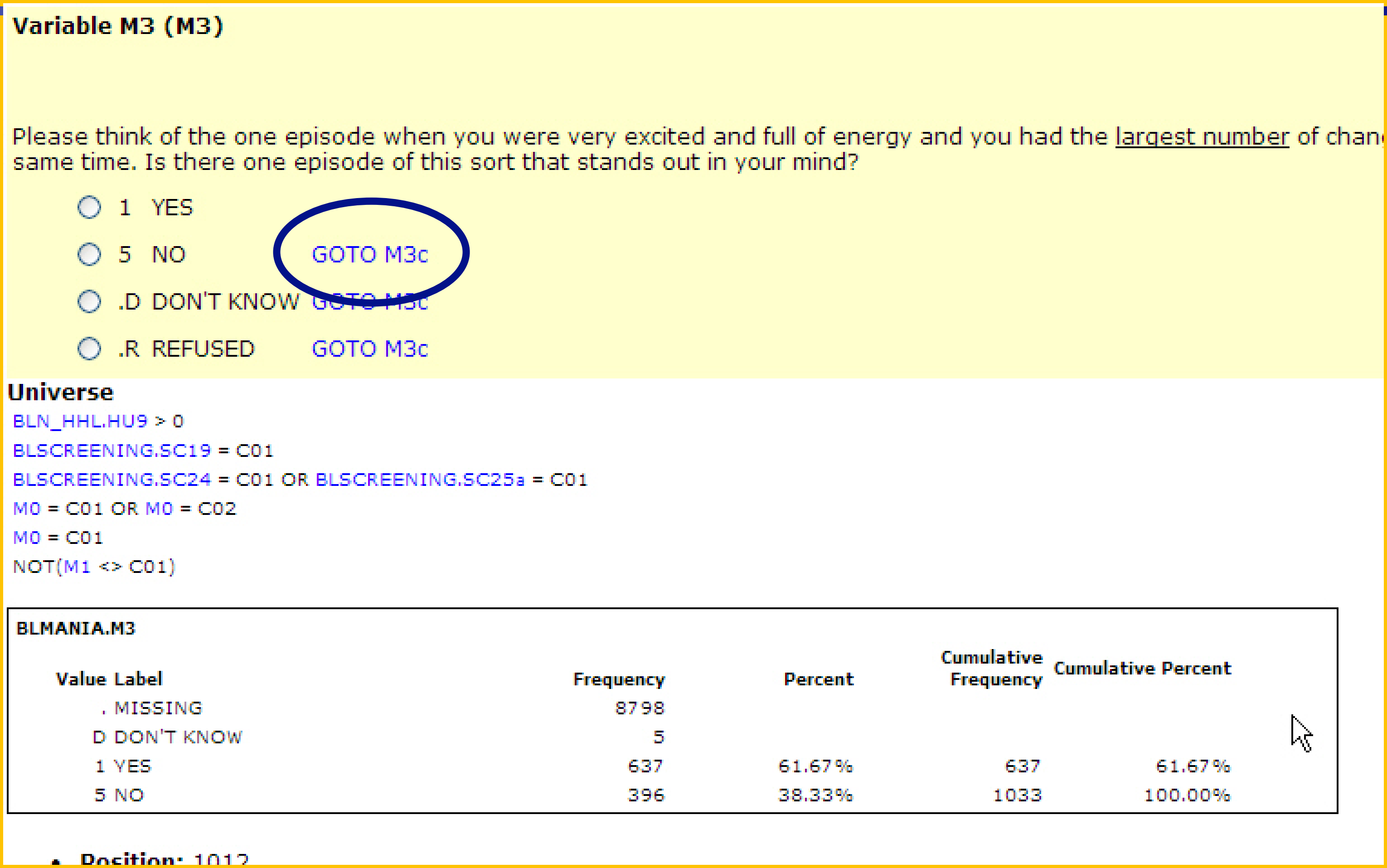
○	1	BIOLOGICAL MOTHER	GOTO CH39_1
○	2	ADOPTIVE MOTHER	GOTO CH39_1
○	3	STEPMOTHER	GOTO CH39_1
○	4	FOSTER MOTHER	GOTO CH39_1
○	5	OTHER FEMALE RELATIVE	GOTO CH39_1
○	6	NANNY/ BABYSITTER	GOTO CH39_1
○	7	NO WOMAN	GOTO CH68
○	8	OTHER (SPECIFY)	
○	.D	DON'T KNOW	GOTO CH68
○	.R	REFUSED	GOTO CH68

CH38.
If there was more than 1 mother figure emphasize ‘most time’.

External Links
QxQ

Universe
BLN_HHL.HU9 > 0
BLSCREENING.SC19 = C01
GROUP = LONG OR GROUP = INTERMEDIATE
CH13 <> C05 OR CH13 <> DK OR CH13 <> RF

Features: Interactive Gotos



Variable M3 (M3)

Please think of the one episode when you were very excited and full of energy and you had the largest number of chan
same time. Is there one episode of this sort that stands out in your mind?

○ 1 YES

○ 5 NO GOTO M3c

○ .D DON'T KNOW GOTO M3c

○ .R REFUSED GOTO M3c

Universe

BLN_HHL.HU9 > 0

BLSCREENING.SC19 = C01

BLSCREENING.SC24 = C01 OR BLSCREENING.SC25a = C01

M0 = C01 OR M0 = C02

M0 = C01

NOT(M1 <> C01)

BLMANIA.M3

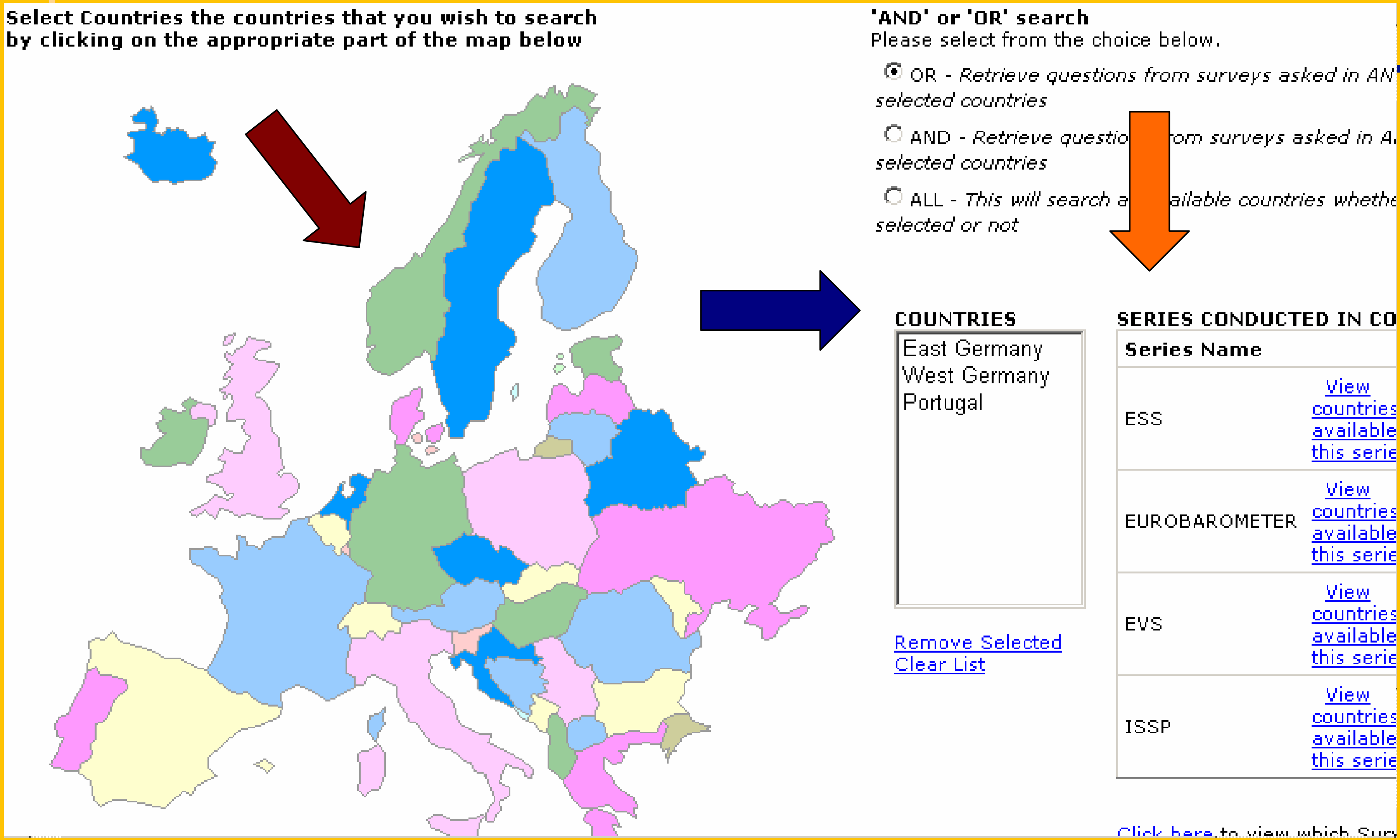
Value Label	Frequency	Percent	Cumulative Frequency	Cumulative Percent
. MISSING	8798			
D DON'T KNOW	5			
1 YES	637	61.67%	637	61.67%
5 NO	396	38.33%	1033	100.00%

- **Position:** 1012

ZUMA ISR

CCESDIS

Don‘t forget analyses!



Select Countries the countries that you wish to search by clicking on the appropriate part of the map below

'AND' or 'OR' search
Please select from the choice below.

● OR - *Retrieve questions from surveys asked in AN selected countries*

○ AND - *Retrieve questio om surveys asked in A selected countries*

○ ALL - *This will search a ailable countries whethe selected or not*

COUNTRIES

East Germany
West Germany
Portugal

Remove Selected
Clear List

SERIES CONDUCTED IN CO

Series Name	
ESS	View countries available this serie
EUROBAROMETER	View countries available this serie
EVS	View countries available this serie
ISSP	View countries available this serie

Click here to view which Surv

Data Stream Paradata

- Definition: data about the data collection process, e.g., call records, cost data, audit trail data, interviewer characteristics, verification/validation data
- Goals:
  - Ongoing process and quality monitoring
  - Inform responsive design decisions
  - Cost modeling
  - Methodological studies
- Examples of Paradata Uses...

Cost Modeling

- Model time to complete interviewing related activities (interviewing, various contact outcomes, traveling, etc.)
- Model interviewer pay rate as a function of sample location, language needs, interviewer experience mix
- Model hours per interview over time/study phase

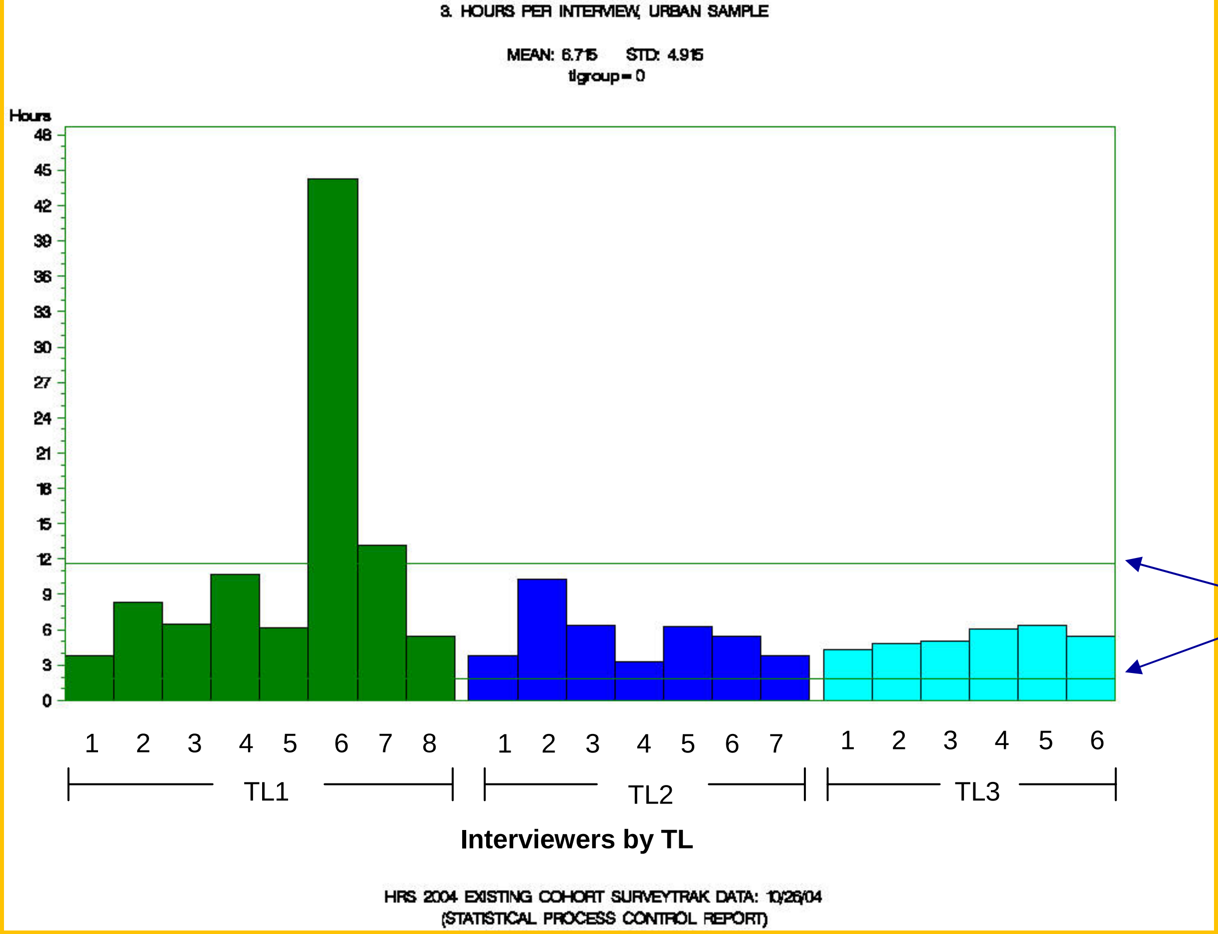
Process and Quality Monitoring

- Examples:
  - Statistical process control charts
  - Identify potential interview falsification trends
    - Length of interviews
    - Number of phone numbers collected

Process and Quality Monitoring

- Focus interviewers' efforts:
  - Tracking age of lines
  - Setting call limitations -- reduce survey costs while being informed about any potential survey bias. Two key criteria:
    - Determining at which point additional calls to a line are inefficient
    - Determining whether respondents cooperating after a certain number of calls are significantly different from others on key indicators

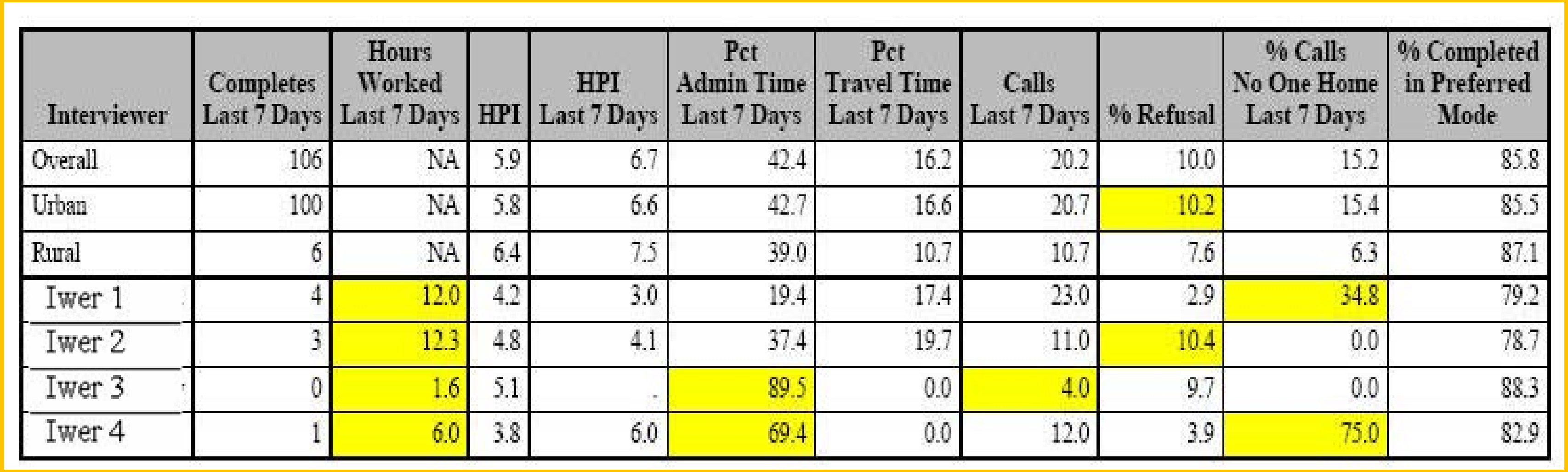
Statistical Control Chart: Hours by Interviewer (HRS)



Control Limits

ZUMA ISR

Interviewer Production Report: Outliers Highlighted (HRS)

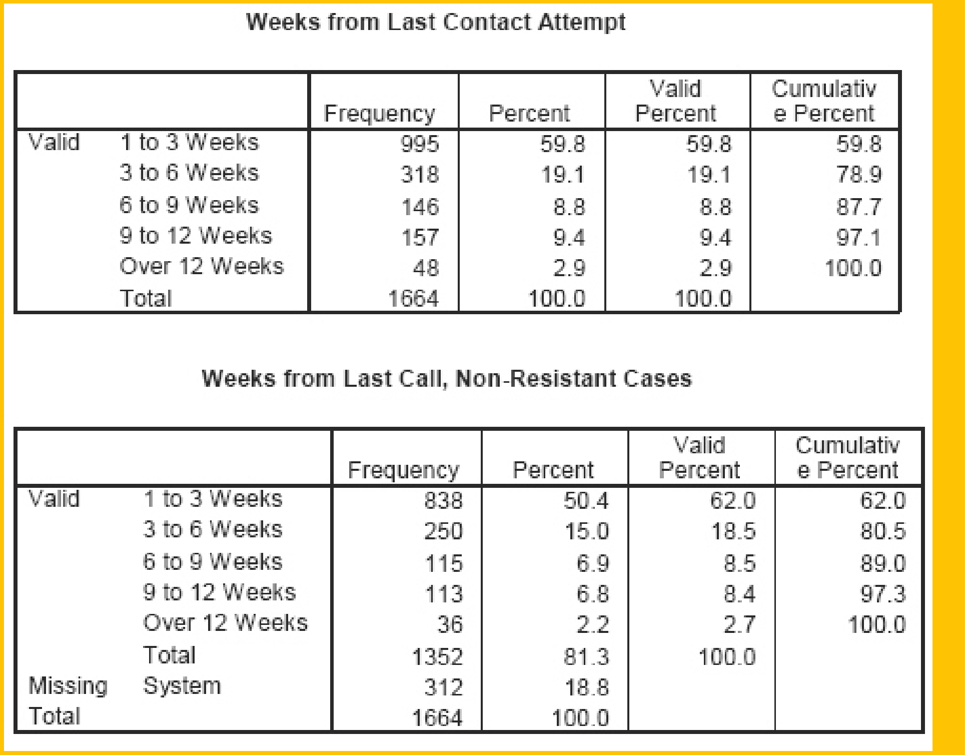


Summary of HRS Interviewer Production for TL1

Interviewer	Completes Last 7 Days	Hours Worked Last 7 Days	HPI	HPI Last 7 Days	Pct Admin Time Last 7 Days	Pct Travel Time Last 7 Days	Calls Last 7 Days	% Refusal	% Calls No One Home Last 7 Days	% Completed in Preferred Mode
Overall	106	NA	5.9	6.7	42.4	16.2	20.2	10.0	15.2	85.8
Urban	100	NA	5.8	6.6	42.7	16.6	20.7	10.2	15.4	85.5
Rural	6	NA	6.4	7.5	39.0	10.7	10.7	7.6	6.3	87.1
Iwer 1	4	12.0	4.2	3.0	19.4	17.4	23.0	2.9	34.8	79.2
Iwer 2	3	12.3	4.8	4.1	37.4	19.7	11.0	10.4	0.0	78.7
Iwer 3	0	1.6	5.1	.	89.5	0.0	4.0	9.7	0.0	88.3
Iwer 4	1	6.0	3.8	6.0	69.4	0.0	12.0	3.9	75.0	82.9

Source: G. Benson & J. Wagner 2003

Aging Lines (NSFG)

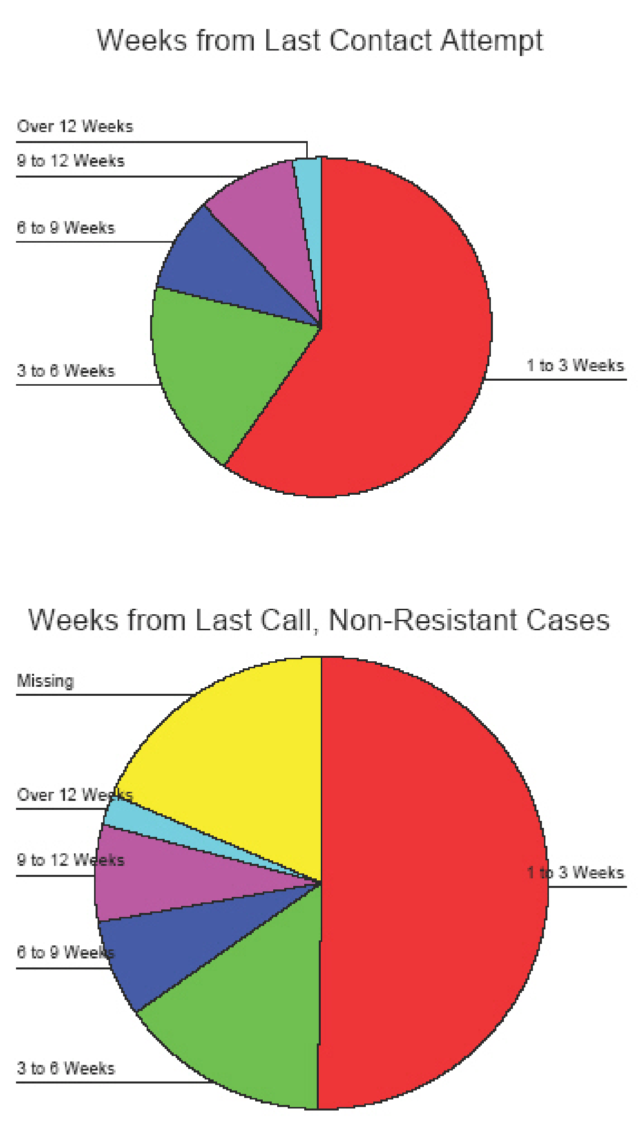


Weeks from Last Contact Attempt

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 to 3 Weeks	995	59.8	59.8	59.8
	3 to 6 Weeks	318	19.1	19.1	78.9
	6 to 9 Weeks	146	8.8	8.8	87.7
	9 to 12 Weeks	157	9.4	9.4	97.1
	Over 12 Weeks	48	2.9	2.9	100.0
	Total	1664	100.0	100.0	

Weeks from Last Call, Non-Resistant Cases

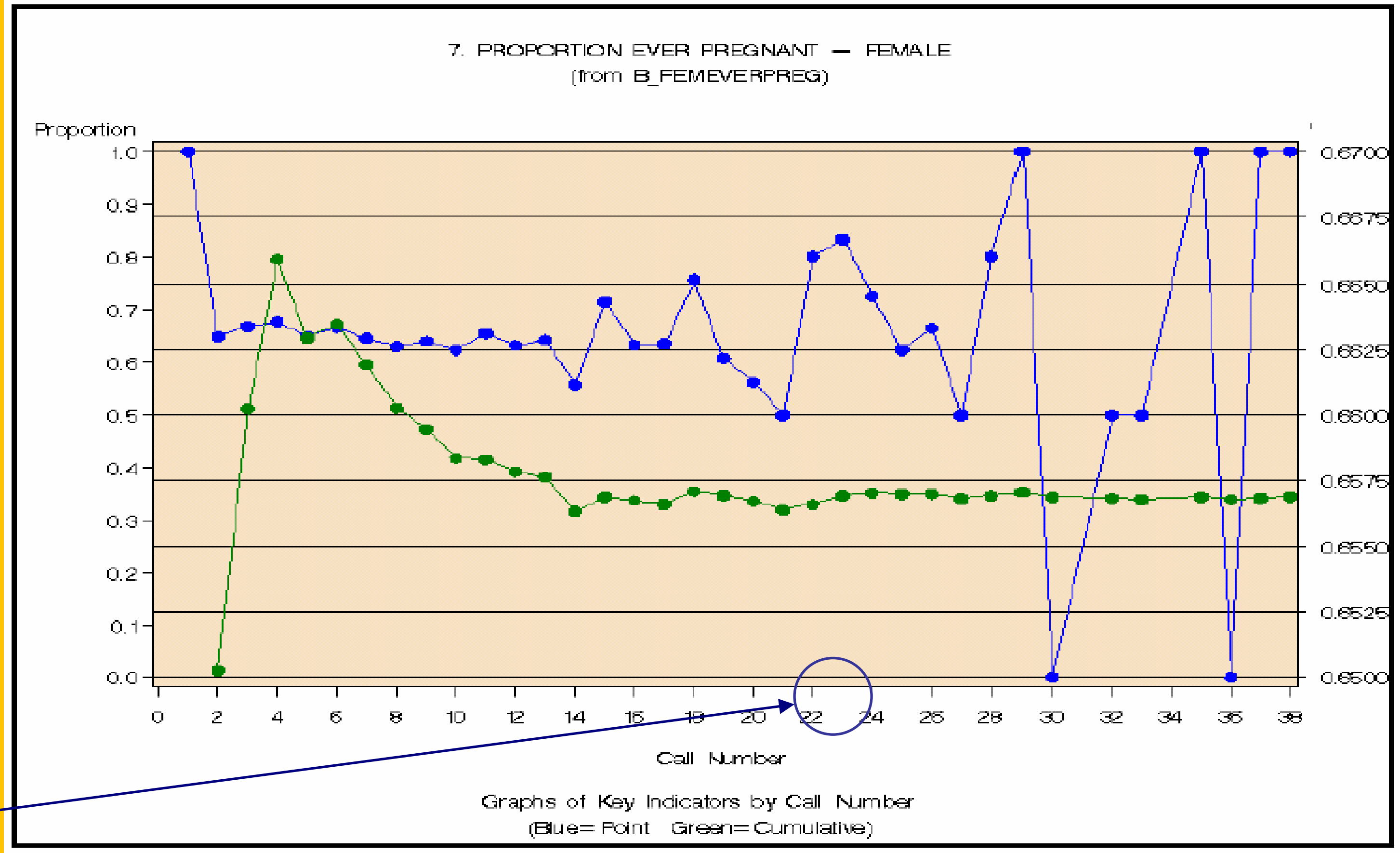
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 to 3 Weeks	838	50.4	62.0	62.0
	3 to 6 Weeks	250	15.0	18.5	80.5
	6 to 9 Weeks	115	6.9	8.5	89.0
	9 to 12 Weeks	113	6.8	8.4	97.3
	Over 12 Weeks	36	2.2	2.7	100.0
	Total	1352	81.3	100.0	
Missing	System	312	18.8		
Total		1664	100.0		



Propensity Model Predicting Likelihood of Next Call Yielding Interview (NSFG)

- Hazard model on call-level data
  - Screener interview model
  - Main interview, given screener, model
- Obtained expected value for each sample case, given model
  - Summed across cases in a segment
  - Used to guide interviewer resource placement
  - Used to stratify segments for double sample selection

Proportion Ever Pregnant by Number Calls (NSFG)



Proportion Stabilizes

Methodological Studies: ADKs

- Instrument design and usability problems
- High incidents of certain behaviors
  - Invoking help & interviewer comments
  - Suppressing edit checks
  - Backing up and reviewing / changing answers

Audit Trail and Keystroke (ADK) Data

- Computer assisted interviewing (CAI) audit trails of survey items visited, with associated keystrokes, user actions, and final values
- Reporting system allows review of specific items and instrument sections, by item/section, interview, or interviewer
- Used to evaluate questionnaire design, interviewer performance, or specific interview problems or issues

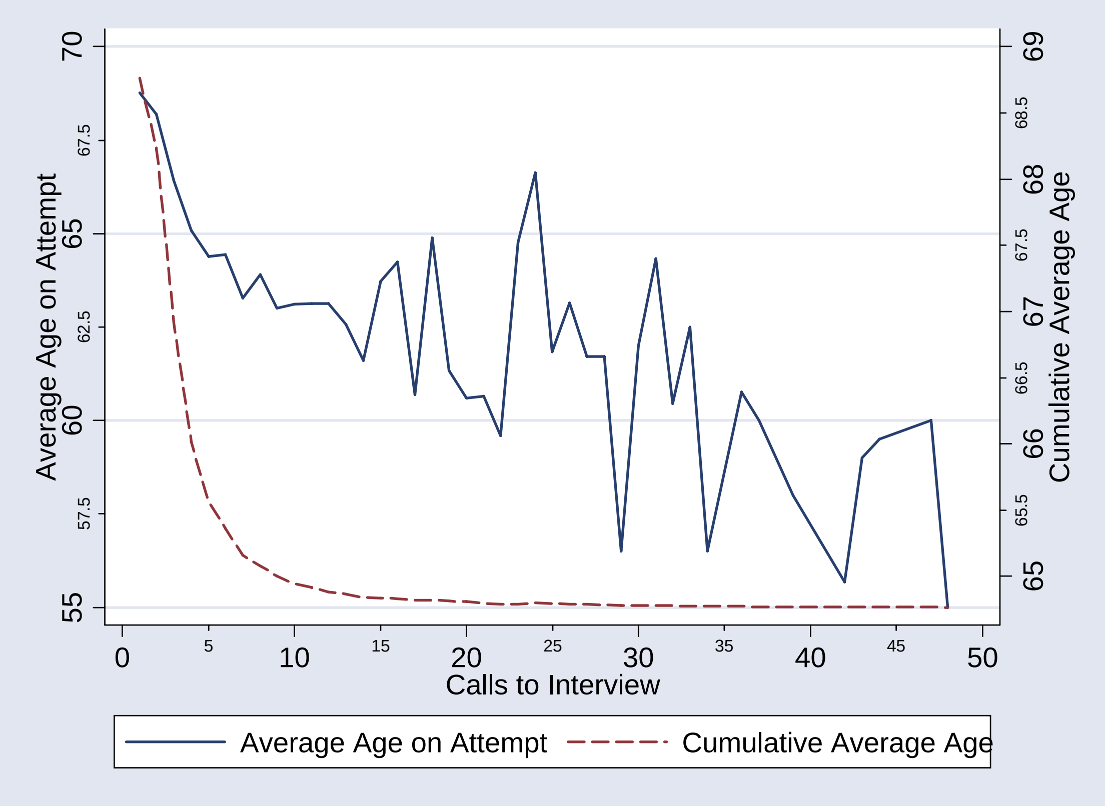
Web Dynamic Reporting System (Web-DRS)

- Uses interviewer-level, case-level, and attempt/dial-level data

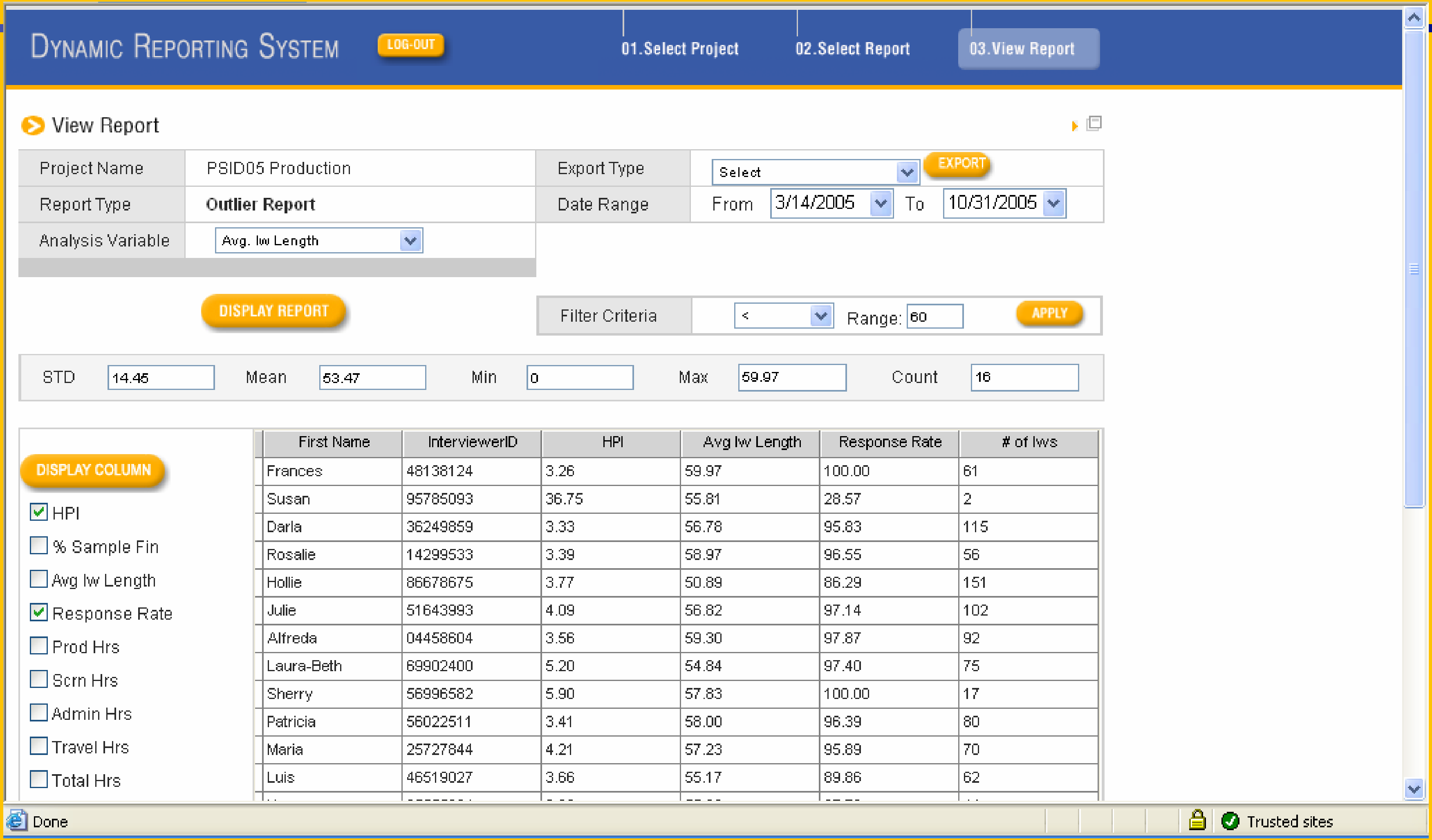
- Four reports
  - Outlier (Interviewer Level)
  - Trend (Project Level)
  - Case Analysis (Sample Line Level)
  - Key Statistics (Call Level)

DRS-Key Stats Report

Sample Report for Retirement Study (Respondent Age)



DRS-Outlier Report



View Report

Project Name	PSID05 Production	Export Type	Select (EXPORT)
Report Type	**Outlier Report**	Date Range	From 3/14/2005 To 10/31/2005
Analysis Variable	Avg. Iw Length		

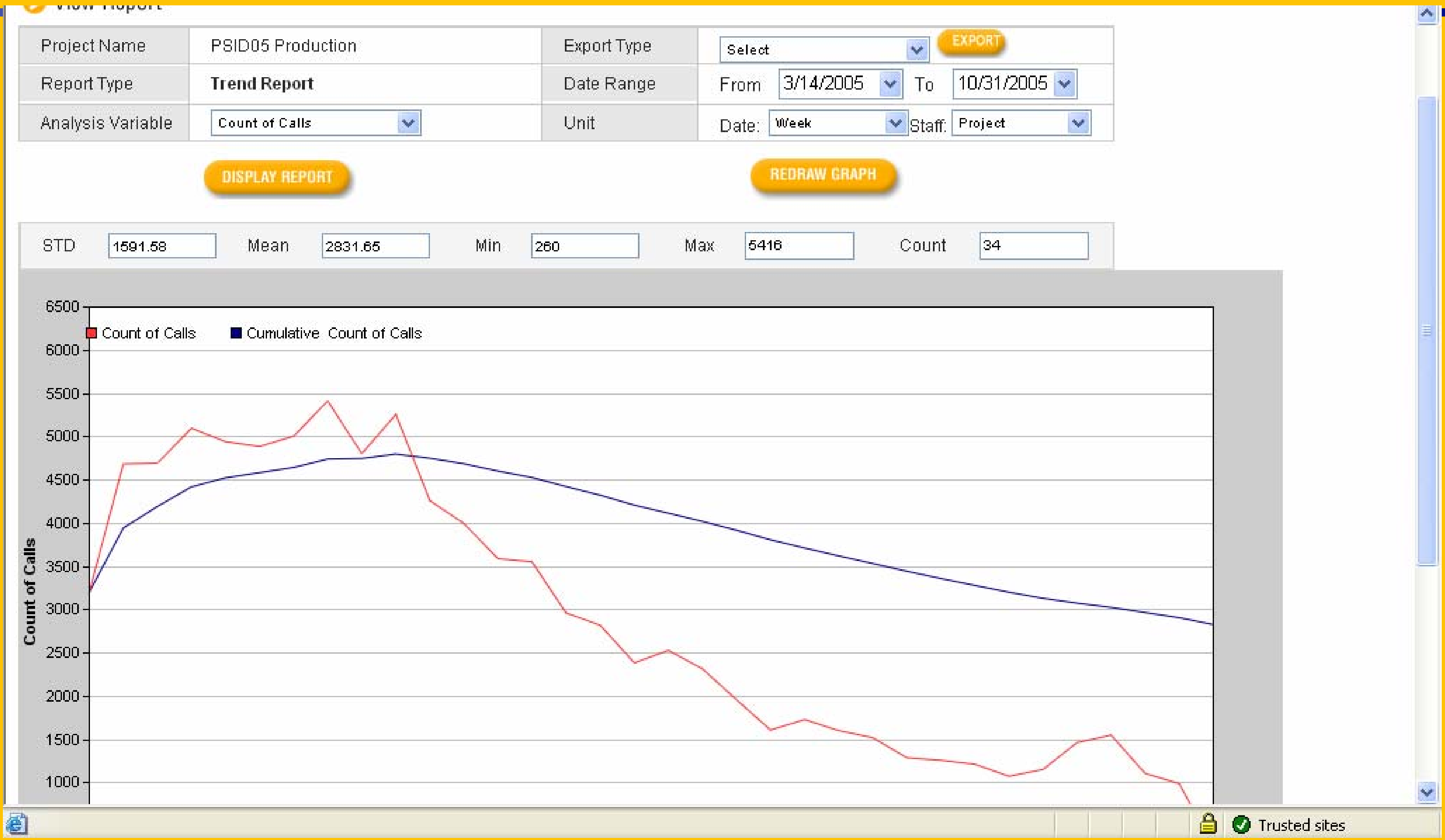
Filter Criteria: < Range: 60

STD 14.45 Mean 53.47 Min 0 Max 59.97 Count 16

- ☑ HPI
- ☐ % Sample Fin
- ☐ Avg Iw Length
- ☑ Response Rate
- ☐ Prod Hrs
- ☐ Scrn Hrs
- ☐ Admin Hrs
- ☐ Travel Hrs
- ☐ Total Hrs

First Name	InterviewerID	HPI	Avg Iw Length	Response Rate	# of Iws
Frances	48138124	3.26	59.97	100.00	61
Susan	95785093	36.75	55.81	28.57	2
Darla	36249859	3.33	56.78	95.83	115
Rosalie	14299533	3.39	58.97	96.55	56
Hollie	86678675	3.77	50.89	86.29	151
Julie	51643993	4.09	56.82	97.14	102
Alfreda	04458604	3.56	59.30	97.87	92
Laura-Beth	69902400	5.20	54.84	97.40	75
Sherry	56996582	5.90	57.83	100.00	17
Patricia	56022511	3.41	58.00	96.39	80
Maria	25727844	4.21	57.23	95.89	70
Luis	46519027	3.66	55.17	89.86	62

DRS-Outlier Report



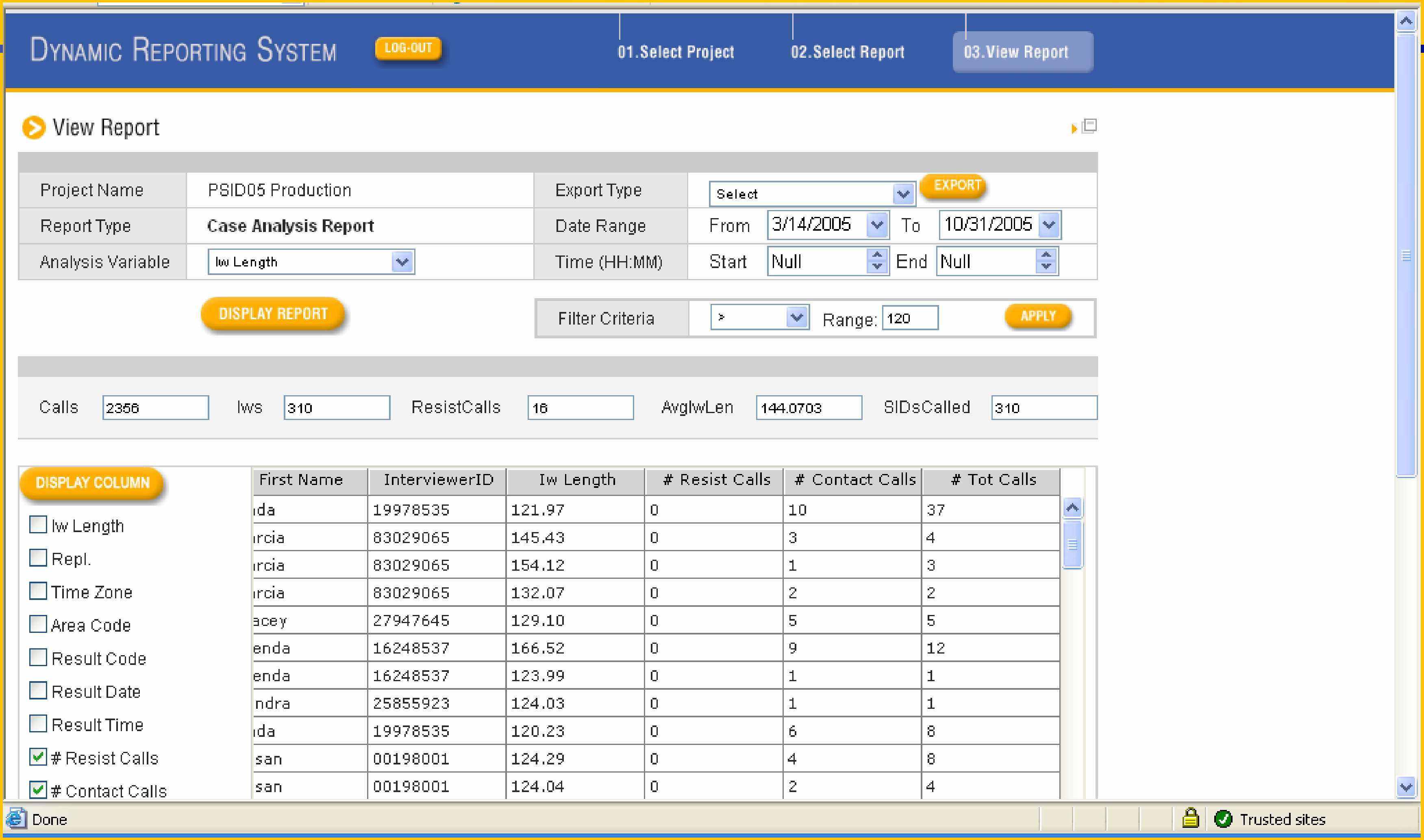
View Report

Project Name	PSID05 Production	Export Type	Select EXPORT
Report Type	**Trend Report**	Date Range	From 3/14/2005 To 10/31/2005
Analysis Variable	Count of Calls	Unit	Date: Week Staff: Project

DISPLAY REPORT REDRAW GRAPH

STD	1591.58	Mean	2831.65	Min	260	Max	5416	Count	34

DRS-Case Analysis Report



View Report

Project Name	PSID05 Production	Export Type	Select EXPORT
Report Type	Case Analysis Report	Date Range	From 3/14/2005 To 10/31/2005
Analysis Variable	Iw Length	Time (HH:MM)	Start Null End Null

Filter Criteria > Range: 120

Calls 2356 Iws 310 ResistCalls 16 AvgIwLen 144.0703 SIDsCalled 310

- ☐ Iw Length
- ☐ Repl.
- ☐ Time Zone
- ☐ Area Code
- ☐ Result Code
- ☐ Result Date
- ☐ Result Time
- ☑ # Resist Calls
- ☑ # Contact Calls

First Name	InterviewerID	Iw Length	# Resist Calls	# Contact Calls	# Tot Calls
ıda	19978535	121.97	0	10	37
ırcia	83029065	145.43	0	3	4
ırcia	83029065	154.12	0	1	3
ırcia	83029065	132.07	0	2	2
acey	27947645	129.10	0	5	5
enda	16248537	166.52	0	9	12
enda	16248537	123.99	0	1	1
ndra	25855923	124.03	0	1	1
ıda	19978535	120.23	0	6	8
san	00198001	124.29	0	4	8
san	00198001	124.04	0	2	4

Done

Trusted sites

Future trends in documentation and archiving

- Collaboration & consortiums
- Increase & improve XML standards for documenting
  - Rollout of DDI draft version 3 IASSIST
    - *International Association for Social Science Information Service and Technology* meeting 2006 May 23-26, 2006, in Ann Arbor
    - Sponsored by ICPSR, UM School of Information, and University Libraries
  - Official release expected at the end 2006

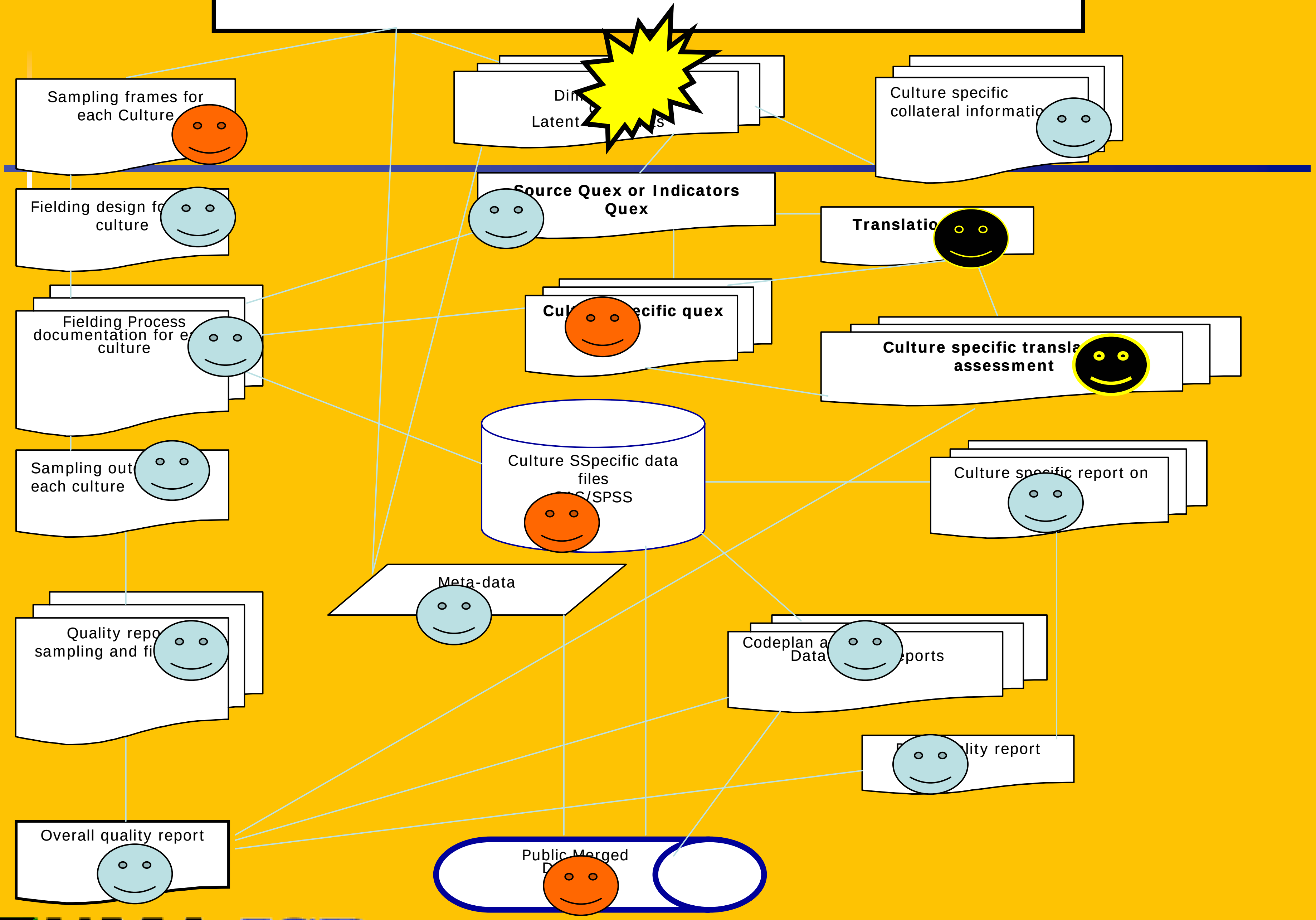
However

“Simple solutions seldom are.”

Alfred North Whitehead

“The best way to predict the future is to invent it.” *Alan Kay*

Total Package Blue Print



Sampling frames for each Culture

Di Latent s

Culture specific collateral informatio

Source Quex or Indicators Quex

Translatio

Fielding design fo culture

Fielding Process documentation for e culture

Cul ecific quex

Culture specific transla assessment

Culture SSpecific data files S S/SPSS

Sampling out each culture

Culture specific report on

Meta-data

Quality repo sampling and fi

Codeplan a Data eports

lity report

Overall quality report

Public Merged D

THANK YOU !

Proceedings of Q2006
European Conference on Quality in Survey Statistics

Comparison of two statistical metadata models: SDMX and CoSSI. How well do they guide the user to proper interpretation of statistical information?

Jaakko Ranta[1]

1. Introduction

In spite of the efforts worldwide no generic widely approved standard for statistical metadata has been reached. Most statistical institutes have developed their metadata processes based on their specific local features. The outcome of this is that comparing the statistics from different sources is usually very difficult even though they in principle describe same phenomenon.

The two metadata models compared in this presentation reach for wider approval as a metadata standard for statistical information:

SDMX (the Statistical Data and Metadata Exchange initiative) is sponsored by BIS, ECB, EUROSTAT, IMF, OECD, UN, and the World Bank. The model to be compared in the paper is described in the SDMX Information Model: UML Conceptual design (Version 2.0.) Another main source used is SDMX Implementors Guide (Version 2.0). Available on the web at: http://www.sdmx.org

CoSSI (Common Structure of Statistical Information) model has been developed in Statistics Finland. Definition descriptions available on the web at: http://www.stat.fi/cossi

The comparison focuses on the capability of the two models to provide the user with rich metadata and to express quality aspects of statistics.

The models are described here only to the extent necessary to clarify the differences from the point of view of our interest.

[1] Jaakko Ranta, Box 3A, Finland, 000022 Statistics Finland, jaakko.ranta@stat.fi

2. Categorisation of Metadata

The categorisation of metadata used in the models differ from each other substantially.

SDMX

- Structural metadata
  Those concepts used in description and identification of statistical data and metadata.
- Reference metadata
  Larger set of concepts that describe and qualify statistical data sets and processing more generally.

CoSSI

- Statistical metadata
  Content-specific metadata necessary for the interpretation of statistical figures.
- Document metadata
  Information about:
  - The producer of document
  - Document’s content
- Processing metadata
  Information for a software to process data.

3. SDMX Model in Brief

In SDMX structural metadata always connect to Data structure definition (Figure 1.), which comprises three types of descriptor concepts:

- dimensions
  both describe and identify the data.
- (data)attributes
  are purely descriptive.
- measures

Each of the descriptor components is assigned a type representation, e.g. a code list, a date, a numeric range, text etc. There always is a code list telling the possible values for each dimension. The Key comprises the Dimensions, whose values in the data set uniquely identify the observed data values. The Group Key comprises a sub set of the dimensions.
The List of Attributes comprises attributes that can be used to give metadata about some part of the data set. Each attribute must be assigned to an identified part of the data set (attachment level): an observation, a key, a group key or a dataset.

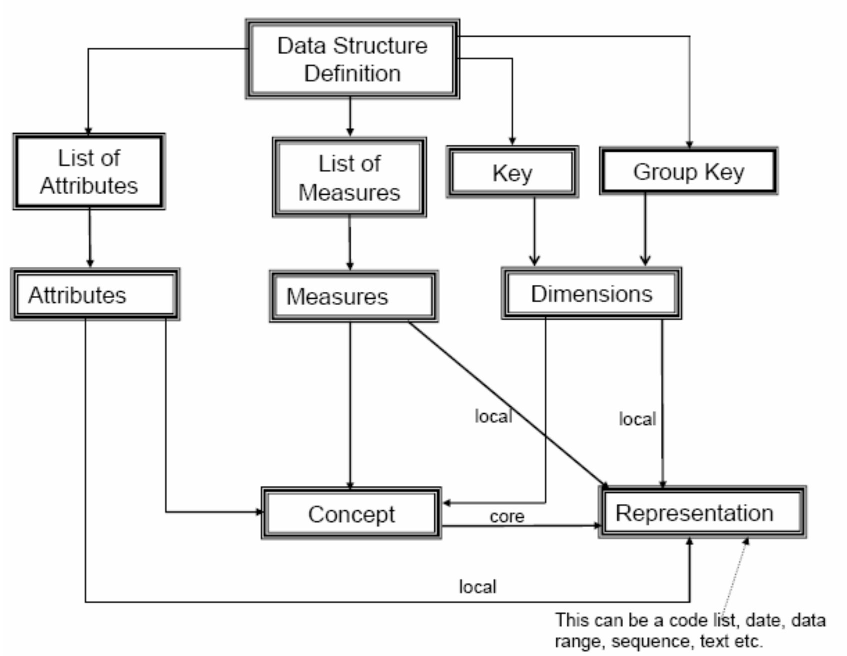


Figure 1: The Data Structure Definition (SDMX Implementors Guide, p. 62)

The List of Measures comprises measures. each of which is a phenomenon for which an observation is relevant.

The data set is linked to data structure definition (Key Family Definition) via Data Flow (Figure 2.)

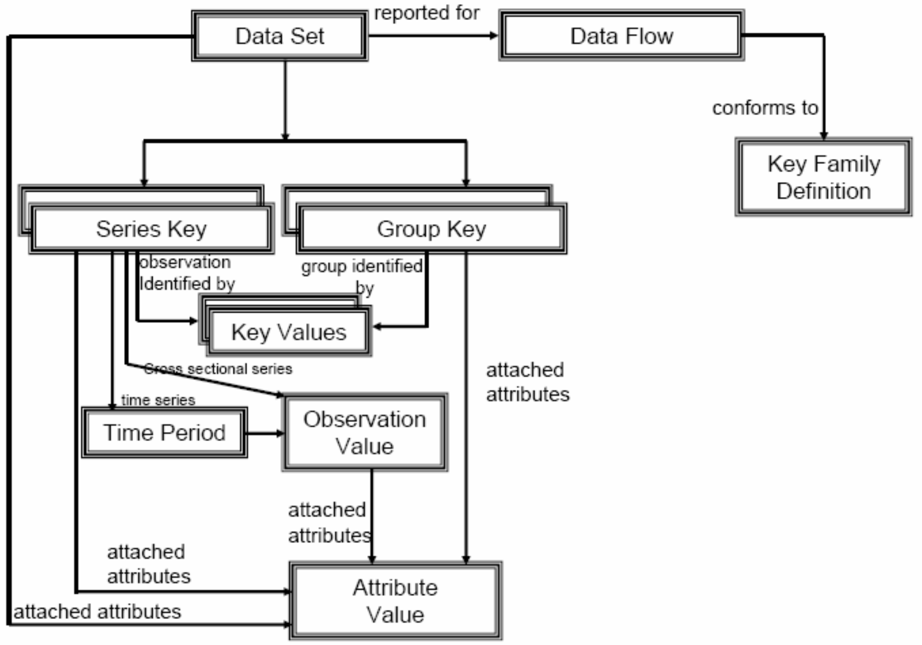
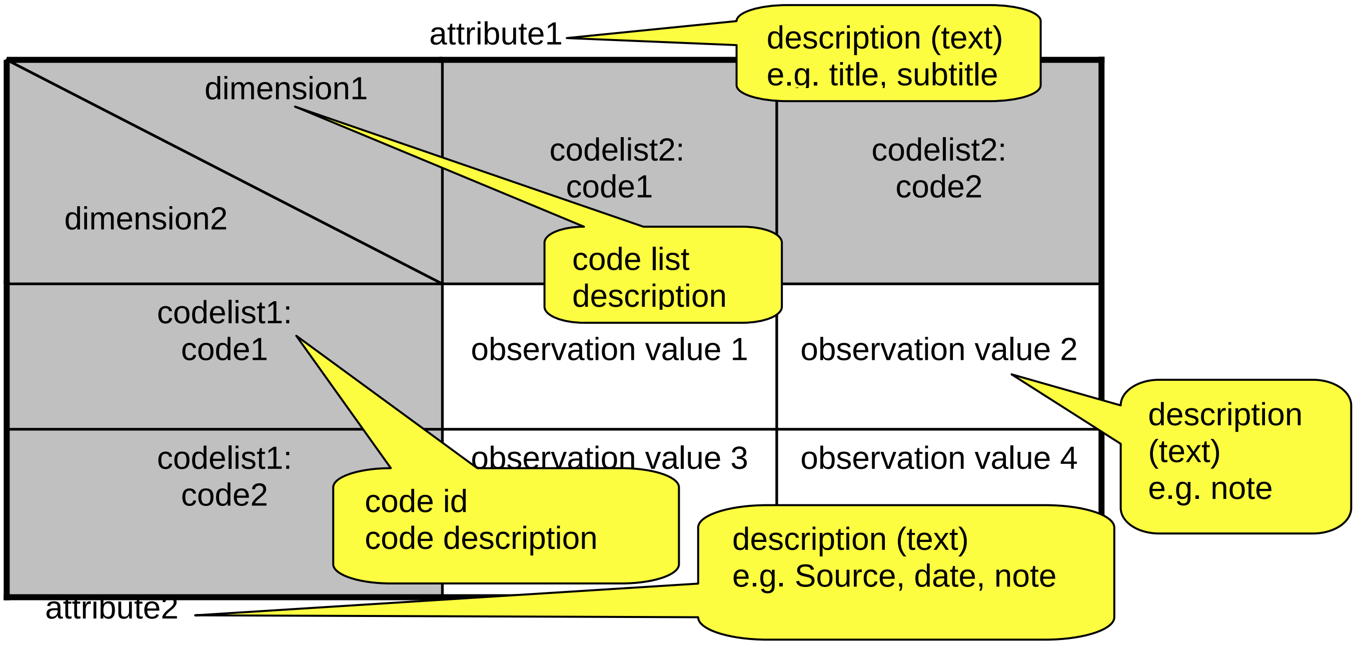


Figure 2: The Data Set (SDMX Implementors Guide, p. 98)

The main structure of the Data Set is a set of Keys and Group Keys. Each Key comprises Key Values, a value for each of the Dimensions defined in data structure Definition (Key Family). For each key there may be one or more Observation values: for time series Observation Value is related to a Time Period, whereas for cross sectional series it is not. Attribute Values can be attached to one of Data Set, Series Key or Group Key.

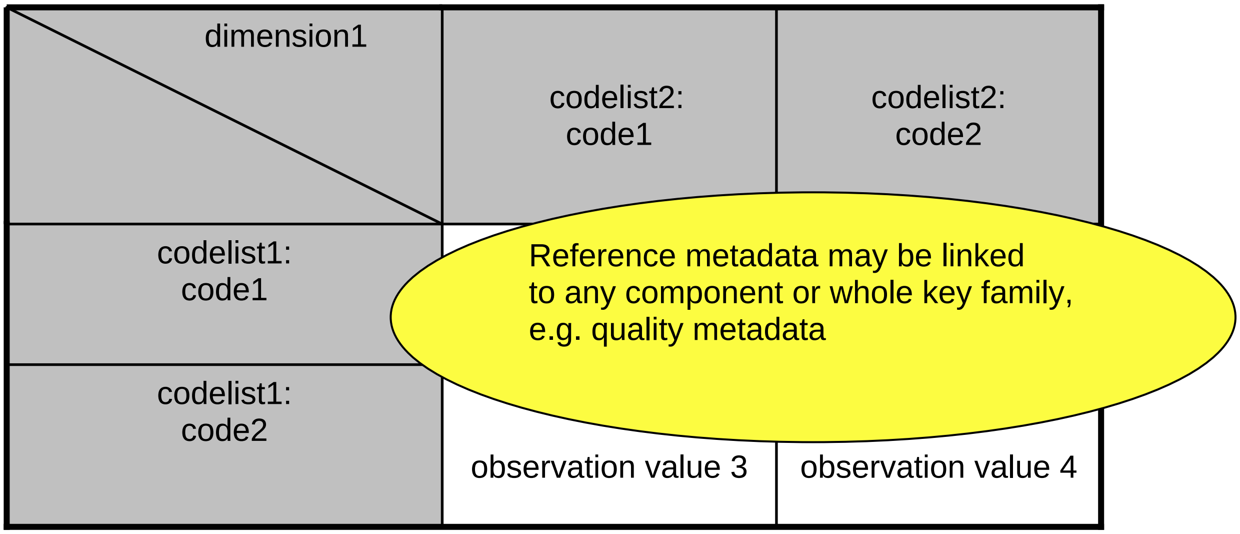
SDMX reference metadata is the metadata not defined in the data structure definition and corresponding data set. SDMX information model is applied to this outside metadata in a similar way as it is applied to data: Metadata structure definition defines the structure of metadata set.
Metadata structure definition defines how to attach metadata to data (Data structure definition or its components).

The tables 1. and 2. show how the metadata can be attached to a statistical table.



dimension1 dimension2	codelist2: code1	codelist2: code2
codelist1: code1	observation value 1	observation value 2
codelist1: code2	observation value 3	observation value 4

Table 1. Structural metadata in a statistical table according to SDMX



dimension1	codelist2: code1	codelist2: code2
codelist1: code1		
codelist1: code2	observation value 3	observation value 4

Reference metadata may be linked to any component or whole key family, e.g. quality metadata

Table 2. Reference metadata in a statistical table according to SDMX

4. CoSSI Model in Brief

Starting points:

- statistical information in modelled, not the real world
- statistical data are defined and describe themselves exhaustively
- structuring of statistical information
- managing statistical information as a single entity

CoSSI is a Modular DTD system:

- document type definitions
- Standards: CALS, XDF(Extended Data Format, developed in NASA), Dublin Core
- XML: one file – data and metadata

In CoSSI on the one hand, statistical information has been defined by using a conceptual analysis, the results from which have been depicted as conceptual models of statistical information and on the other hand, an analysis has been made of different forms of organising statistical data and presenting statistical information, which has been used to specify basic models for presenting statistical data. The outcome is described in figure 3. Structural models of data and related data models have been produced for concept models and different forms of organising data, and definitions for these have been implemented in the CoSSI model as multi-level hierarchical (so-called tree-structured) data models . The data models have been documented as XML DTD definitions.

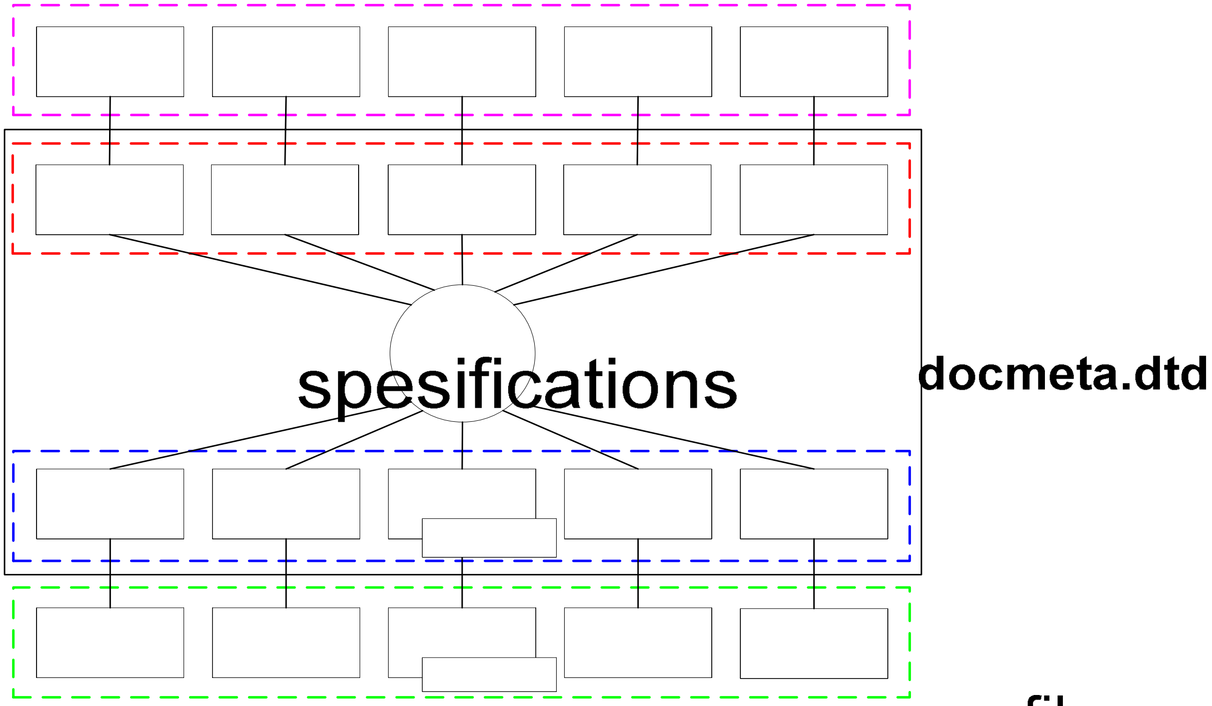


Figure 3. Common Structure of Statistical Information (CoSSI) - parts and entity

As an example of the concept models of the upper part of the figure 3 the logical concept model of statistical metadata is illustrated in Figure 4.

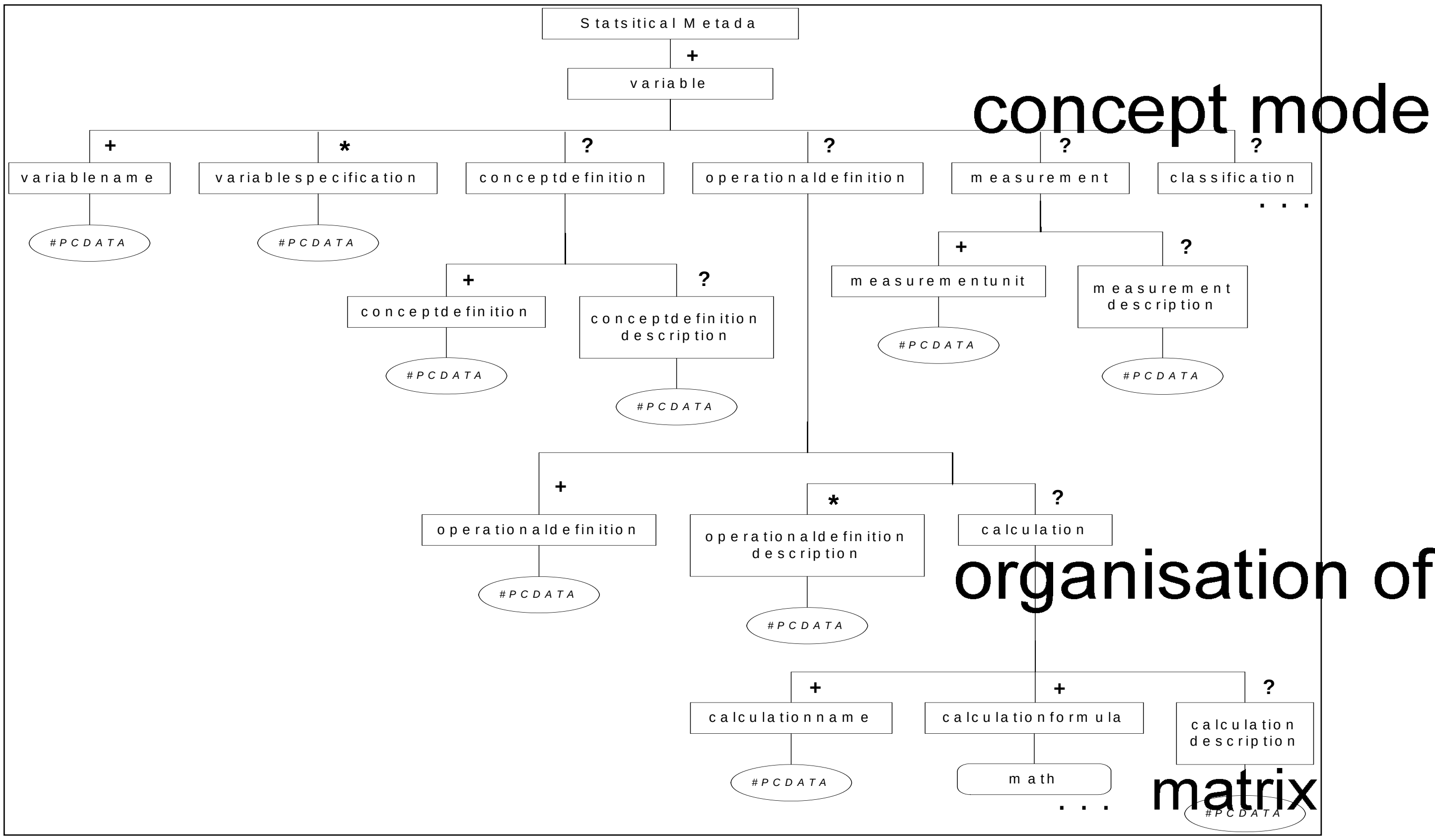
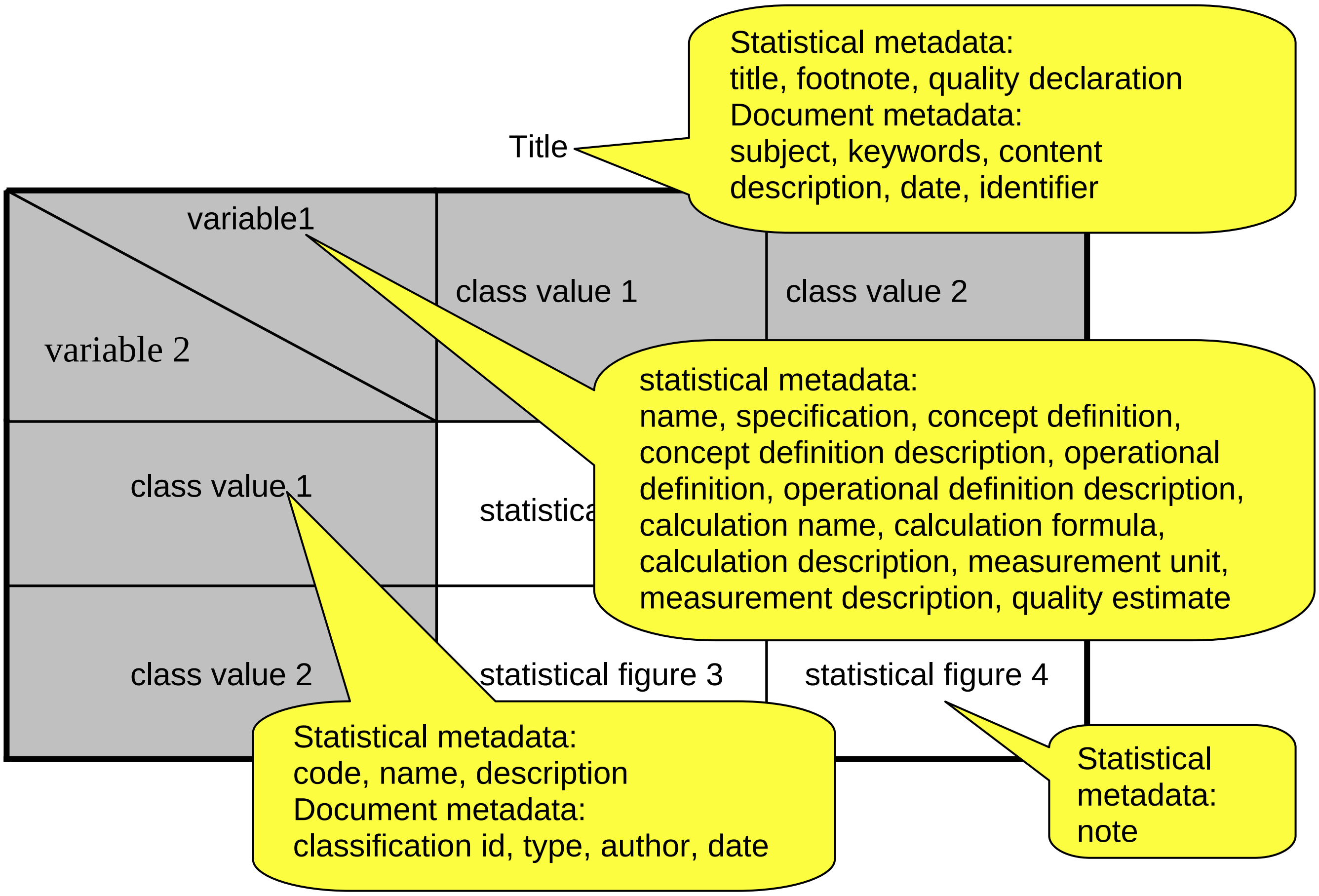


Figure 4. The logical data model of statistical metadata

The table 3 show how the CoSSI model metadata can be attached to a statistical table.



Statistical metadata:
title, footnote, quality declaration
Document metadata:
subject, keywords, content description, date, identifier

statistical metadata:
name, specification, concept definition, concept definition description, operational definition, operational definition description, calculation name, calculation formula, calculation description, measurement unit, measurement description, quality estimate

Statistical metadata:
code, name, description
Document metadata:
classification id, type, author, date

Statistical metadata:
note

Table 3. Metadata in a statistical table according to CoSSI

5 Conclusions

5.1 Generality of the models

<u>SDMX</u>

- SDMX could be describes as super generic. It is open for any kind of data structure definitions and metadata structure definitions.
- To be able to use the model in a rational way, an agreement about the data and metadata structure of statistics is needed among the parties wanting to share information.
- A special XML schema or dtd is needed for each data set and corresponding data structure definition.

<u>CoSSI</u>

- In CoSSI the elements of metadata are fixed. They are defined in the logical concept model and implemented in the corresponding dtd.
- Just one dtd is needed for each type of organisation of data, e.g. table.dtd, matrix.dtd).
- CoSSI is still open for expansion.
- Not all metadata elements need to be used, if the metadata is not available.

5.2 Entity of data and metadata

In SDMX structural metadata is attached to data, but reference metadata is in one or more metadata sets outside of data set. Linking of reference metadata to data is made from metadata set, not from data set.

In CoSSI tables or matrixes and variables in them are directly attached to corresponding metadata.

5.3. Richness and expandability of metadata

SDMX

- The structural metadata is somewhat limited in quantity and deepness. Any number of attributes can be added, but they always are attached to the data at the same hierarchical level.

- There is no limit how much reference metadata there is in the separate metadata sets. The necessity to define a metadata structure definition for each metadata set makes it rather heavy and restrictive procedure.

-

CoSSI

- The metadata elements are designed to cover the metadata needs as far as possible, but If needed, the model and the dtd can expanded both horizontally and vertically.

The figure 5. illustrates how the models cover the entity of metadata connected a phenomenon. As In CoSSI the elements of metadata are fixed there is no need to define them in each case one by one. In SDMX must be defined for each data set separately.

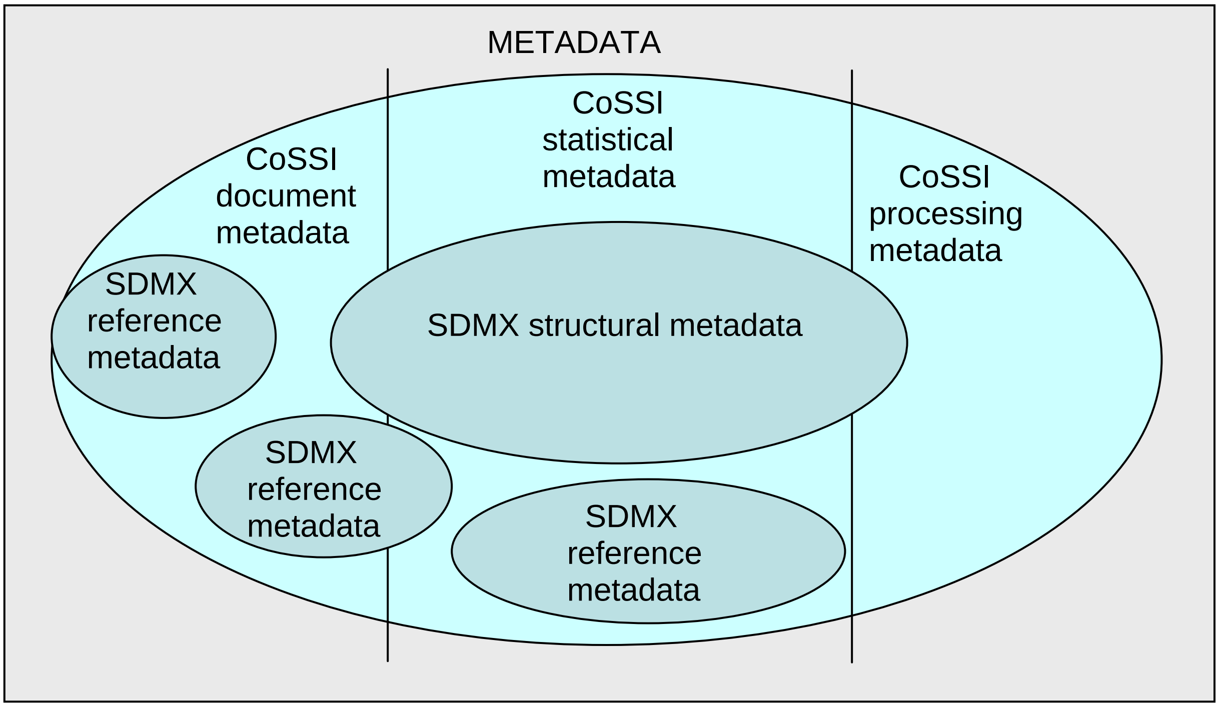


Figure 5. Metadata connected to the statistical description of a phenomenon.

5.4. Transparency of metadata

For the users to be able to evaluate the usefulness of statistical data all the relevant statistical metadata should be obtainable, e.g.

- about how a survey was defined and what asked
- about quality aspects

In SDMX reference metadata is the way to deliver this kind of information. The problem is that, there is no formalised way to attach this kind of metadata and no way to directly point to that metadata set from the presentation of statistical information, e.g. statistical table.

To CoSSI some formalisations have been or are to be added: quality declaration as an additional module (quality declaration.dtd) and quality estimate as a vertical expansion of statistical metadata module (statmeta.dtd), see Rouhuvirta (2006).

5.5. Mapping between the models

The mapping between the models as such is not possible, because of the generality of SDMX. The model of SDMX doesn’t have any statistical specific information. All statistic specific information has to be defined using domains. So how the mappings should be done depends on the way domain specifications have been realised in practice. The richness of CoSSI information content makes the mappings quite possible even for the most complicated domain specifications.

References

Framework for SDMX Technical Standards (Version 2.0), (2005). Available on the web at: http://www.sdmx.org

Rouhuvirta, H. (2001), "On The Structuring of Statistical Information", Originally presented at First MetaNet Conference, Voorburg, 2001, also available on the Internet at: http://www.stat.fi/org/tut/dthemes/papers/structuring_statistical_information_2001.pdf

Rouhuvirta, H. (2006), "Methodologically Oriented Metadata - Framework for Enhanced Quality Information of Statistics", Proceedings of European Conference on Quality in Survey Statistics 2006.

Rouhuvirta, H. and Lehtinen, H. (2003), "Common Structure of Statistical Information (CoSSI) - Definition Descriptions", 2nd December 2003, Version 0.91, Statistics Finland 2003, also available on the Internet at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf

SDMX Information Model: UML Conceptual design (Version 2.0.), (2005). Available on the web at: http://www.sdmx.org

SDMX Implementors Guide (Version 2.0), (2005). Available on the web at: http://www.sdmx.org

Comparison of two statistical metadata models:

SDMX and CoSSI.

How well do they guide the user to proper interpretation of statistical information?

- In spite of the efforts worldwide no generic widely approved standard for statistical metadata has been reached.

- Most statistical institutes have developed their metadata processes based on their specific local features.

- The outcome of this is that comparing the statistics from different sources is usually very difficult even though they in principle describe same phenomenon.

The two metadata models compared in this presentation reach for wider approval as a metadata standard for statistical information:

- SDMX (the Statistical Data and Metadata Exchange initiative) is sponsored by BIS, ECB, EUROSTAT, IMF, OECD, UN, and the World Bank. The model to be compared in the paper is described in the SDMX Information Model: UML Conceptual design (Version 2.0.) Another main source used is SDMX Implementors Guide (Version 2.0). Available on the web at: http://www.sdmx.org

- CoSSI (Common Structure of Statistical Information) model has been developed in Statistics Finland. Definition descriptions available on the web at: http://www.stat.fi/cossi

- The comparison focuses on the capability of the two models to provide the user with rich metadata and to express quality aspects of statistics.
- The models are described here only to the extent necessary to clarify the differences from the point of view of our interest.

The categorization of metadata used in the models differ from each other substantially.

SDMX

- **Structural metadata**

  Those concepts used in description and identification of statistical data and metadata
- **Reference metadata**

  Larger set of concepts that describe and qualify statistical data sets and processing more generally

CoSSI

- **Statistical metadata**

  Content-specific metadata necessary for the interpretation of statistical figures.
- **Document metadata**

  Information about:

  - The producer of document

  - Document's content
- **Processing metadata**

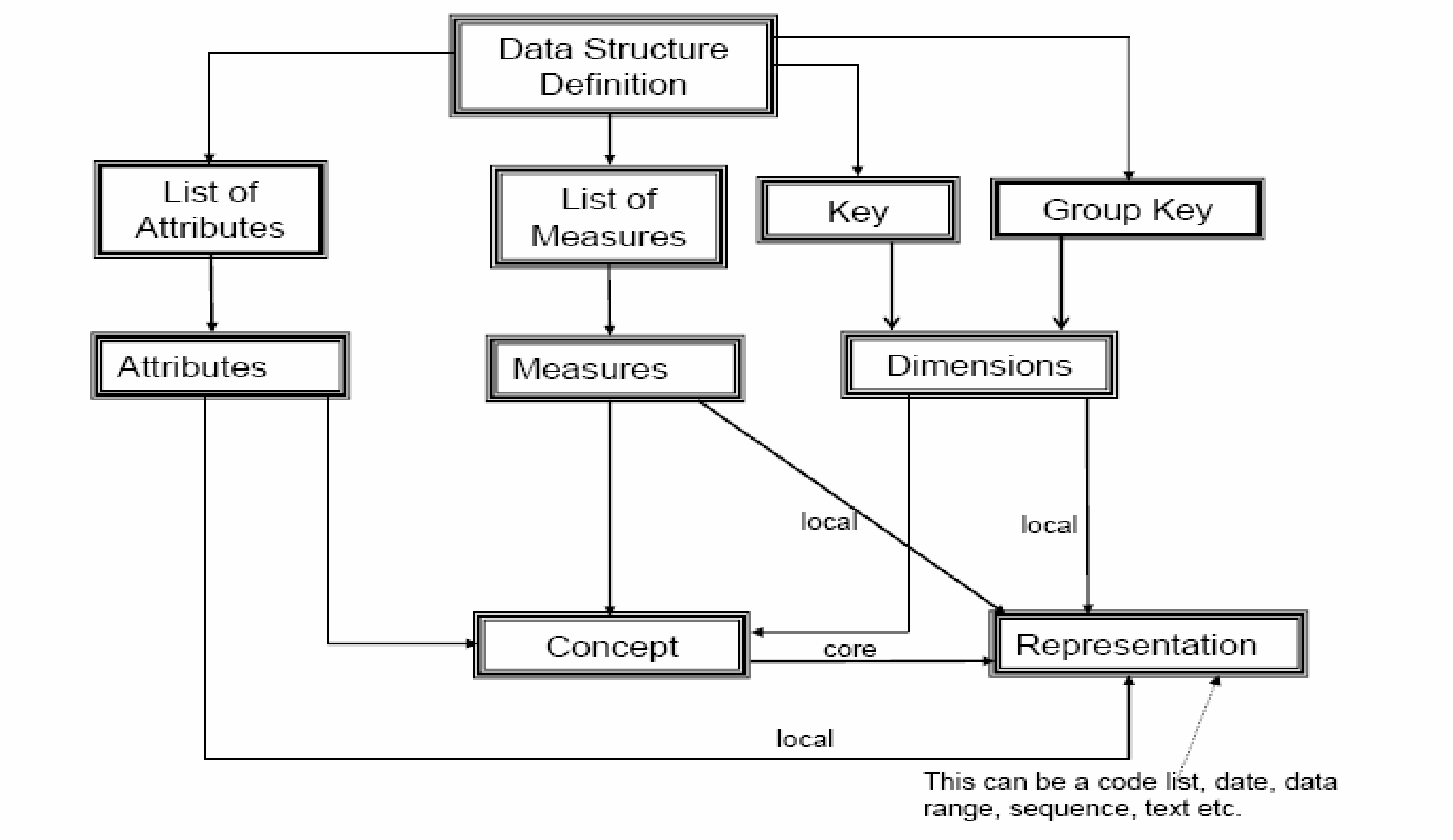
  - Information for a software to process data

In SDMX structural metadata always connect to **Key family** (Data structure definition).
Key family comprises three types of descriptor concepts:

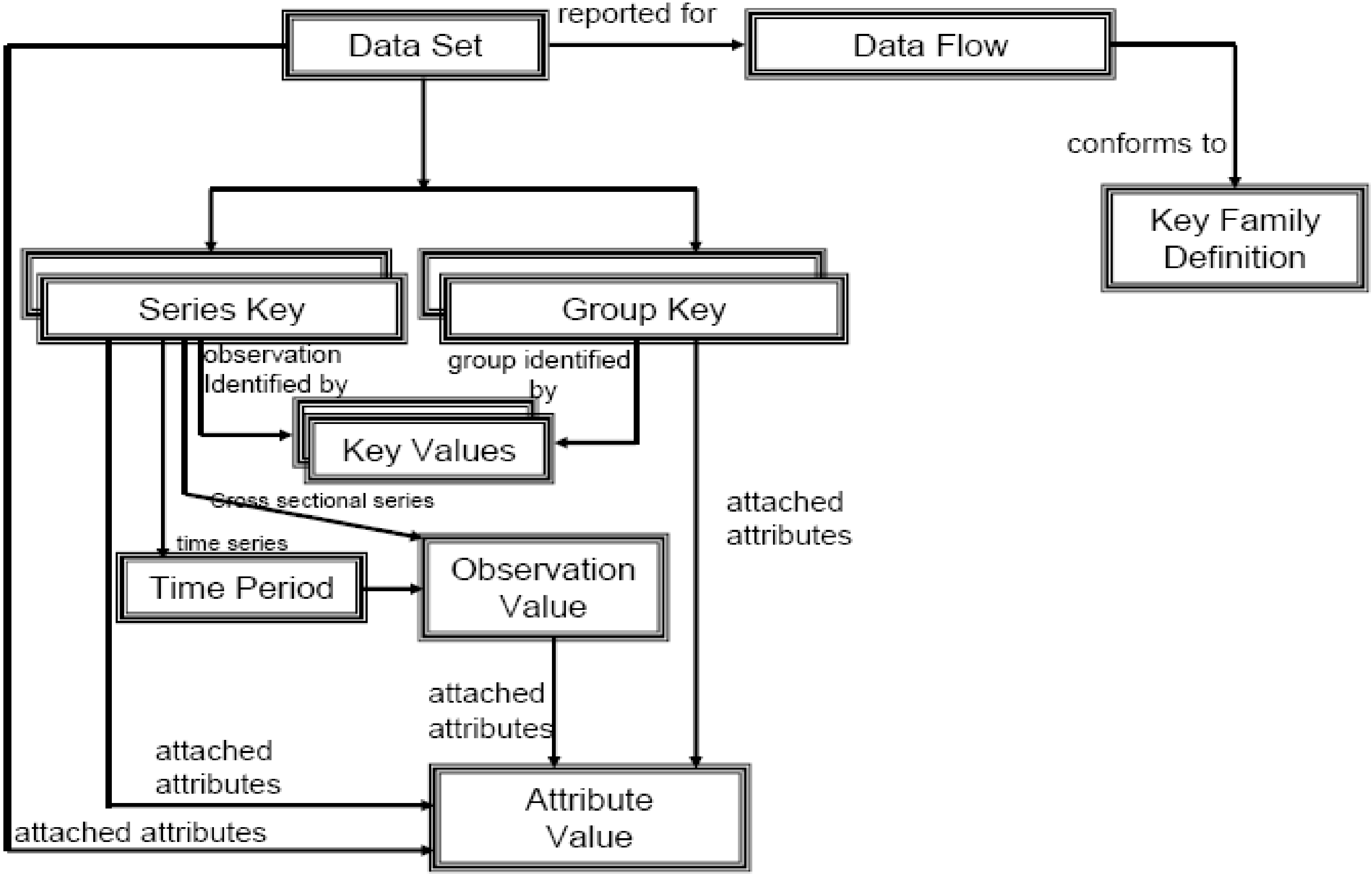
- **dimensions**
  both describe and identify the data.
- **(data) attributes**
  are purely descriptive.
- **measures**

- Each of the descriptor components is assigned a type representation, e.g. a code list, a date, a numeric range, text etc.
- There always is a code list telling the possible values for each dimension.

Data structure definition (Key Family)



Data Set

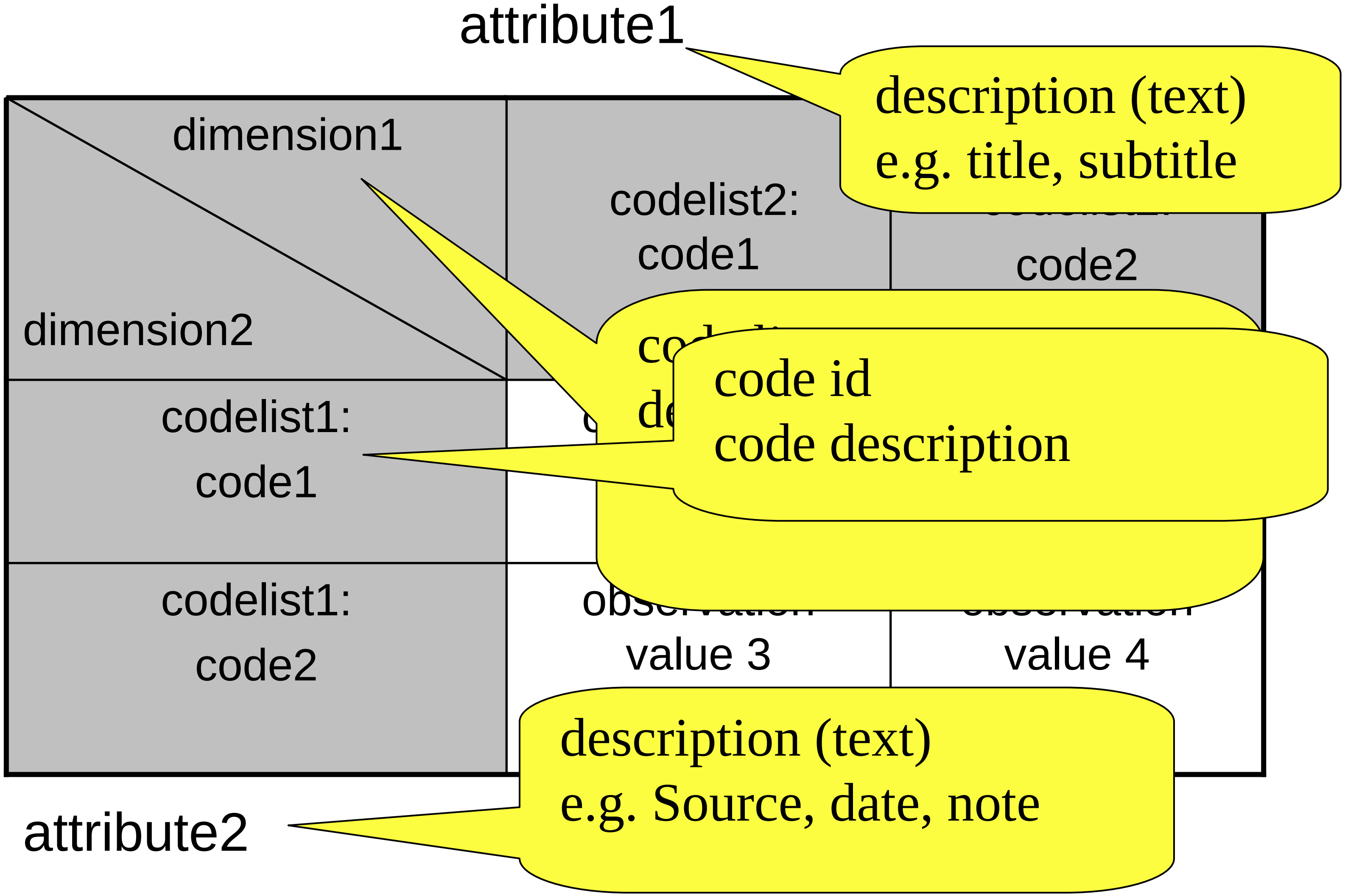

Data Set
reported for
Data Flow
conforms to
Key Family Definition
Series Key
Group Key
observation Identified by
group identified by
Key Values
Cross sectional series
time series
Time Period
Observation Value
attached attributes
attached attributes
attached attributes
attached attributes
Attribute Value

SDMX reference metadata is the metadata not defined in key family and corresponding data set.
SDMX information model is applied to this outside metadata in a similar way as it is applied to data:

- **Metadata structure definition** defines the structure of
- **metadata set.**

Metadata structure definition defines how to attach metadata to data (key family or key family components).

Metadata (structural) in statistical table according to SDMX



attribute1

description (text)
e.g. title, subtitle

dimension1

dimension2

codelist2:
code1

code2

co... li...
de...

code id
code description

codelist1:
code1

codelist1:
code2

observation
value 3

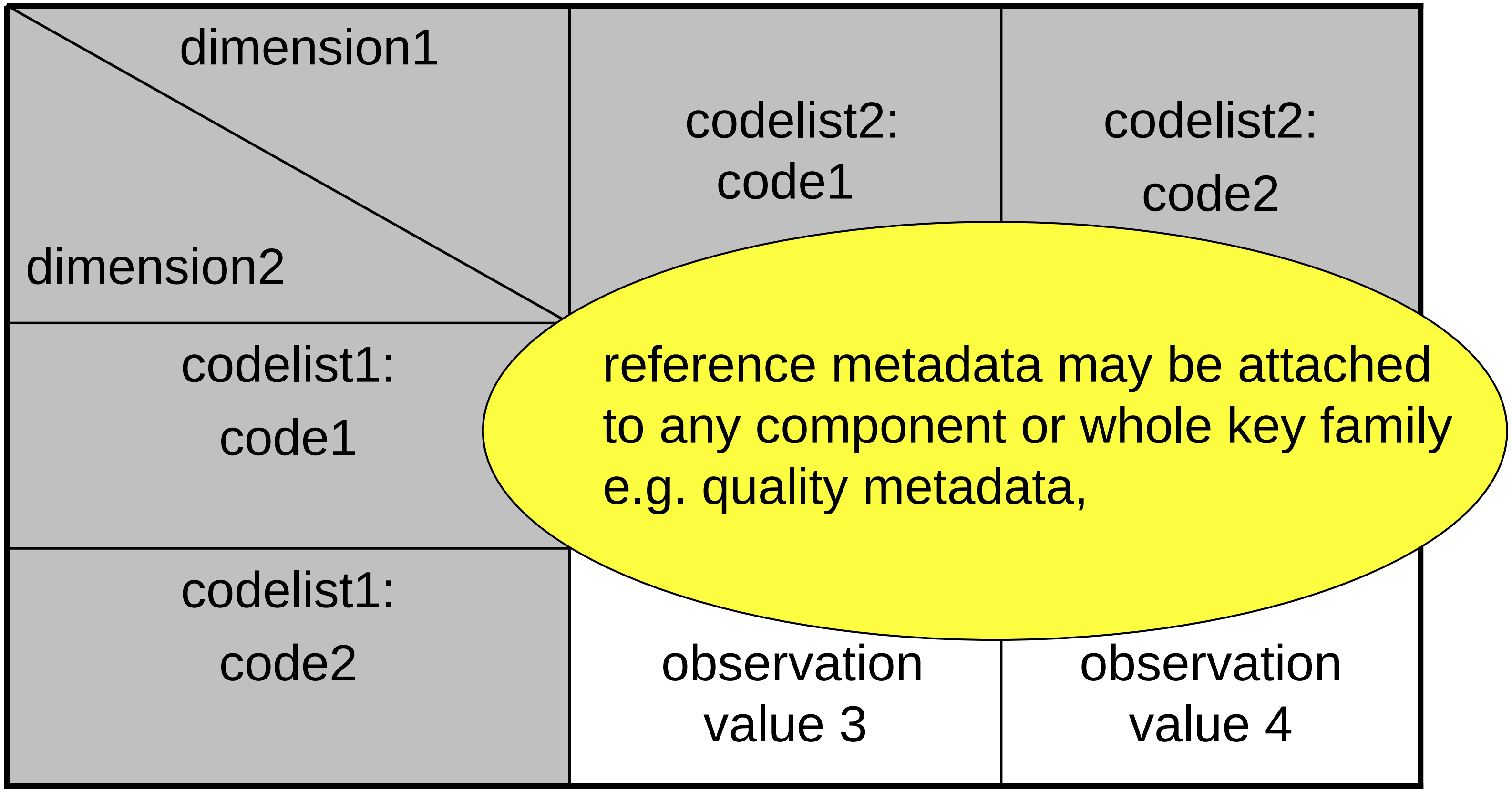
observation
value 4

description (text)
e.g. Source, date, note

attribute2

Statistical table according to SDMX

attribute1 (e.g. Title)



dimension1 dimension2	codelist2: code1	codelist2: code2
codelist1: code1		
codelist1: code2	observation value 3	observation value 4

reference metadata may be attached
to any component or whole key family
e.g. quality metadata,

attribute2 (e.g. Source)

CoSSI Statistical Information model

Conceptual modelling of statistical information

Starting points:

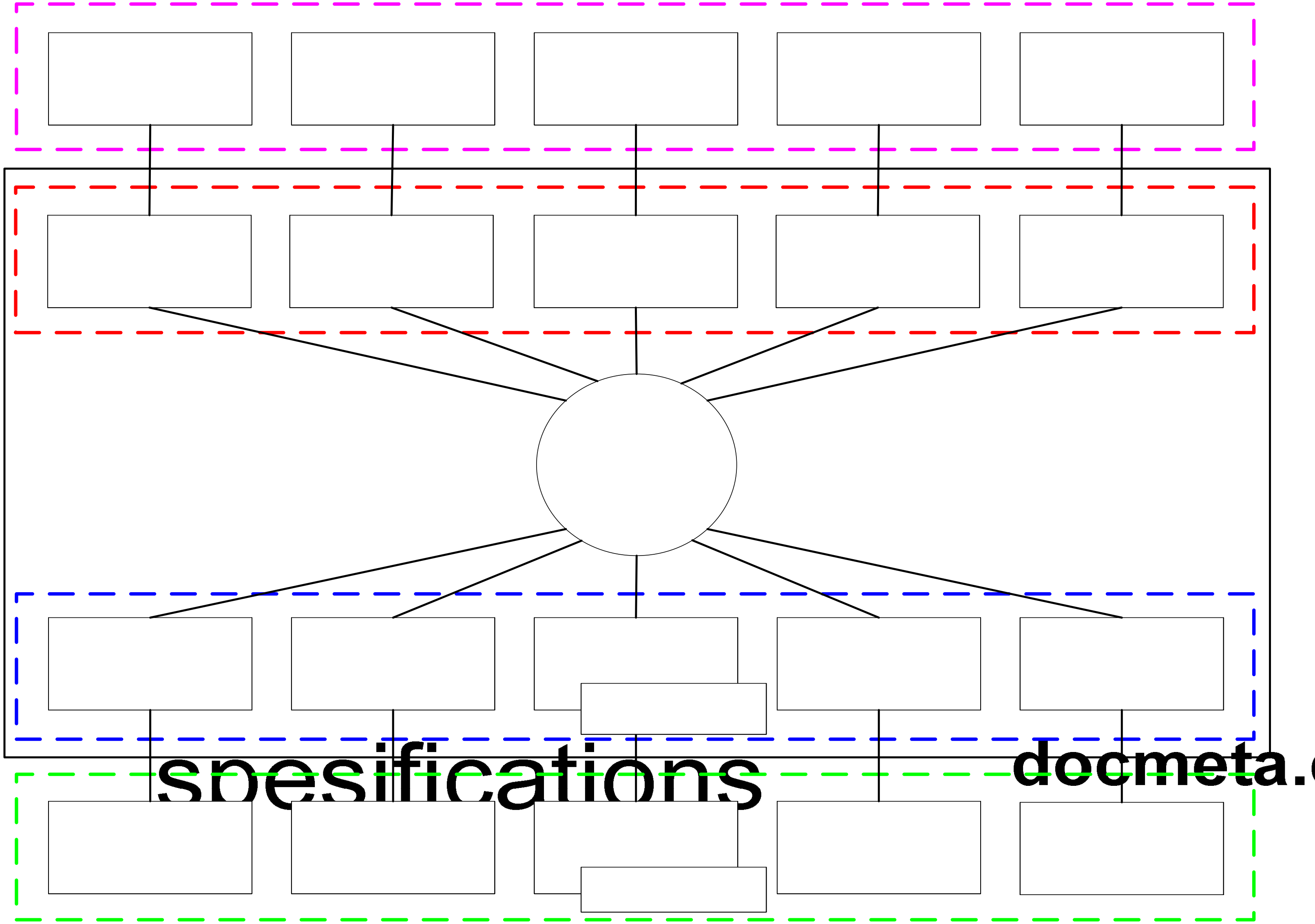
- statistical information in modelled, not the real world
- statistical data are defined and describe themselves exhaustively
- structuring of statistical information
- managing statistical information as a single entity

CoSSI Implementation

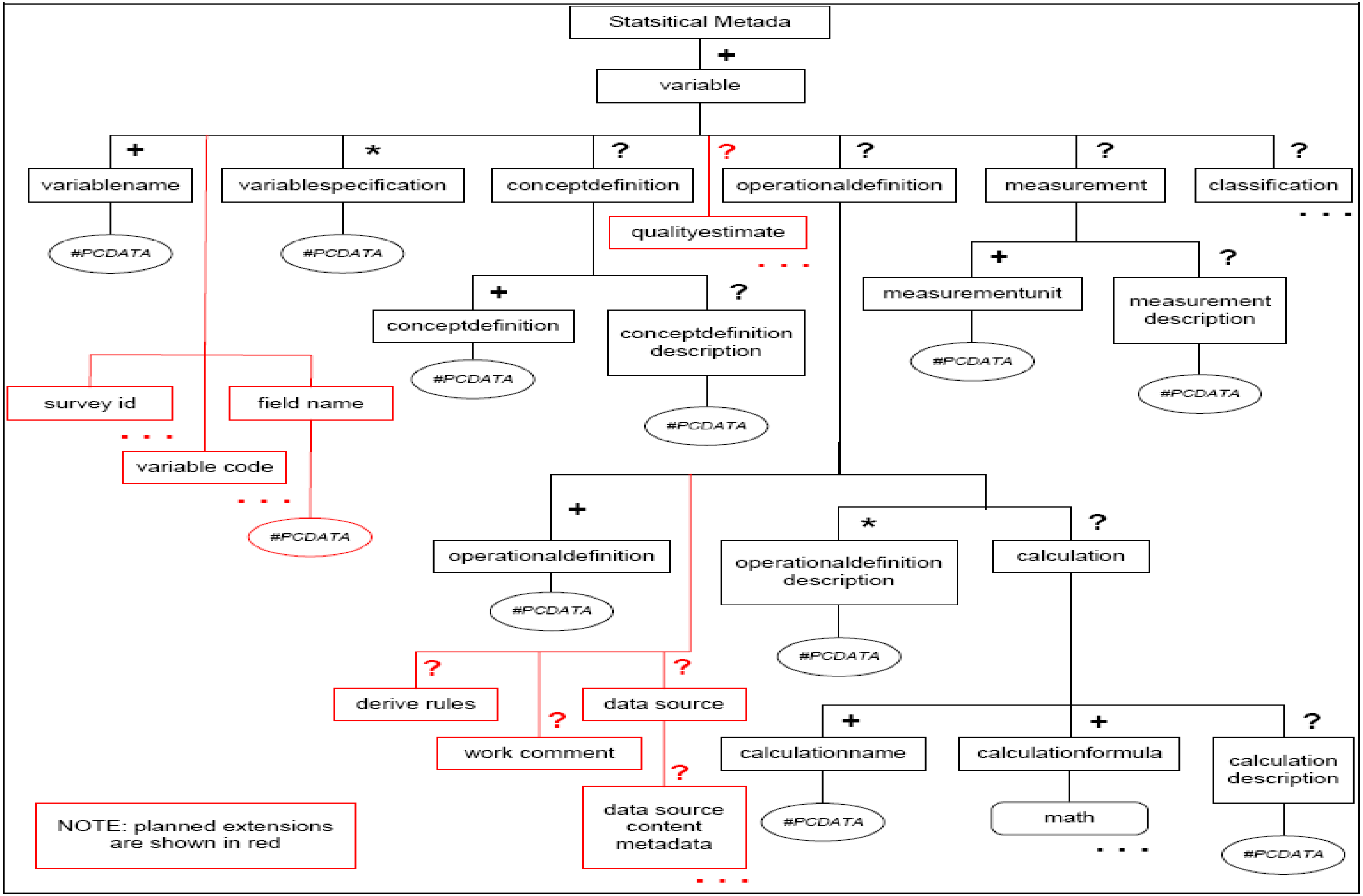
Modular DTD system

- document type definitions
- Standards: CALS, XDF, Dublin Core
- XML: one file – data and metadata

Common Structure of Statistical Information (CoSSI) – parts and entity

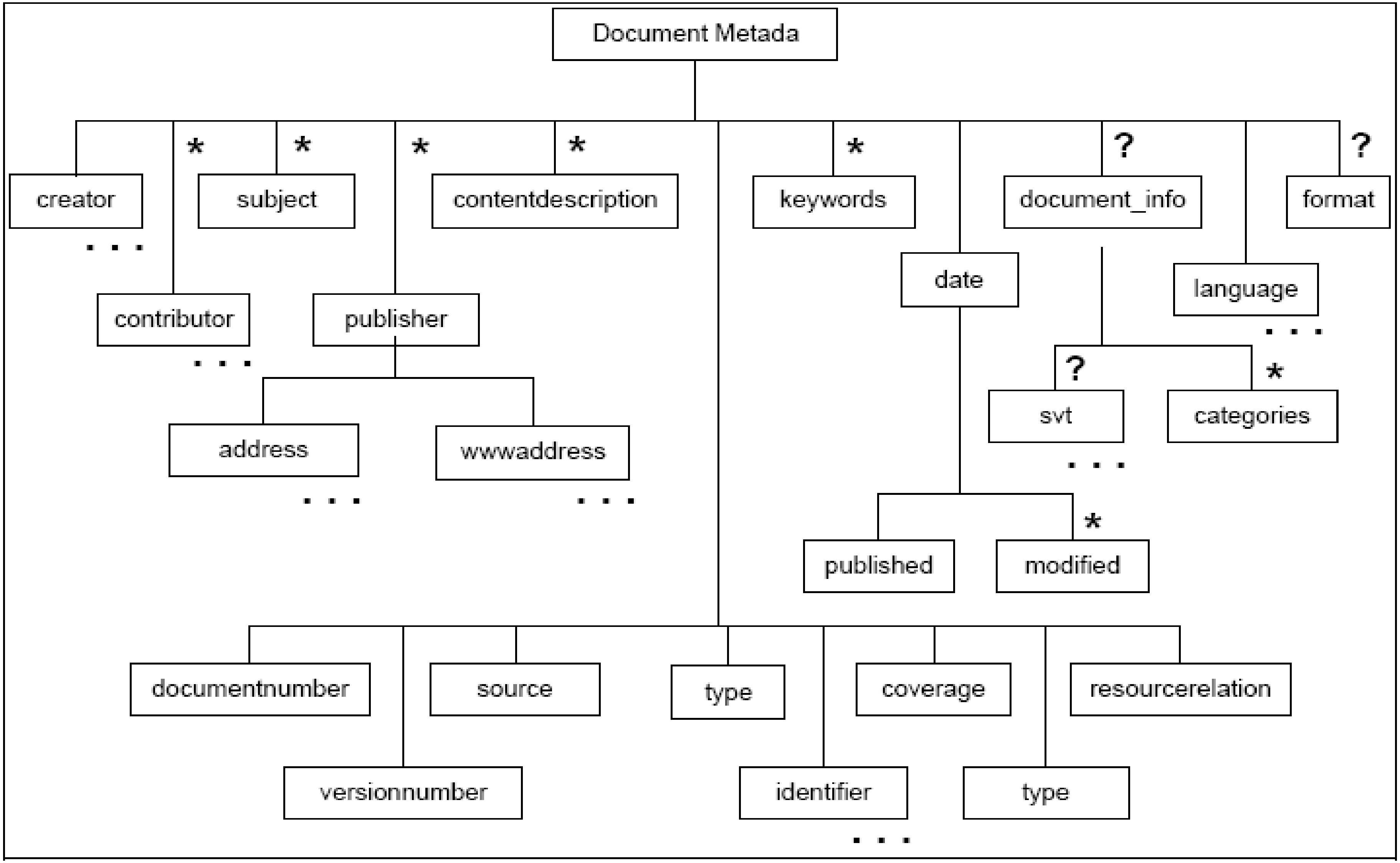


Statistical Metadata – Logical Concept Model

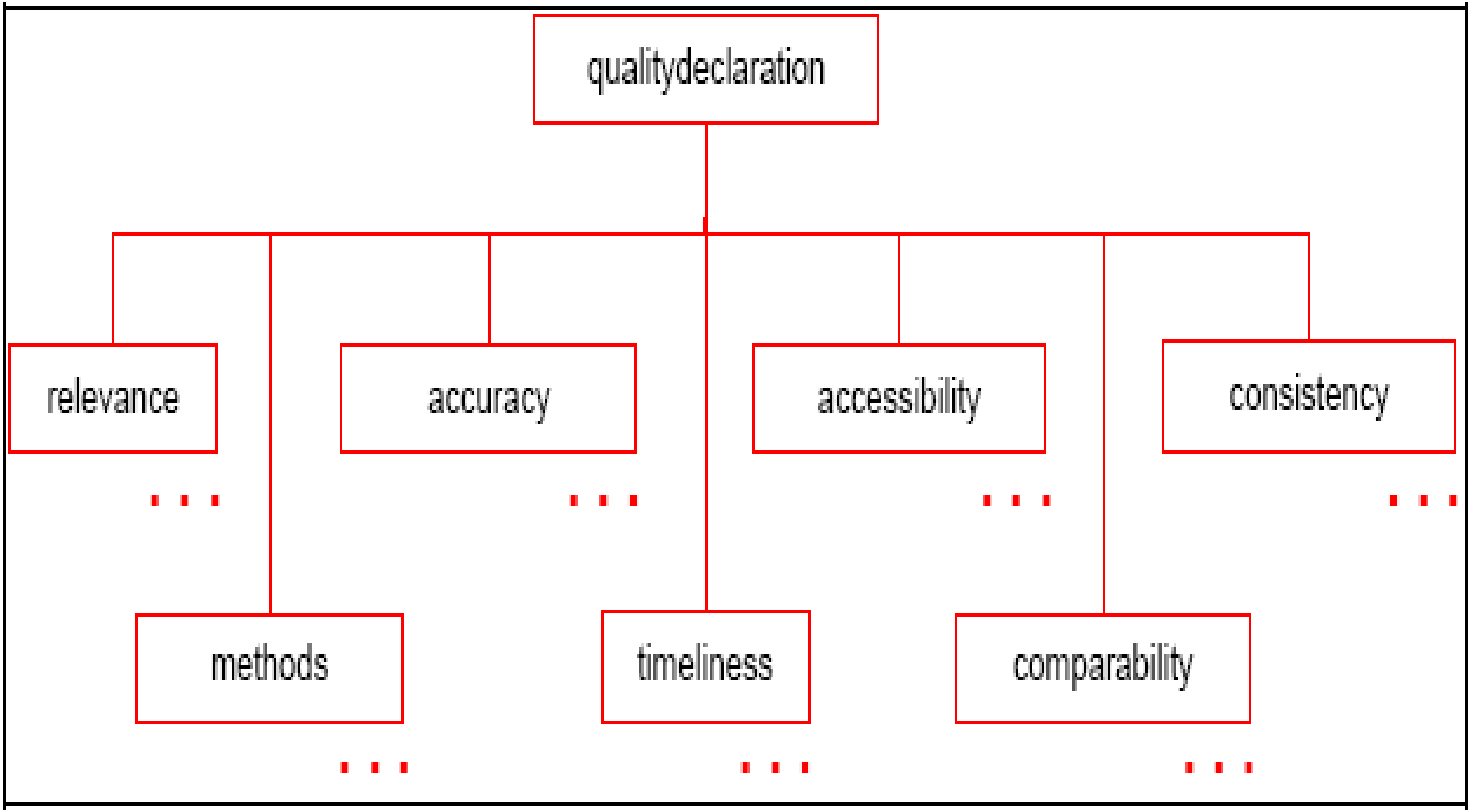


Document Metadata – Logical Concept Model

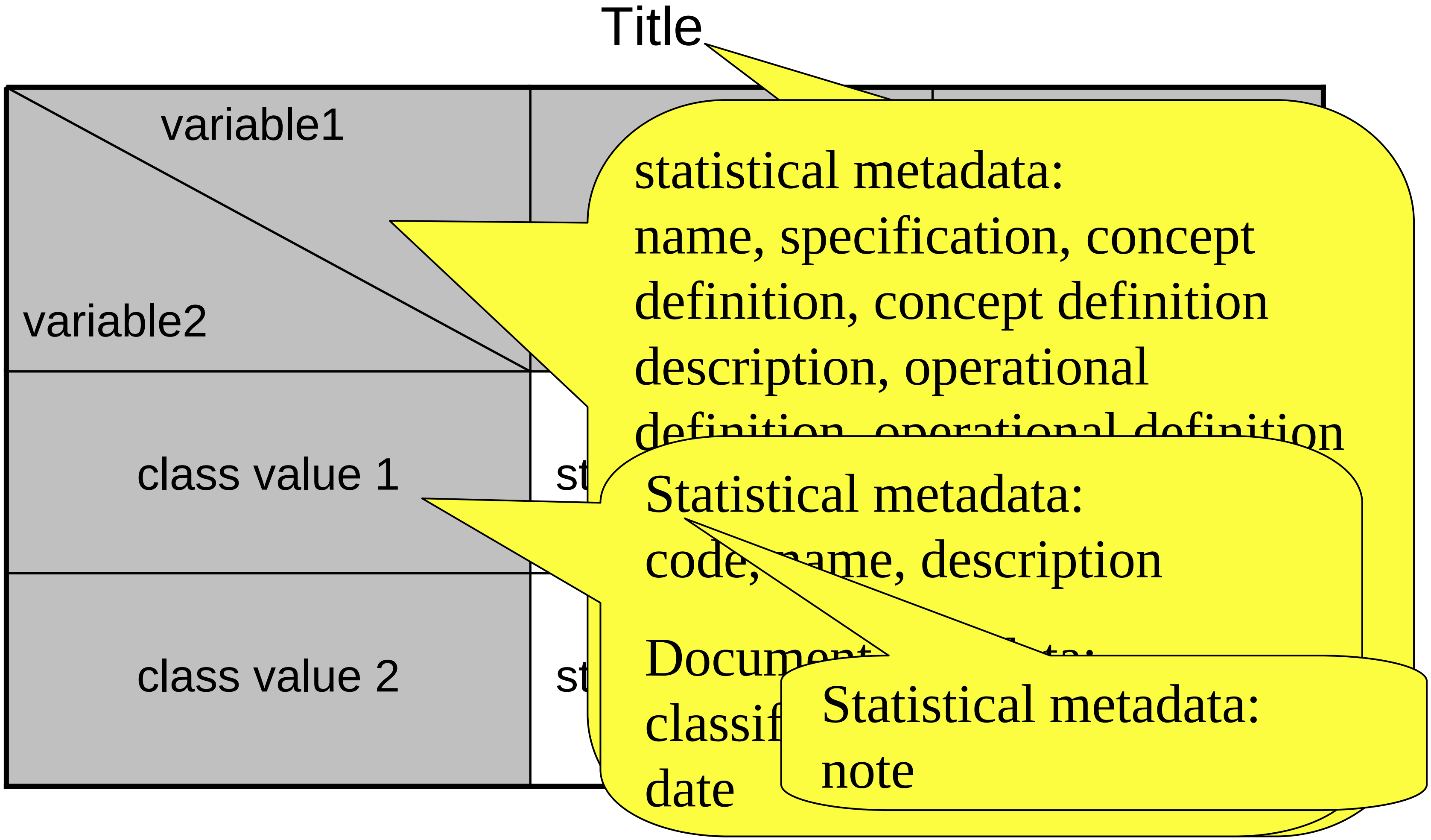
(Dublin Core compatible)



Quality Declaration – Logical Concept Model



Metadata in statistical table according to CoSSI



Title

variable1

variable2

class value 1

st

class value 2

st

statistical metadata:
name, specification, concept definition, concept definition description, operational definition, operational definition

Statistical metadata:
code, name, description

Document
classif
date

Statistical metadata:
note

Some conclusions

Generality of the models

SDMX

- SDMX could be describes as super generic. It is open for any kind of key family definitions and metadata structure definitions.
- To be able to use the model in a rational way, an agreement about the data and metadata structure of statistics is needed among the parties wanting to share information.
- A special XML schema or dtd is needed for each data set and corresponding key family

Some conclusions

Generality of the models

CoSSI

- In CoSSI the elements of metadata are fixed. They are defined in the logical concept model and implemented in the corresponding dtd.
- Just one dtd is needed for each type of organisation of data, e.g. table.dtd, matrix.dtd).
- CoSSI is still open for expansion.
- Not all metadata elements need to be used, if the metadata is not available.

Some conclusions

Entity of data and metadata

- In SDMX structural metadata is attached to data, but reference metadata is in one or more metadata sets outside of data set. Linking of reference metadata to data is made from metadata set, not from data set.
- In CoSSI tables or matrixes and variables in them are directly attached to corresponding metadata.

Some conclusions

Richness and expandability of metadata

- In SDMX structural metadata is somewhat limited in quantity and deepness. Any number of attributes can be added, but they always are attached to the data at the same hierarchical level.

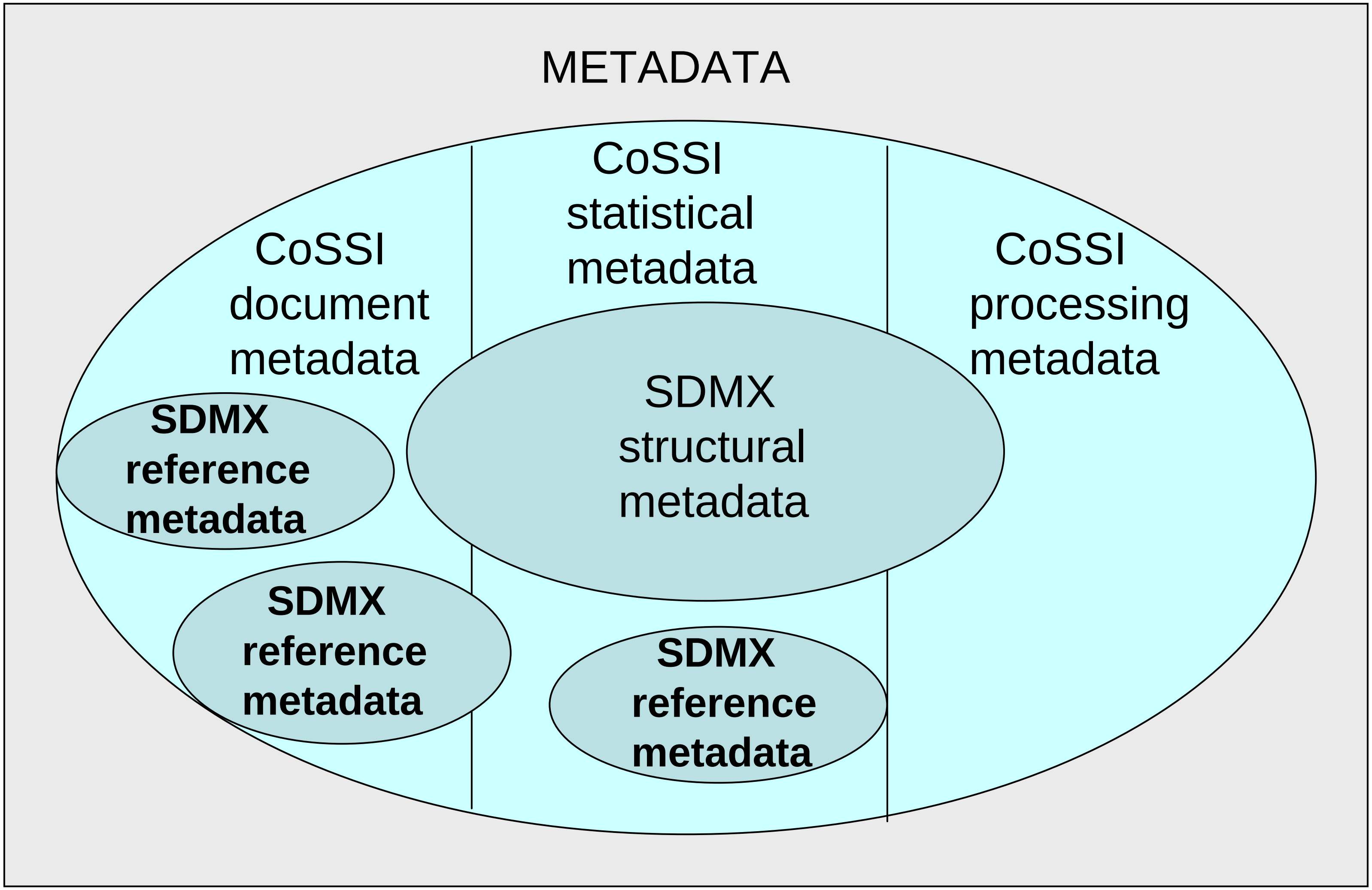
  There is no limit how much reference metadata there is in the separate metadata sets. The necessity to define a metadata structure definition for each metadata set makes it rather heavy and restrictive procedure.

Some conclusions

Richness and expandability of metadata

- In CoSSI the metadata elements are designed to cover the metadata needs as far as possible, but If needed, the model and the dtd can expanded both horizontally and vertically.

Metadata connected to the statistical description of a phenomena



Some conclusions

Transparency of metadata

For the users to be able to evaluate the usefulness of statistical data all the relevant statistical metadata should be obtainable, e.g.

- about how a survey was defined and what asked
- about quality aspects

Some conclusions

Transparency of metadata

- In SDMX reference metadata is again the way to deliver this kind of information. The problem is that, there is no formalized way to attach this kind of metadata and no way to directly point to that metadata set from the presentation of statistical information, e.g. statistical table.
- To CoSSI some formalizations have been or are to be added: quality declaration as an additional module (quality declaration.dtd) and quality estimate as a vertical expansion of statistical metadata module (statmeta.dtd).

Proceedings of Q2006
European Conference on Quality in Survey Statistics

Methodologically Oriented Metadata – Framework for Enhanced Quality Information of Statistics

Heikki Rouhuvirta[1]

1. Introduction

Users of statistical information increasingly demand accurate contentual metadata describing the content of statistical information as well as more detailed information about the quality of statistical data and statistical figures when determining the usability of the statistical figures for their own purposes. From metadata and quality information users of statistical information require concreteness, illustrativeness, interpretability and usability, where both statistical metadata and quality information can be easily searched and viewed in the same context and in the same place from which the users search and receive numerical statistical data into their use.

Broader implementation of metadata and quality information create a set of problems. Some problems are due to the quality indicators of the actual statistical information and their illustration, some others are caused by connecting quality indicator information to metadata information and metadata systems - particularly as the present metadata systems are not specially designed to be applied to that purpose - and yet others by the distribution of quality information, when only a few distribution techniques enable integrated distribution of statistical figures, metadata and quality information.

In searching for usable solutions it is necessary to consider methodological questions connected to the definition of quality indicators and illustration of methodological issues, questions related to the technique and technology for connecting quality information to statistical metadata, and technical and practical questions related to the distribution of extended quality information. It is also necessary to define those information structures of statistical information that enable combination of quality information to numerical data and simultaneous integrated distribution of numerical information and quality-oriented metadata.

As a result of an analysis of the information structure of quality descriptions a definition is given, on the basis of which separate indicator information could be managed systematically and technically. In this connection the requirements set by different metadata systems on quality information are also viewed and it is considered whether the description of statistical metadata of the CoSSI (Common Structure of Statistical Information) model could be extended so that it could also include information (i.e. values) on the indicators of data quality.

[1] Heikki Rouhuvirta, Box 2V, Finland, 000022 Statistics Finland, heikki.rouhuvirta@stat.fi

At the end of this paper there are presented some possibilities for practical implementation on the basis of the experiences gained by Statistics Finland on the processing of statistical metadata. Production and management of quality information emerge as the key questions here. In the scope of the information structure outlined on the basis of preliminary results it is possible to produce such quality indicator information that can be connected as a metadata description to numerical statistical information and distributed to users integrated into numerical information, for example by utilising the XML technology in Internet distribution of statistical information.

2. Conceptual Modelling of Statistical Information

Statistics Finland has been starting to implement a statistical metadata concept based on Statistical Information Model called CoSSI (Common Structure of Statistical Information)[2].

In modelling of statistical information the methodological starting point of the definition of metadata is that in the conceptualisation of the contentual description of statistical information use is made as far as possible of the concepts characteristic of statistical information, the concepts and concept structures it contains and the logic that allows sufficiently multifaceted and complex concept structures for an exhaustive description of the information content[3].

As the used description method of CoSSI allows implementation of complicated structure descriptions, the procedure does not have essentially any factors that would per se somehow force to contract or limit the contentual description.

Results obtained when defining statistical information by setting out from the above points of departure are described in the adjacent figure (Figure 1). On the one hand, statistical information has been defined by using a conceptual analysis, the results from which have been depicted as conceptual models of statistical information and, on the other, an analysis has been made of different forms of organising statistical data and presenting statistical information, which has been used to specify basic models for presenting statistical data. Structural models of data and related data models have been produced for concept models and different forms of organising data, and definitions for these have been implemented in the CoSSI model as multi-level hierarchical (so-called tree-structured) data models[4]. The data models have been documented as XML DTD definitions. The basic method used in the implementation was the ”From model to markup” approach.

[2] Technical description of CoSSI is contained in the definition, see Rouhuvirta and Lehtinen (2003).

[3] More details on the foundations and points of departure for the structuring of statistical information, as well as the requirements set on the system for describing it, see Rouhuvirta 2004a.

[4] On demands imposed on the hierarchy of statistical data, see Rouhuvirta 2004b

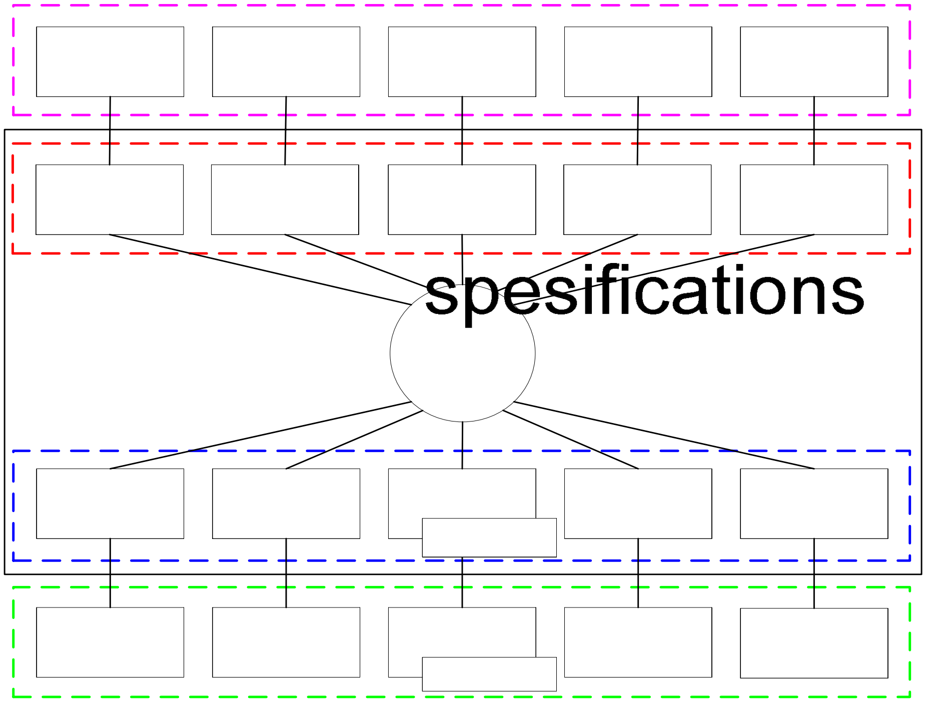


Figure 1. Common Structure of Statistical Information (CoSSI) – parts and entity

Basic models for organising statistical data (statistical data files, tables, etc.) are characterised by the fact that the definitions of different forms of organising statistical data allow presentation of the same information content irrespective of the form. Thus, the scope of the data content is not a criterion on the choice of organisation of data but other factors relating to the processing of the data determine the form of organisation that will serve best the production of statistics and the dissemination of statistical information in each case.

In CoSSI, all information describing the content, defining, etc., of produced data represent metadata. The following typology of metadata has been used as the metadata frame in CoSSI:

(1) Statistical metadata that are content-specific and necessary for the interpretation of numerical statistical data.
(2) Metadata relating to the identification and archiving of datafiles, which form document metadata.
(3) Metadata concerning processing, of which some belong to statistical metadata as statistical and methodological process data and some belong to the process description as technical metadata required by the used applications.
(4) Technical metadata concerning the process, which contain the technical data required by applications and the metadata used or created in the steering of the project.

On the one hand, data obtained from diverse sources for statistical purposes, such as descriptions of data in administrative registers, are based on the own, specific logic of each data source and, on the other, on the availability of data and on the possibility of converting the data into a form where the descriptive information can be electronically attached to the source data and thereby utilised in the production of statistics. Descriptions of source data do not as such form an independent area of their own deviating from statistical metadata, but the descriptive information of the source file is ”included” in one way or another in the statistical metadata as part of the description of the content of the final statistical information[5].

The metadata definitions specifying and describing the contents of statistical information have been technically gathered into the following modules in the CoSSI model:

(1) file metadata (docmeta.dtd)
(2) quality evaluation (qualitydeclaration.dtd)
(3) metadata on statistical information content (statmeta.dtd)
(4) metadata on inquiry (question.dtd)
(5) metadata on register information (e.g. Taxmeta.dtd)
(6) process metadata (e.g. procmeta.dtd).

The defining module can be used combined with each other, or as entities supplementing each other dependent of the situation and data description requirements.

[5] An example of how the descriptive data of an administrative register is handled in the CoSSI framework, see Rouhuvirta, Lehtinen, Karevaara, Laavola, Harlas (2004).
An example of how to attach the description of a register into statistical metadata, see Rouhuvirta (2005).

3. Statistical Metadata

In all situations, the way of processing statistical information is eventually based on the fact that, on the one hand, we have observation units, which in statistics production are also called statistical units. However, on the other hand, besides identification of the observation unit, we also have information produced with different measurement methods on the characteristic of the said unit, which we here refer to as variables for short. This structural characteristic of statistical information (data) is utilised in the CoSSI model to attach and anchor statistical metadata to a variable. Thus, the task of statistical metadata is to describe exhaustively the content and characteristics of the variable for the needs of both producers and users of statistics (see Figure 2).

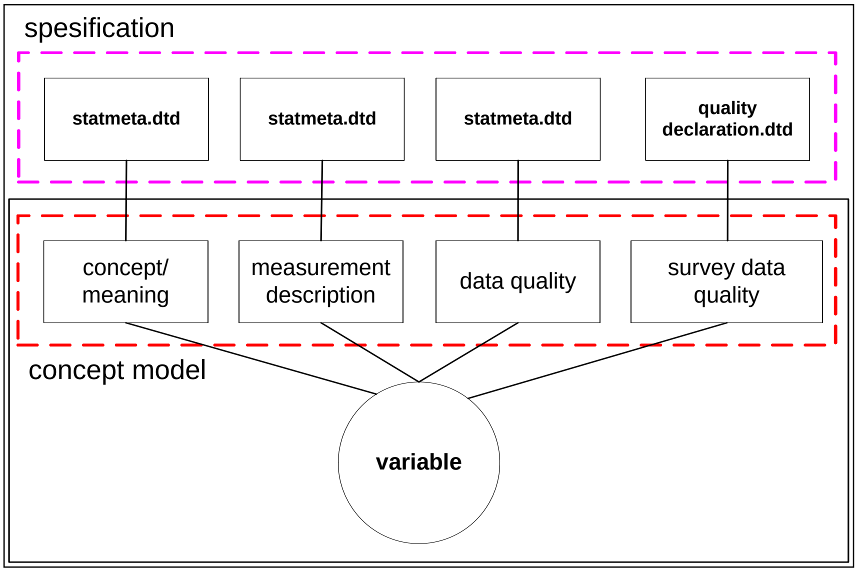


Figure 2. Components of the concepts of statistical metadata and their definitions in CoSSI

Some of the qualitative information on statistical data describe the characteristics of a variable and some the nature of the entire statistical datafile, and the information is not overlapping or summary in all respects. The metadata relating to a datafile cannot be simplified or assigned to the quality descriptions of the variables it contains but require their own overall quality, i.e. a separate examination of the material entity formed in a certain way. Because of the information relating to quality has been divided into two components and assigned, on the one hand, to the variable insofar as it describes the quality of the variable and, on the other, to the datafile insofar at it describes its characteristics. File-specific quality evaluations are presented in quality descriptions appended to the files.

Variable-centredness also brings the practical benefit that the same metadata description can be used unchanged, and even in the same physical format, in different production stages and in all forms of organising statistical data. This way, many adaptations of the syntax or structure of metadata can be avoided, which might otherwise be necessary for productional reasons.

Variable-centredness is a foundation that ensures that metadata are transferred with data to wherever the data are transferred to during statistics production. Irrespective of how the measurement values of measured observation characteristics are handled in different stages of statistics production, the metadata remain the same provided the statistical data themselves are not manipulated in a way that affects their interpretation. Variable-centredness is also a basis whereby descriptions of the contents of administrative and other similar files that are used as sources of statistical data can be combined with the statistical information formed from them in production and in certain cases also with the final description of the statistical information in its dissemination.

The description of statistical information at the unit level is comprised of the documentation describing the data of the statistics, which contains statistical metadata by variable and a quality description that contains general methodological descriptions and quality evaluations relating to the data. The variable-specific descriptions of statistical metadata can be supplemented with application-specific process metadata descriptions, in which the technical information required by the application, such as length of data record or its number or character format, can be attached to the metadata descriptions.

4. Data Models of Metadata

The basic conceptual model of statistical metadata is described as a logical data model in the adjacent figure (see Figure 3). The basic, main concepts of statistical metadata relate to the conceptual defining of the content of a variable and to the defining of the measured characteristics. The meaning of a variable is described in a conceptual definition and the matters relating to the measurement in the operational definition of the variable. If the variable is a summary one or one formed in some other manner, the formula used in its formation can be attached to the description of the metadata.

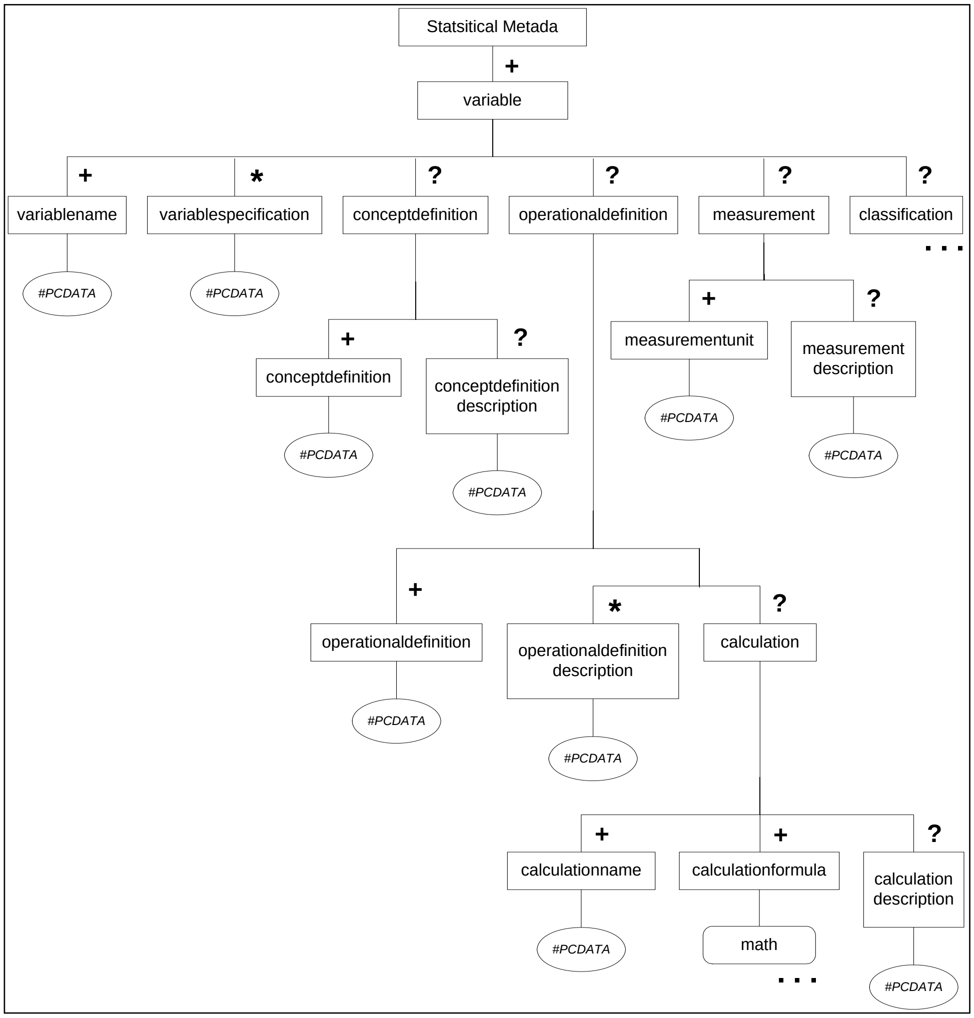


Figure 3. Logical data model of statistical metadata[6]

The main elements of statistical metadata descriptions and their purposes of use are presented in Table 1.

[6] The logical models presented in this context are indicative and detailed, normative data models have been described in the CoSSI definition, see Rouhuvirta and Lehtinen (2003).

Table 1. Basic elements of the concept model of statistical metadata and their purposes of use

Item	Purpose of use
variable name	The name of a variable. Variable name element is used for conceptual naming of variables in natural language. Variable name is not meant to be a code or an abbre-
variable specification	Variable specification is used when the naming of a variable is not enough to describe it. Variable specification element gives a more specific description of the vari-
conceptdefinition	Conceptual definition element contains the conceptual definition of a variable.
conceptdefinition description	Conceptual definition description is used when the information in the conceptual definition element is not enough to clarify the conceptual aspects of a variable.
operationaldefinition	Operational definition element contains a written operational definition of a variable.
operdefinition description	Description element of an operational definition includes a written description of the operational definition. The description is given in natural language. Description of an operational definition is used when the information in an operational definition element is not enough to clarify the operational aspects of a variable.
calculationname	Name of a calculation. If a calculation is given it must be named. It is possible to give the name of the used method without giving the actual calculation formula. The name can be a generic or a case-specified name of the method.
calculationformula	The actual calculation formula is given here in MathML format.
calculation description	Describes a calculation formula. Calculation description is used when the information in the calculation name and calculation formula is not enough to clarify the composing of a variable.
measurementunit	This element names the measurement unit of a variable. The measurement unit is given as a standardised Finnish abbreviation (at Statistics Finland).
measurement description	The description is used to clarify the measurement if the measurement unit is not clear enough.

The central component of statistical metadata is description of the classification of the variables. In the concept model, matters relating to used classifications have been described in two ways. On the one hand, the used classification standard can be identified or, alternatively, the used category values and their importance can also be described.

After initial implementation of the definition of the basic elements of statistical metadata it has become clear that the set of concepts relating to statistical metadata must be enlarged by both conceptualisation of the qualitative data on individual variables and the data that are necessary to steer the processing of statistical data. The steering data assist the communication of the professionals working in production and the realisation of division of responsibilities.

Extensions made at the first stage to the logical data model of statistical data to serve the said purposes are presented in the adjacent figure (see Figure 4)[7].

[7] The extensions shown in this context are due to be implemented into version 2.0 of the CoSSI definition during 2006.

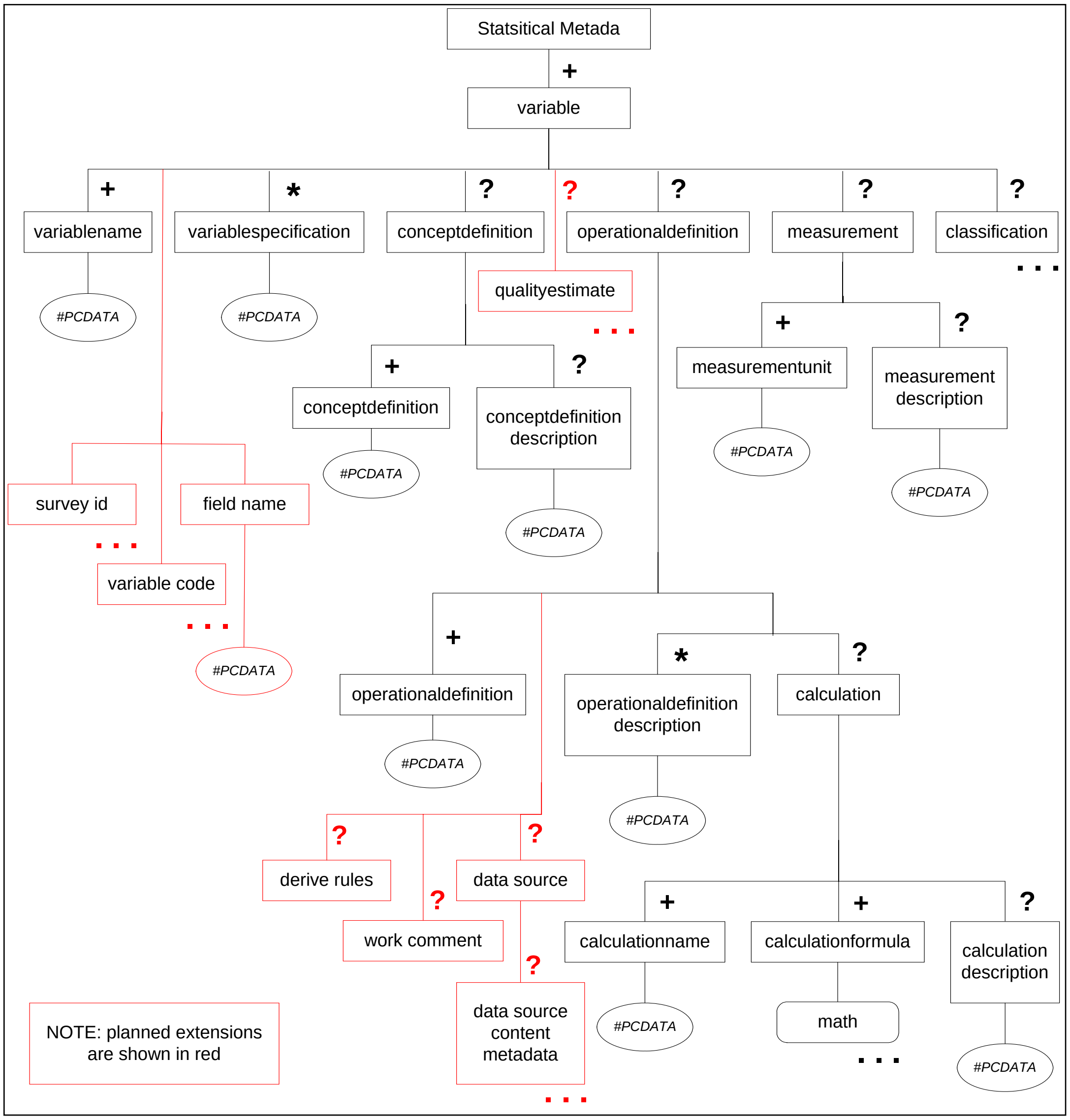


Figure 4. Extensions of the logical data model of statistical metadata

The new elements attached to a variable for production purposes are:

-field name

Technical identifier: Technical identifier enables the use of short names of variables in different information technology environments where the use of long, natural names is not always possible because of technical reasons. In addition to, and separate from this, a variable has a universal identifier (ID).

-survey id

Survey identifier: Statistics departments collect data for more than one statistical survey simultaneously and/or form different ”versions” of a variable. To help identification in production, a variable can be given a survey identifier in the form of a set of characters.

-variable code

Variable code: Variable code facilitates denoting hierarchical variables so that the code can be used in data processing and output. An example would be the long list of income categories in income distribution statistics, in which income variables have been given numerical codes with which different income concepts can be summed up semi-automatically. In this respect, the numerical codes of the income categories could be compared to the sets of codes used in regional classifications. The numerical codes of the variables are included as separate elements in the model because their purpose of use is different from that of the technical or ID codes.

-derive rules

Derive rule: Derive rule is a productional element into which the compilers of statistics can record in their own way in statistical jargon the rule by which a variable is formed.

The intention is that these preliminary expressions will be used to develop an operative definition of a variable, which can be registered as an operational definition of the variable in terms of its accuracy and understandability while at the same time retaining the original derive rule expressed in statistical jargon for productional purposes. The derive rule functions at the same time as a common definition document for the compiler of statistics and for the application developer/programmer.

-work comment

Work comment: Work comment is intended to be used in the supervision of the work of compilers of statistics and as a production check list.

-data source

Source of data on variable: A link or direct reference can be given to a variable to either an external data source, such as an administrative register, or to a question in a question database of data collected in-house.

-metadata on the content of source data

Description of the content of source data: Description of the content of an external data source can be attached here, if the structural description of the data is known, as is the case with taxation data if they have been described according to taxmeta.dtd or if the description of a question relating to survey data collected in-house is in the format specified in question.dtd. A description complying with question.dtd contains the original question text and the values and descriptions of the reply alternatives to it.

The above-described data are primarily meant for production purposes, and it is not the intention to include them in systems for disseminating statistical information. Descriptive information on data sources, which in itself is public information but whose inclusion in the dissemination of statistical information is subject to a separate agreement with the original data producer is, of course, a borderline case. This procedure must be followed irrespective of the fact that the data themselves can be used for statistical purposes. In an ideal case, description of the input data can be included as part of statistical metadata when understandably expressed.

4.1 Quality specifications for Variable

Once the data are completed the purpose of the data evaluating the quality a variable is to help the users of statistical information to assess and use the statistics correctly. For that we need contentual information, methodological information and quality information.

Traditional used quality indicators are for instance:

-Sampling Errors: CV's, variances, confidence intervals…
-Nonsampling Errors
-Coverage: Rate, bias …
-Non-response: Non-response rate, bias, imputation rate, imputation impact, …
-Response: Measurement error, collection mode effects, bias…
-Processing: Keying errors, editing impact…
-Modelling: Variance, bias…

But some new quality measures are under work (Lavallée, 2005) and possible to use:

-Quality Profiles
-combined rates (imputation)

A preliminary data model for data evaluating variable quality is presented in the adjacent figure (see Figure 5).

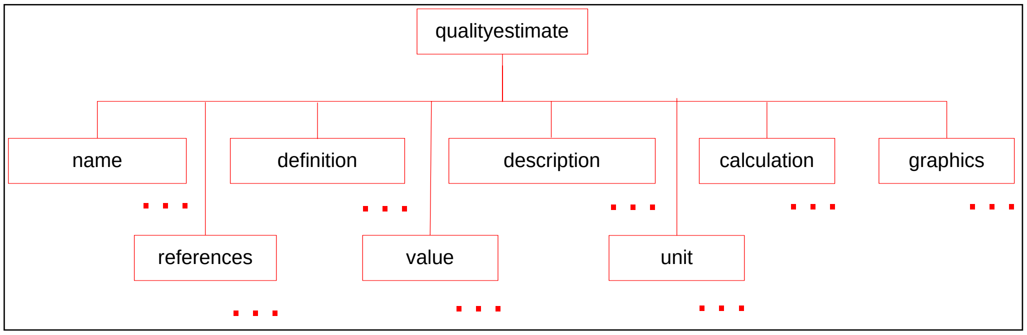


Figure 5. Extension of the logical model of statistical metadata – quality evaluation of variable

The quality evaluation data attached to a variable comprise the following elements:

-quality estimate/quality index/indicator [quality judgement],
-name of parameter,
-method or formula for calculating parameter,
-definition of description of parameter,
-reference to possible methodological source,
-description or interpreting instructions,
-calculation result/value/result value,
-calculation unit of parameter and
-graphic depiction or presentation of result or result value.

4.2 Quality specifications for Statistical Datafile

The quality description of a datafile follows the approved manner for producing quality descriptions at Statistics Finland. The concept model of the planned quality description is presented in Figure 6.

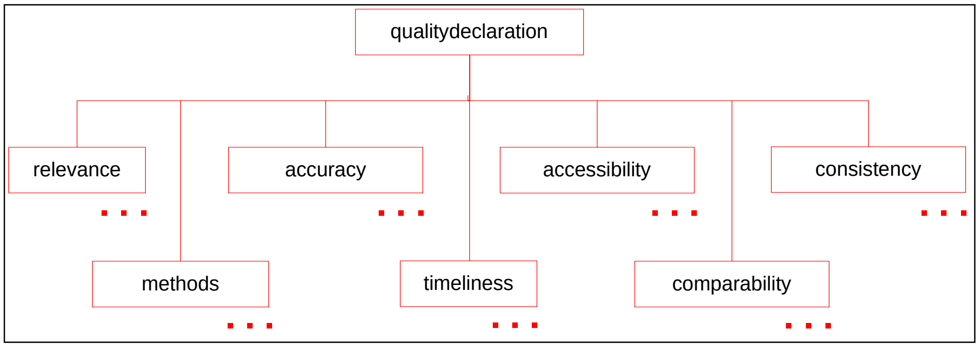


Figure 6. Logical data model of quality description

Purposes of use for the central concept elements of a quality description are presented in Table 2.

Table 2. Basic elements of the concept model of a quality description and their purposes of use

	Item	Purpose of use
1.	relevance	Relevance of statistical data
1.1.		Produce a detailed summary of the product's information content and end use. Identify the phenomenon that this set of statistics is designed to describe and explain its history.
1.2.		Introduce concepts that are important to understanding the statistics, classifications used or object of study, to identifying the data collector and informants.
1.3.		Describe the any acts, decrees and recommendations upon which the statistics are based.
1.4.		Assess the relevance of the statistical information produced in relation to customer needs, and how any changes in the phenomenon concerned have been taken into account in compiling the statistics.
2.	methods	Description of the methods used in statistical surveys
2.1.		Describe methods precisely, e.g.the methods applied, i.e. the population of the statistics, the materials used, the survey design (census survey or sample survey), the sampling design, data collection method, editing, imputation, the use of weighting coefficients in sample surveys and estimation methods required by the final results
2.2.		Justify the methods used and any changes made (including an assessment of the impacts of those changes upon time series).
2.3.		Methods descriptions identify the data sources used in statistics production (also for auxiliary information).
2.4.		Review the whole process of statistical survey.
3.	accuracy	Accuracy of information
3.1.		Demonstrate that the statistics measure what they are supposed to measure.
3.2.		Report on all facts that may have a bearing on the reliability of the statistics. Also mention key uncertainty factors, i.e. possible sampling and non-sampling errors.

.3.3.		Estimate the correspondence between the target population and the population of interest and the quality of the sampling frame used.
3.4.		Describe the main uncertainty factors, i.e. possible sources of error, and assess their impacts on the estimates published: – Sampling errors, – Non-sampling errors: – Coverage error, – Measurement error, – Processing error, – Non-response error.
3.5.		Using the main classifications employed in the statistics, tabulate statistical parameters for the estimates, such as standard deviations that take the sampling design into account, mean square errors (MSE) and parameters estimating the efficiency of the sampling design (deff)
3.6.		Interpret tables produced in 3.5.
4.	timeliness	Timeliness and promptness of the information published
4.1.		Indicate the point of time or period that the statistics describe.
4.2.		Indicate whether the information is preliminary or final.
4.3.		Where necessary examine how time series data have changed over time (e.g. on account of seasonal adjustment).
5.	accessibility	Accessibility and clarity of information
5.1.		For statistics where the data constitute comparable time series, indicate the length of the time series available.
6.	comparability	Comparability of statistics
6.1.		Describe the comparability of the statistics over time and with other materials.
6.2.		Examine changes that have affected comparability and their significance, e. g. in the statistics production process, survey design concepts and classifications.
7.	consistency	Consistency
7.1.		Assess the consistency of the statistics in comparison with other statistics on the same subject. For example, examine differences in their concepts and data collection processes and assess their impacts.

5. Some Conclusions

According to our initial experiences, application of a structured information model to the conceptualisation of statistical metadata has opened new possibilities for exploiting the new technologies developed for easy handling of information in text format. XML can be regarded as representing such technologies. More efficient processing of text-format information facilitates management of richer statistical metadata. This aspect can be exploited equally well in the production of statistics and in the dissemination of statistical information. Example how to connect textual, numerical and graphical quality information to statistical tables is presented in Appendix 1 and example XML files in Appendix 2.

The practical benefits brought by the application of structuring to the management of metadata include the ease with which the data model can be expanded and the extensions can be technically implemented. The editability can be utilised to develop as consistent and all-embracing content specification as possible for statistical metadata.

Besides the scope of its content and its flexibility, as a frame of reference for metadata the CoSSI model examined here differs quite essentially from the metadata systems conventionally used at Statistics Finland in that
-it makes it possible to change from decentralised to centralised management of metadata in which the producer of statistics can control the correctness of the metadata concerning the data material, and which can also be used in the dissemination of statistical information, and
-when receiving numerical statistical data, in the same connection the users also receive the metadata that are essential in their interpretation, and instead of untargeted metadata in separate reference volumes or other similar sources, metadata can be presented immediately adjacent to numerical statistical data.

Examined from the perspective of production, the point of departure in the modelling and organisation of metadata could have been its attachment to the numerical value of data. However, the now implemented attachment of metadata to the variable instead of the data value facilitated the use of a simpler and more informative data model relative to the scope of the data content, and simplified and rationalised the management of data. Attachment of statistical metadata to the structure of statistical data through a variable is technically considerably simpler to implement than linking of metadata to individual data values in an information system, whatever the available production technology.

In practice, the structured model of statistical information represents a real alternative as a frame of reference for statistical metadata, both in respect of its approach and concept defining, and at the moment there appears to be no specific need to change the developed basic solution. Indeed, the needs for further development concern primarily extensions of the data content along the lines described above. Moreover, we are endeavouring to improve the functionality of the technical solutions of statistics production so as to make the use of statistical metadata effortless and easy in different stages of production. These kinds of solutions serving the production of statistics include creation of user interfaces with statistical metadata into such tabulation applications as SAS and SuperStar.

References

Lavallée, P. (2005), “Quality Indicators when Combining Survey Data and Administrative Data How to adjust current indicators in the presence of administrative data?”, paper presented at the Statistics Canada’s 22nd International Methodology Symposium, Methodological Challenges for Future Information Needs, Ottawa, Canada.

Rouhuvirta H. (2004a), “An alternative approach to metadata – CoSSI and modelling of metadata”, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at:

http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

Rouhuvirta, H. (2004b), “Structuring of statistical information and statistical metadata”, Codacmos 2004. Project IST-2001-38636

Rouhuvirta, H. (2005), Conceptual Modelling Of Administrative Register Information And Xml - Taxation Metadata As An Example. UNECE Work Session on Statistical Data Editing, Ottawa 2005. Also available on the Internet at: http://www.unece.org/stats/documents/2005/05/sde/wp.3.e.pdf

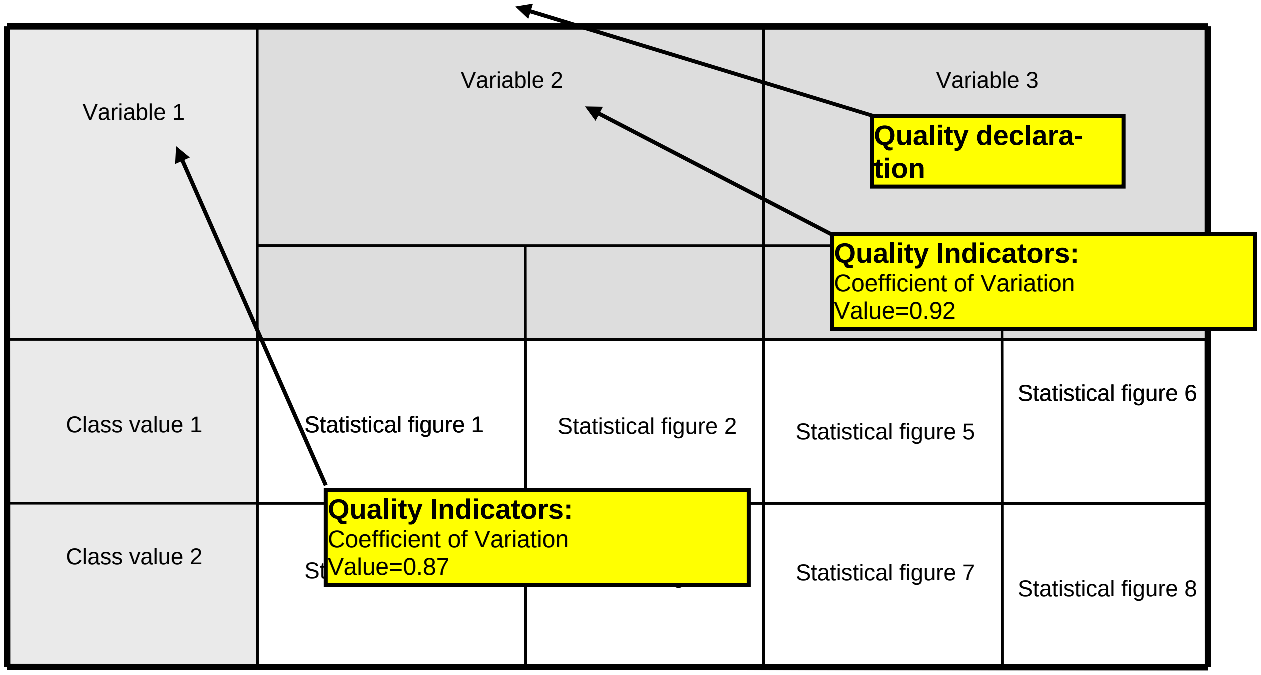
Rouhuvirta, H. and Lehtinen, H. (2003), “Common Structure of Statistical Information (CoSSI) - Definition Descriptions”, 2nd December 2003, Version 0.91, Statistics Finland 2003. Also available on the Internet at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf

Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S. (2004), Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf

Appendix 1.

Example Statistical Tables Including Information on Quality

Table 1. Statistical Metadata in a informative statistical table (I)



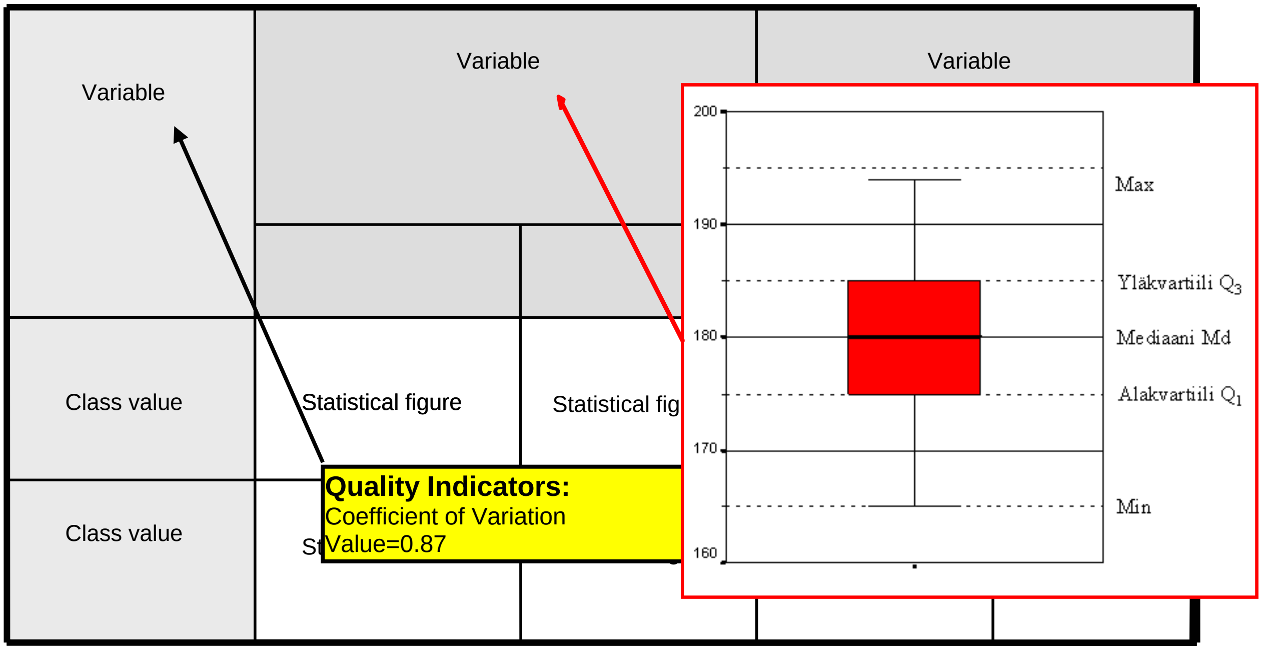
Variable 1	Variable 2		Variable 3	
Class value 1	Statistical figure 1	Statistical figure 2	Statistical figure 5	Statistical figure 6
Class value 2	St		Statistical figure 7	Statistical figure 8

Quality declara-tion

Quality Indicators:
Coefficient of Variation
Value=0.92

Quality Indicators:
Coefficient of Variation
Value=0.87

Table 1. Statistical Metadata in an informative statistical table (II)



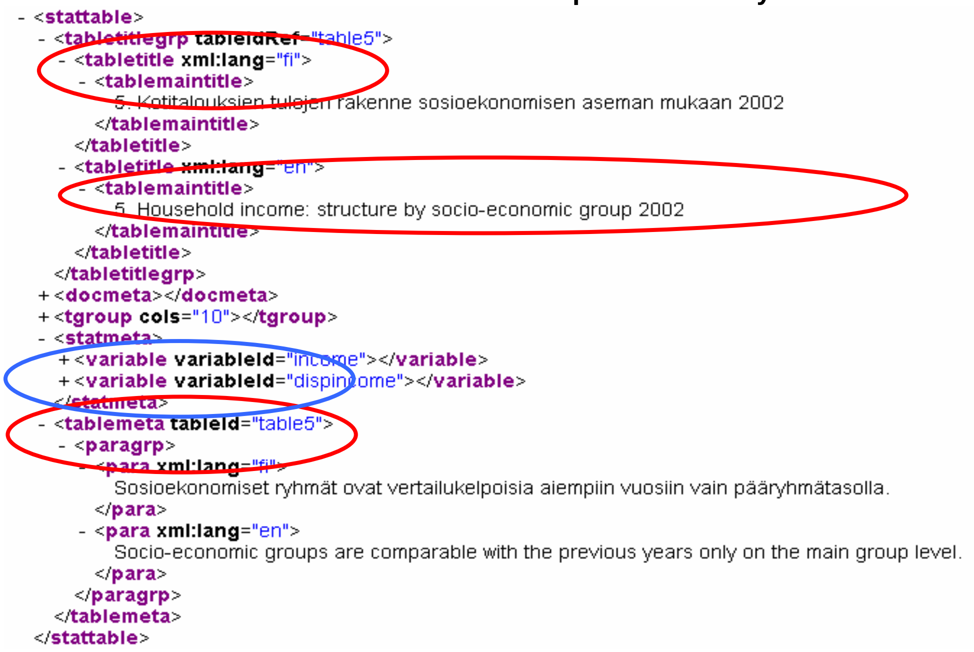
Variable	Variable		Variable	
Class value	Statistical figure	Statistical fig		
Class value	St			

Quality Indicators:
Coefficient of Variation
Value=0.87

Appendix 2.

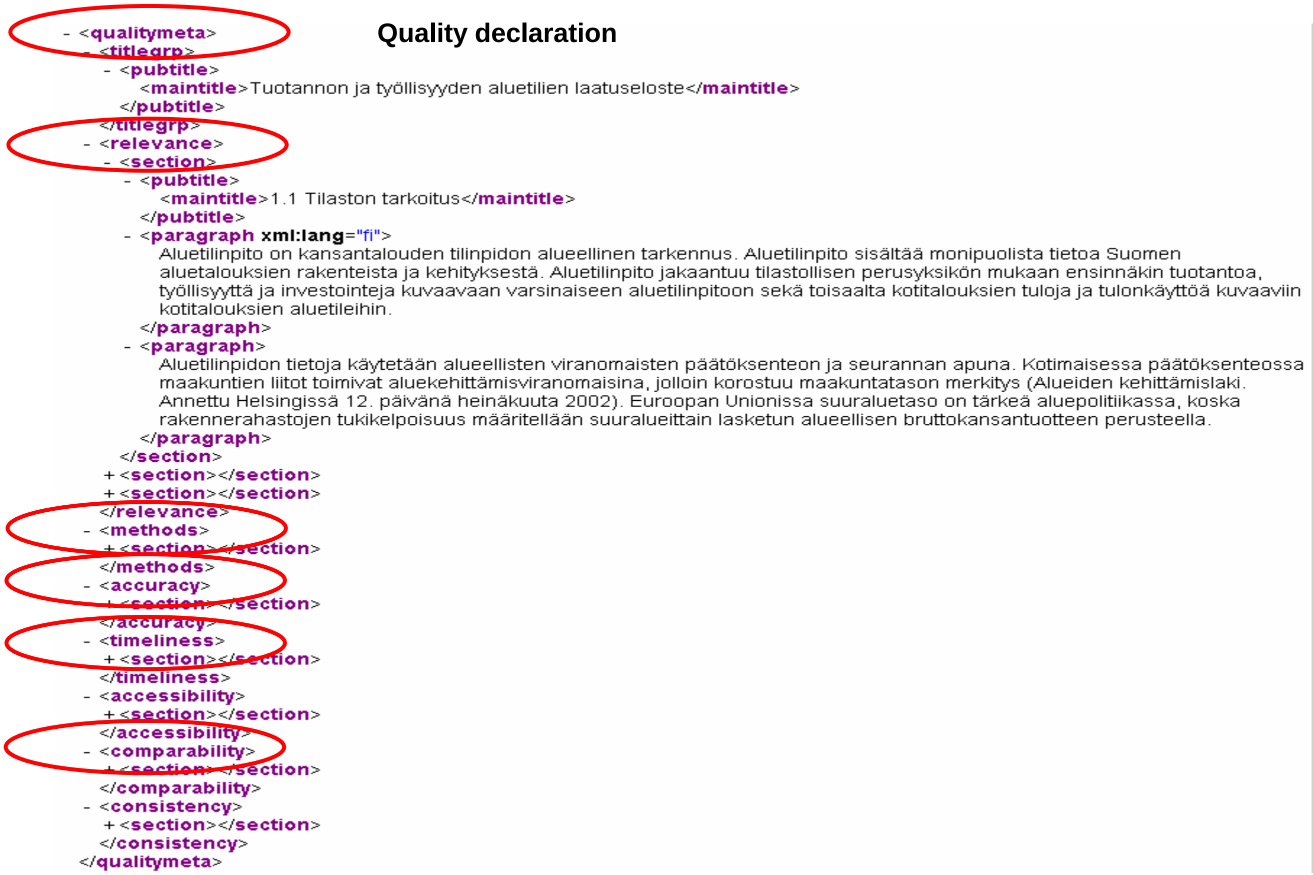
Example Statistical Tables XML Files Including Information on Quality

Statistical Table – parts and entity



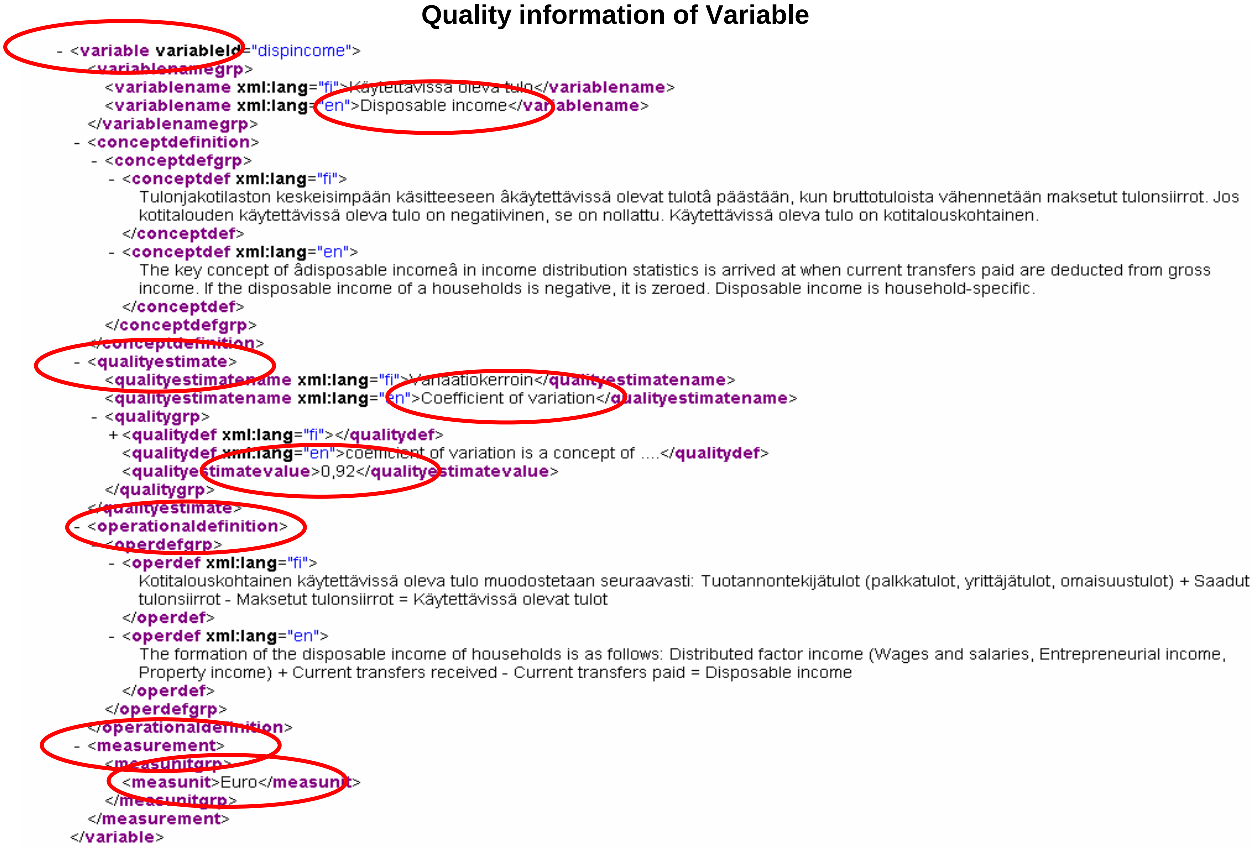
```
- <stattable>
  - <tabletitlegrp tableIdRef="table5">
    - <tabletitle xml:lang="fi">
      - <tablemaintitle>
          5. Kotitalouksien tulojen rakenne sosioekonomisen aseman mukaan 2002
        </tablemaintitle>
      </tabletitle>
    - <tabletitle xml:lang="en">
      - <tablemaintitle>
          5. Household income: structure by socio-economic group 2002
        </tablemaintitle>
      </tabletitle>
    </tabletitlegrp>
  + <docmeta></docmeta>
  + <tgroup cols="10"></tgroup>
  - <statmeta>
    + <variable variableId="income"></variable>
    + <variable variableId="dispincome"></variable>
    </statmeta>
  - <tablemeta tableId="table5">
    - <paragrp>
      - <para xml:lang="fi">
          Sosioekonomiset ryhmät ovat vertailukelpoisia aiempiin vuosiin vain pääryhmätasolla.
        </para>
      - <para xml:lang="en">
          Socio-economic groups are comparable with the previous years only on the main group level.
        </para>
      </paragrp>
    </tablemeta>
  </stattable>
```

Quality declaration



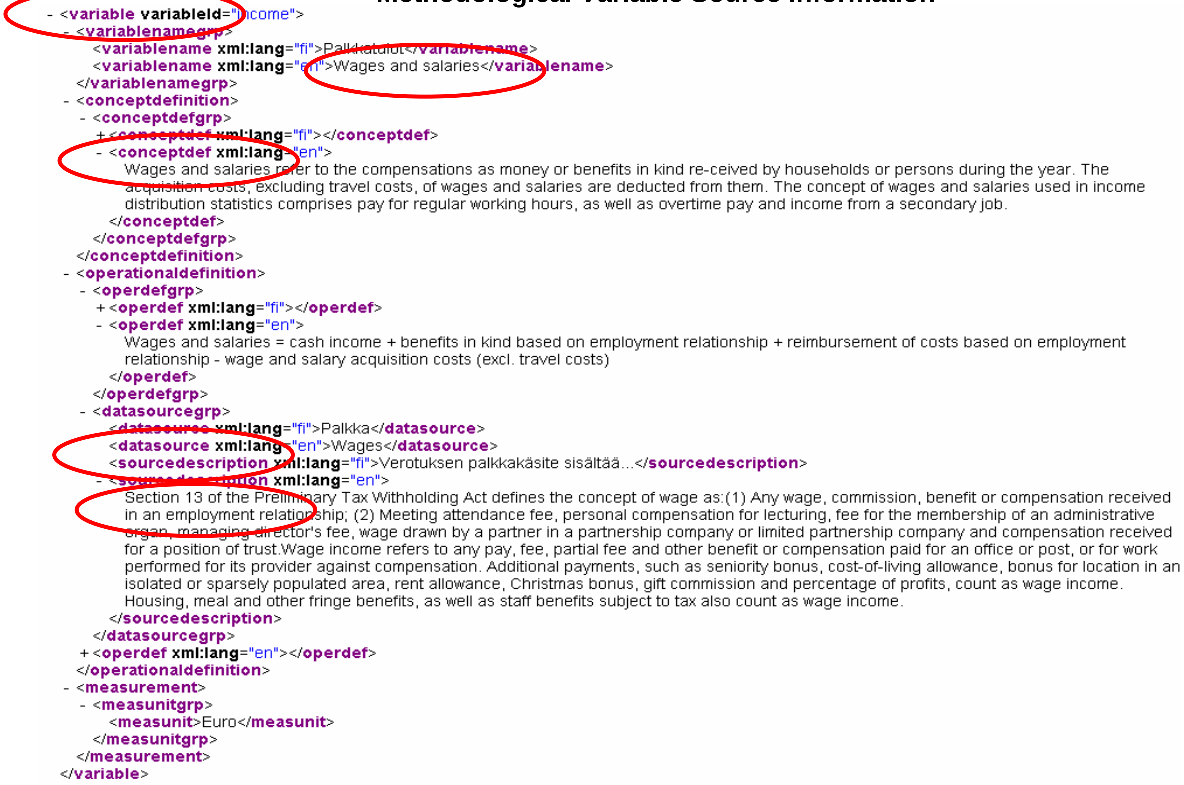
```
- <qualitymeta>
  - <titlegrp>
    - <pubtitle>
        <maintitle>Tuotannon ja työllisyyden aluetilien laatuseloste</maintitle>
      </pubtitle>
    </titlegrp>
  - <relevance>
    - <section>
      - <pubtitle>
          <maintitle>1.1 Tilaston tarkoitus</maintitle>
        </pubtitle>
      - <paragraph xml:lang="fi">
          Aluetilinpito on kansantalouden tilinpidon alueellinen tarkennus. Aluetilinpito sisältää monipuolista tietoa Suomen
          aluetalouksien rakenteista ja kehityksestä. Aluetilinpito jakaantuu tilastollisen perusyksikön mukaan ensinnäkin tuotantoa,
          työllisyyttä ja investointeja kuvaavaan varsinaiseen aluetilinpitoon sekä toisaalta kotitalouksien tuloja ja tulonkäyttöä kuvaaviin
          kotitalouksien aluetileihin.
        </paragraph>
      - <paragraph>
          Aluetilinpidon tietoja käytetään alueellisten viranomaisten päätöksenteon ja seurannan apuna. Kotimaisessa päätöksenteossa
          maakuntien liitot toimivat aluekehittämisviranomaisina, jolloin korostuu maakuntatason merkitys (Alueiden kehittämislaki.
          Annettu Helsingissä 12. päivänä heinäkuuta 2002). Euroopan Unionissa suuraluetaso on tärkeä aluepolitiikassa, koska
          rakennerahastojen tukikelpoisuus määritellään suuralueittain lasketun alueellisen bruttokansantuotteen perusteella.
        </paragraph>
      </section>
    + <section></section>
    + <section></section>
    </relevance>
  - <methods>
    + <section></section>
    </methods>
  - <accuracy>
    + <section></section>
    </accuracy>
  - <timeliness>
    + <section></section>
    </timeliness>
  - <accessibility>
    + <section></section>
    </accessibility>
  - <comparability>
    + <section></section>
    </comparability>
  - <consistency>
    + <section></section>
    </consistency>
  </qualitymeta>
```

Quality information of Variable



```
- <variable variableId="dispincome">
  <variablenamegrp>
    <variablename xml:lang="fi">Käytettävissä oleva tulo</variablename>
    <variablename xml:lang="en">Disposable income</variablename>
  </variablenamegrp>
- <conceptdefinition>
  - <conceptdefgrp>
    - <conceptdef xml:lang="fi">
        Tulonjakotilaston keskeisimpään käsitteeseen âkäytettävissä olevat tulotâ päästään, kun bruttotuloista vähennetään maksetut tulonsiirrot. Jos
        kotitalouden käytettävissä oleva tulo on negatiivinen, se on nollattu. Käytettävissä oleva tulo on kotitalouskohtainen.
      </conceptdef>
    - <conceptdef xml:lang="en">
        The key concept of âdisposable incomeâ in income distribution statistics is arrived at when current transfers paid are deducted from gross
        income. If the disposable income of a households is negative, it is zeroed. Disposable income is household-specific.
      </conceptdef>
    </conceptdefgrp>
  </conceptdefinition>
- <qualityestimate>
    <qualityestimatename xml:lang="fi">Variaatiokerroin</qualityestimatename>
    <qualityestimatename xml:lang="en">Coefficient of variation</qualityestimatename>
  - <qualitygrp>
    + <qualitydef xml:lang="fi"></qualitydef>
      <qualitydef xml:lang="en">coefficient of variation is a concept of ....</qualitydef>
      <qualityestimatevalue>0,92</qualityestimatevalue>
    </qualitygrp>
  </qualityestimate>
- <operationaldefinition>
  - <operdefgrp>
    - <operdef xml:lang="fi">
        Kotitalouskohtainen käytettävissä oleva tulo muodostetaan seuraavasti:  Tuotannontekijätulot (palkkatulot, yrittäjätulot, omaisuustulot) + Saadut
        tulonsiirrot - Maksetut tulonsiirrot = Käytettävissä olevat tulot
      </operdef>
    - <operdef xml:lang="en">
        The formation of the disposable income of households is as follows: Distributed factor income (Wages and salaries, Entrepreneurial income,
        Property income) + Current transfers received - Current transfers paid = Disposable income
      </operdef>
    </operdefgrp>
  </operationaldefinition>
- <measurement>
    <measunitgrp>
      <measunit>Euro</measunit>
    </measunitgrp>
  </measurement>
</variable>
```

Methodological Variable Source Information



```
- <variable variableId="income">
  - <variablenamegrp>
      <variablename xml:lang="fi">Palkkatulot</variablename>
      <variablename xml:lang="en">Wages and salaries</variablename>
    </variablenamegrp>
  - <conceptdefinition>
    - <conceptdefgrp>
      + <conceptdef xml:lang="fi"></conceptdef>
      - <conceptdef xml:lang="en">
          Wages and salaries refer to the compensations as money or benefits in kind re-ceived by households or persons during the year. The
          acquisition costs, excluding travel costs, of wages and salaries are deducted from them. The concept of wages and salaries used in income
          distribution statistics comprises pay for regular working hours, as well as overtime pay and income from a secondary job.
        </conceptdef>
      </conceptdefgrp>
    </conceptdefinition>
  - <operationaldefinition>
    - <operdefgrp>
      + <operdef xml:lang="fi"></operdef>
      - <operdef xml:lang="en">
          Wages and salaries = cash income + benefits in kind based on employment relationship + reimbursement of costs based on employment
          relationship - wage and salary acquisition costs (excl. travel costs)
        </operdef>
      </operdefgrp>
    - <datasourcegrp>
        <datasource xml:lang="fi">Palkka</datasource>
        <datasource xml:lang="en">Wages</datasource>
        <sourcedescription xml:lang="fi">Verotuksen palkkakäsite sisältää...</sourcedescription>
      - <sourcedescription xml:lang="en">
          Section 13 of the Preliminary Tax Withholding Act defines the concept of wage as:(1) Any wage, commission, benefit or compensation received
          in an employment relationship; (2) Meeting attendance fee, personal compensation for lecturing, fee for the membership of an administrative
          organ, managing director's fee, wage drawn by a partner in a partnership company or limited partnership company and compensation received
          for a position of trust.Wage income refers to any pay, fee, partial fee and other benefit or compensation paid for an office or post, or for work
          performed for its provider against compensation. Additional payments, such as seniority bonus, cost-of-living allowance, bonus for location in an
          isolated or sparsely populated area, rent allowance, Christmas bonus, gift commission and percentage of profits, count as wage income.
          Housing, meal and other fringe benefits, as well as staff benefits subject to tax also count as wage income.
        </sourcedescription>
      </datasourcegrp>
    + <operdef xml:lang="en"></operdef>
    </operationaldefinition>
  - <measurement>
    - <measunitgrp>
        <measunit>Euro</measunit>
      </measunitgrp>
    </measurement>
  </variable>
```

Statistics Finland

Methodologically Oriented Metadata - Framework for Enhanced Quality Information of Statistics

Heikki Rouhuvirta, Statistical Methodology R&D

heikki.rouhuvirta@stat.fi

Cardiff, 26-27 April 2006

The typology of metadata

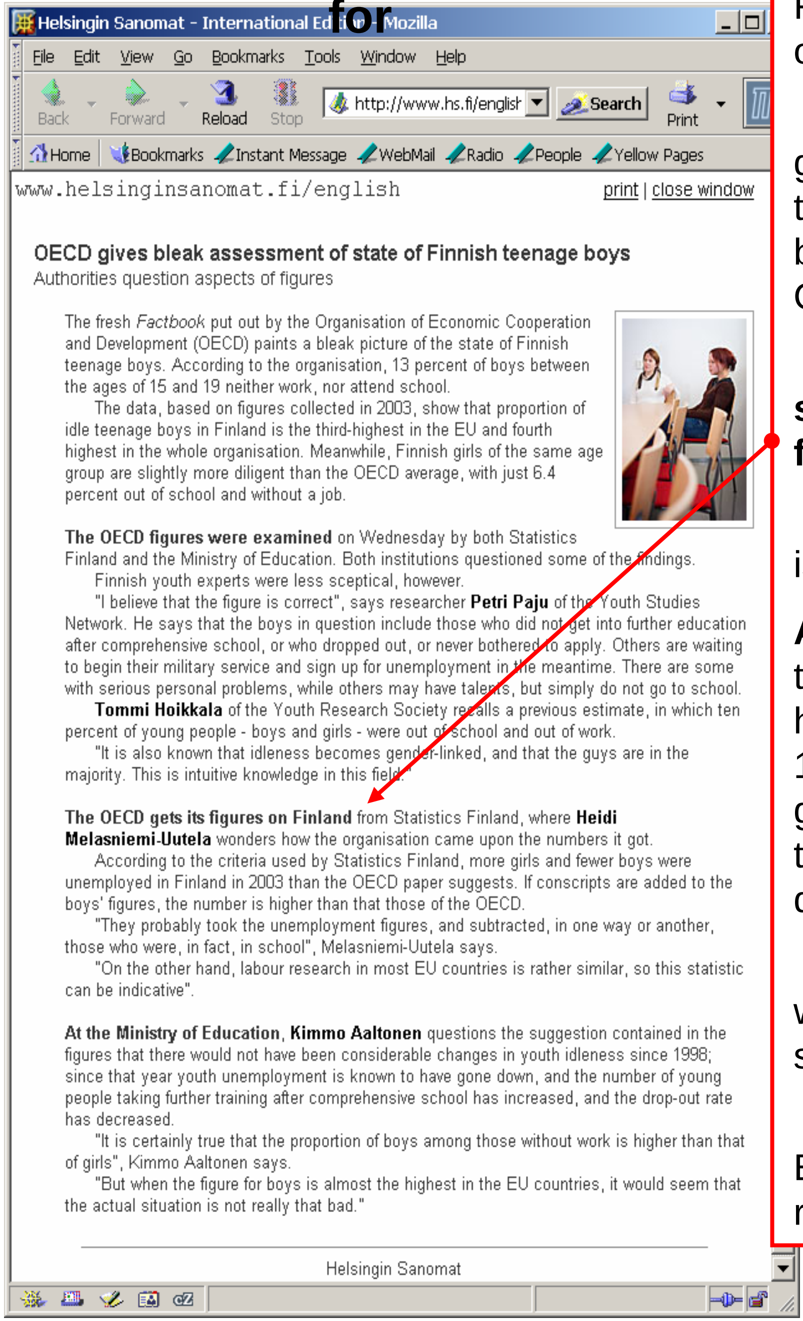
- (1) Statistical metadata that are content-specific and necessary for the interpretation of numerical statistical data.
- (2) Metadata relating to the identification and archiving of datafiles, which form document metadata.
- (3) Metadata concerning processing, of which some belong to statistical metadata as statistical and methodological process data and some belong to the process description as technical metadata required by the used applications.
- (4) Technical metadata concerning the process, which contain the technical data required by applications and the metadata used or created in the steering of the project.

Statistical metadata

- is the most important part of metadata

Statistical Metadata refers to all the data vital for interpreting numerical statistical information.

=> Statistical metadata – what for


Helsingin Sanomat - International Edition - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.hs.fi/english
Search
Print
Home Bookmarks Instant Message WebMail Radio People Yellow Pages
print | close window

www.helsinginsanomat.fi/english

OECD gives bleak assessment of state of Finnish teenage boys

Authorities question aspects of figures

The fresh *Factbook* put out by the Organisation of Economic Cooperation and Development (OECD) paints a bleak picture of the state of Finnish teenage boys. According to the organisation, 13 percent of boys between the ages of 15 and 19 neither work, nor attend school.

The data, based on figures collected in 2003, show that proportion of idle teenage boys in Finland is the third-highest in the EU and fourth highest in the whole organisation. Meanwhile, Finnish girls of the same age group are slightly more diligent than the OECD average, with just 6.4 percent out of school and without a job.

The OECD figures were examined on Wednesday by both Statistics Finland and the Ministry of Education. Both institutions questioned some of the findings.

Finnish youth experts were less sceptical, however.

"I believe that the figure is correct", says researcher **Petri Paju** of the Youth Studies Network. He says that the boys in question include those who did not get into further education after comprehensive school, or who dropped out, or never bothered to apply. Others are waiting to begin their military service and sign up for unemployment in the meantime. There are some with serious personal problems, while others may have talents, but simply do not go to school.

Tommi Hoikkala of the Youth Research Society recalls a previous estimate, in which ten percent of young people - boys and girls - were out of school and out of work.

"It is also known that idleness becomes gender-linked, and that the guys are in the majority. This is intuitive knowledge in this field."

The OECD gets its figures on Finland from Statistics Finland, where **Heidi Melasniemi-Uutela** wonders how the organisation came upon the numbers it got.

According to the criteria used by Statistics Finland, more girls and fewer boys were unemployed in Finland in 2003 than the OECD paper suggests. If conscripts are added to the boys' figures, the number is higher than that those of the OECD.

"They probably took the unemployment figures, and subtracted, in one way or another, those who were, in fact, in school", Melasniemi-Uutela says.

"On the other hand, labour research in most EU countries is rather similar, so this statistic can be indicative".

At the Ministry of Education, **Kimmo Aaltonen** questions the suggestion contained in the figures that there would not have been considerable changes in youth idleness since 1998; since that year youth unemployment is known to have gone down, and the number of young people taking further training after comprehensive school has increased, and the drop-out rate has decreased.

"It is certainly true that the proportion of boys among those without work is higher than that of girls", Kimmo Aaltonen says.

"But when the figure for boys is almost the highest in the EU countries, it would seem that the actual situation is not really that bad."

Helsingin Sanomat

The OECD gets its figures on Finland from Statistics Finland, where **Heidi Melasniemi-Uutela** wonders how the organisation came upon the numbers it got.

According to the criteria used by Statistics Finland, more girls and fewer boys were unemployed in Finland in 2003 than the OECD paper suggests. If conscripts are added to the boys' figures, the number is higher than that those of the OECD.

"They probably took the unemployment figures, and subtracted, in one way or another, those who were, in fact, in school", Melasniemi-Uutela says.

"On the other hand, labour research in most EU countries is rather similar, so this statistic can be indicative".

At the Ministry of Education, Kimmo Aaltonen questions the suggestion contained in the figures that there would not have been considerable changes in youth idleness since 1998; since that year youth unemployment is known to have gone down, and the number of young people taking further training after comprehensive school has increased, and the drop-out rate has decreased.

"It is certainly true that the proportion of boys among those without work is higher than that of girls", Kimmo Aaltonen says.

"But when the figure for boys is almost the highest in the EU countries, it would seem that the actual situation is not really that bad."

=> And what should it be like …

Statistical Metadata

- Contentual information
- Methodological information
- Quality information

1. Traditional quality indicators

- Sampling Errors: CV’s, variances, confidence intervals…
- Nonsampling Errors
  - Coverage: Rate, bias …
  - Nonresponse: Non-response rate, bias, imputation rate, imputation impact, …
  - Response: Measurement error, collection mode effects, bias…
  - Processing: Keying errors, editing impact…
  - Modelling: Variance, bias…

2. Possible new quality measures

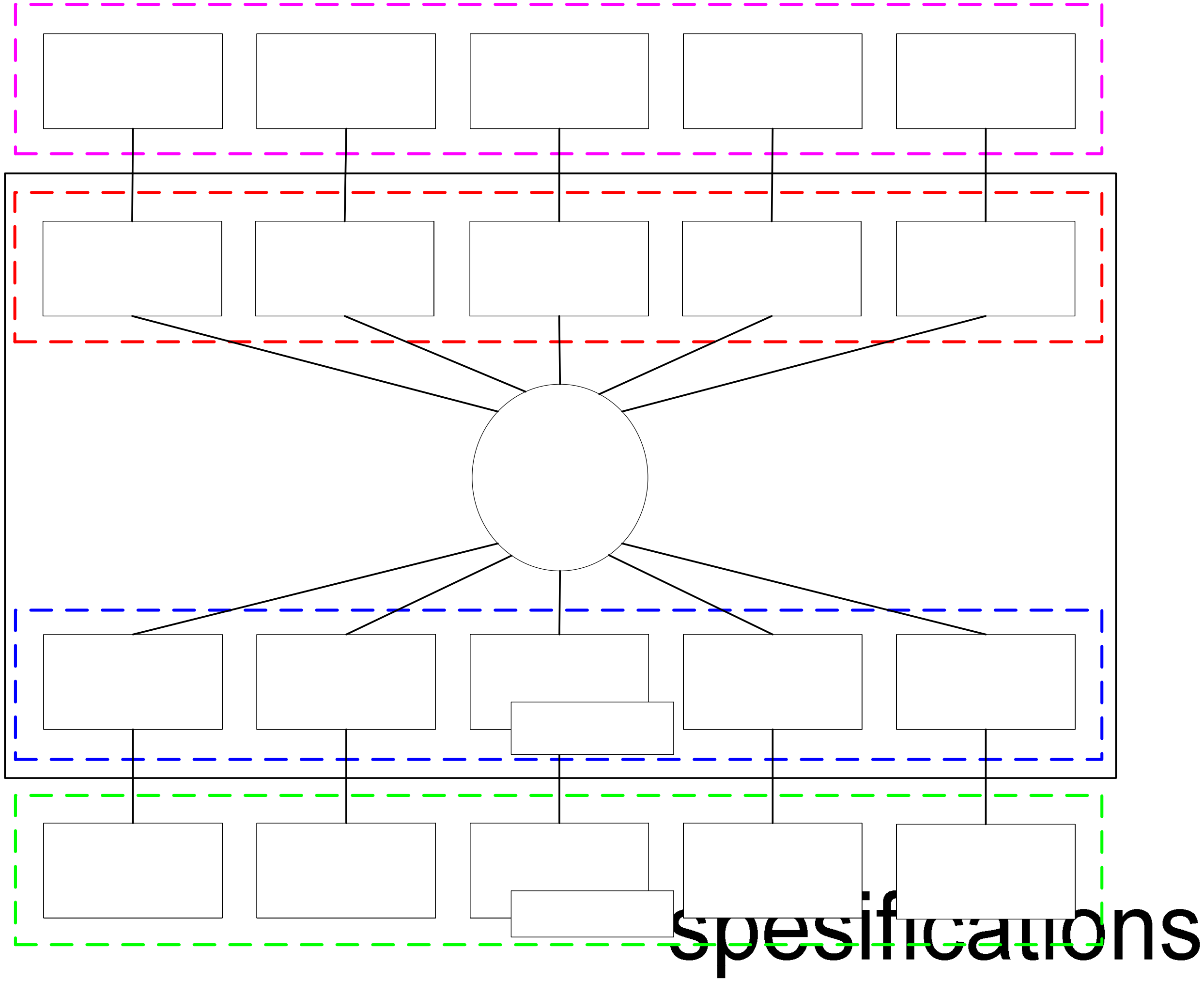
- Quality Profiles (Lavallée, 2005)
  - combined rates (imputation)

How information on quality can be attached to numerical statistics

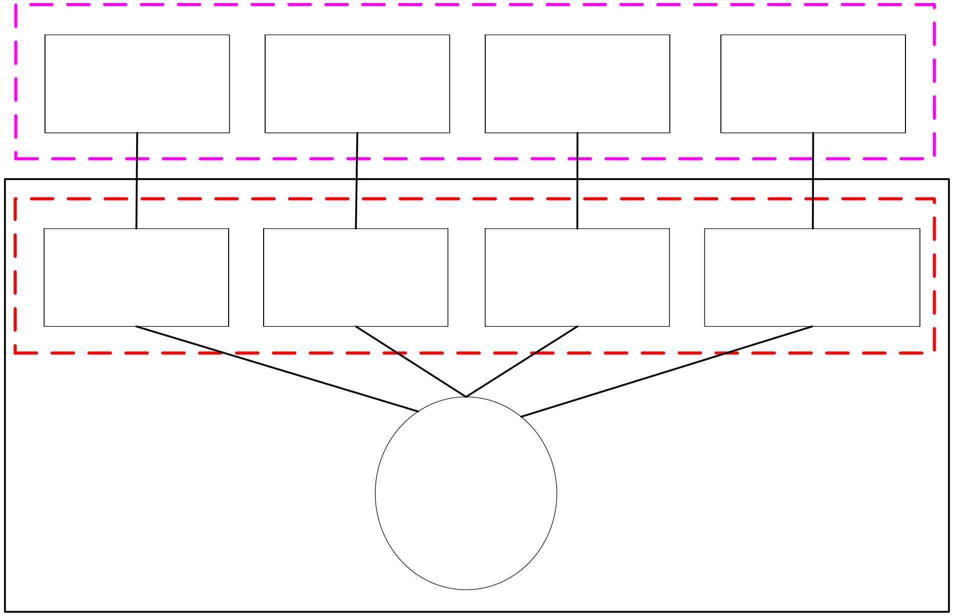
Common Structure of Statistical Information – CoSSI

CoSSI: http://www.stat.fi/cossi

Common Structure of Statistical Information (CoSSI) – parts and entity

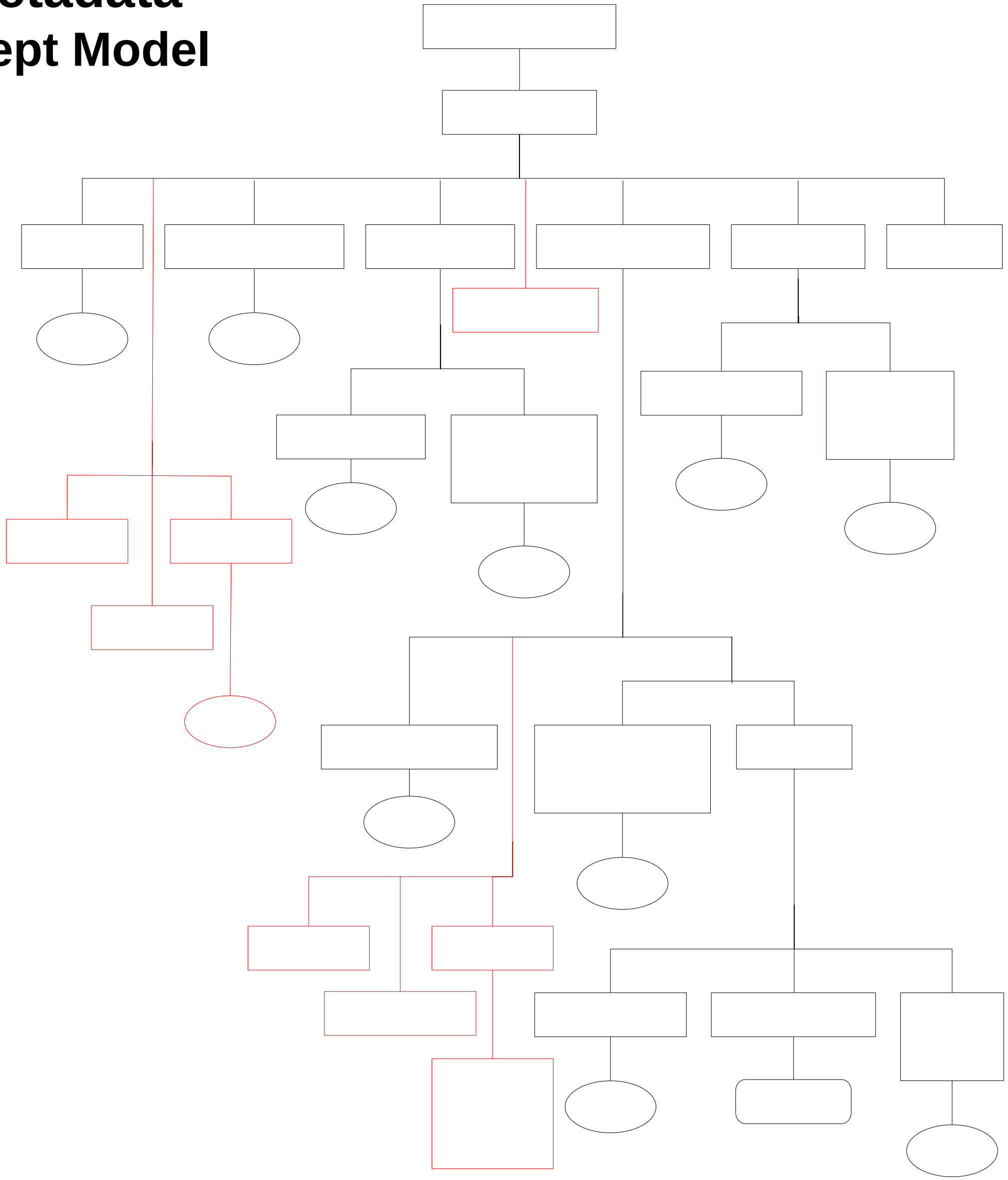


Statistical metadata variable centric concepts in CoSSI



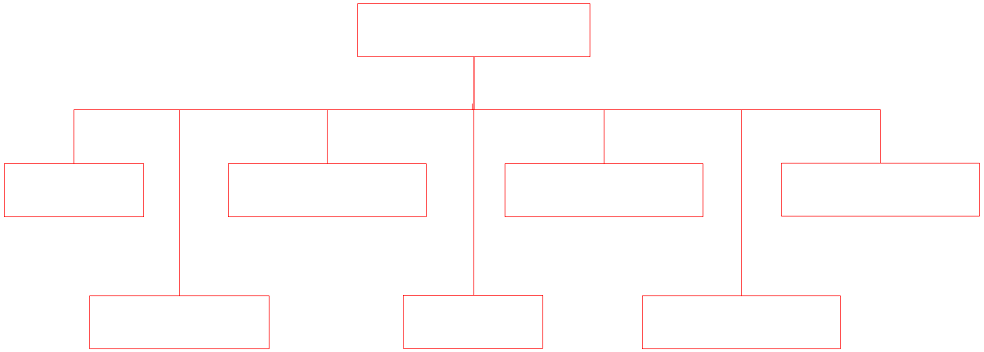
sp

Statistical Metadata - Logical Concept Model (I)

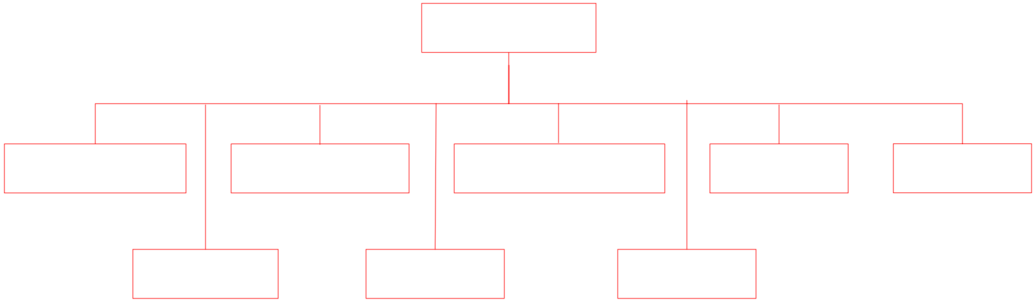


variable

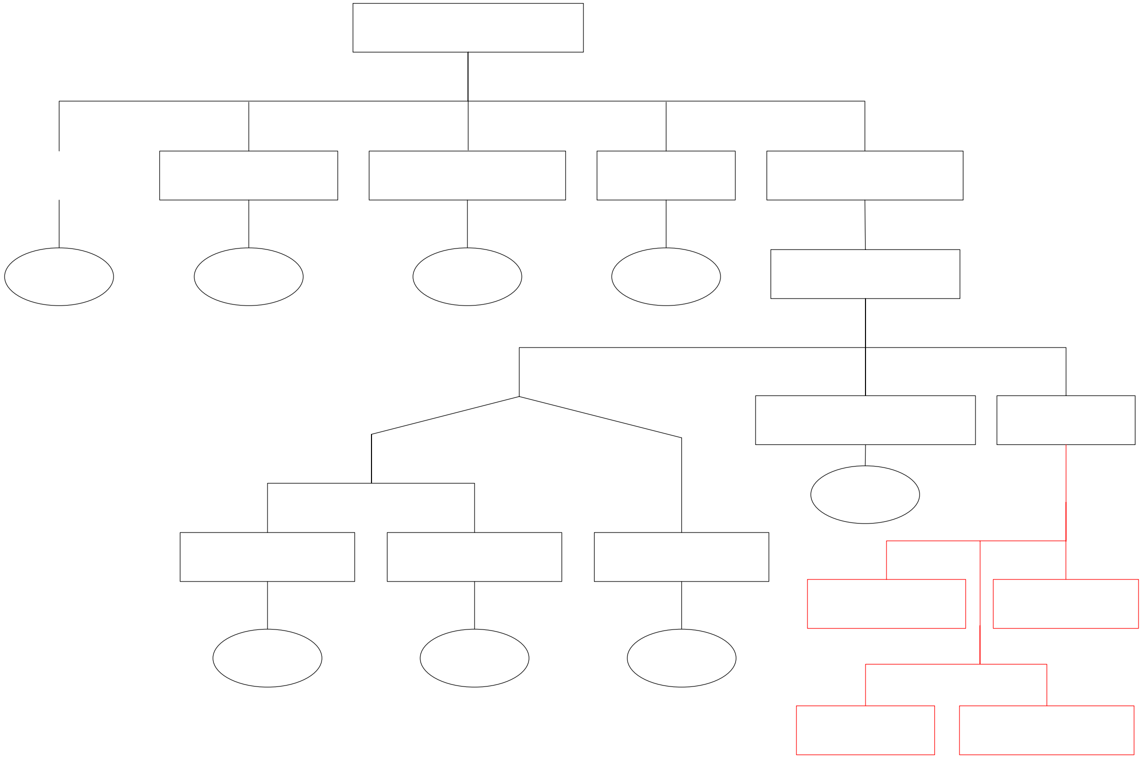
Statistical Metadata - Logical Concept Model (II)



Statistical Metadata - Logical Concept Model (III)

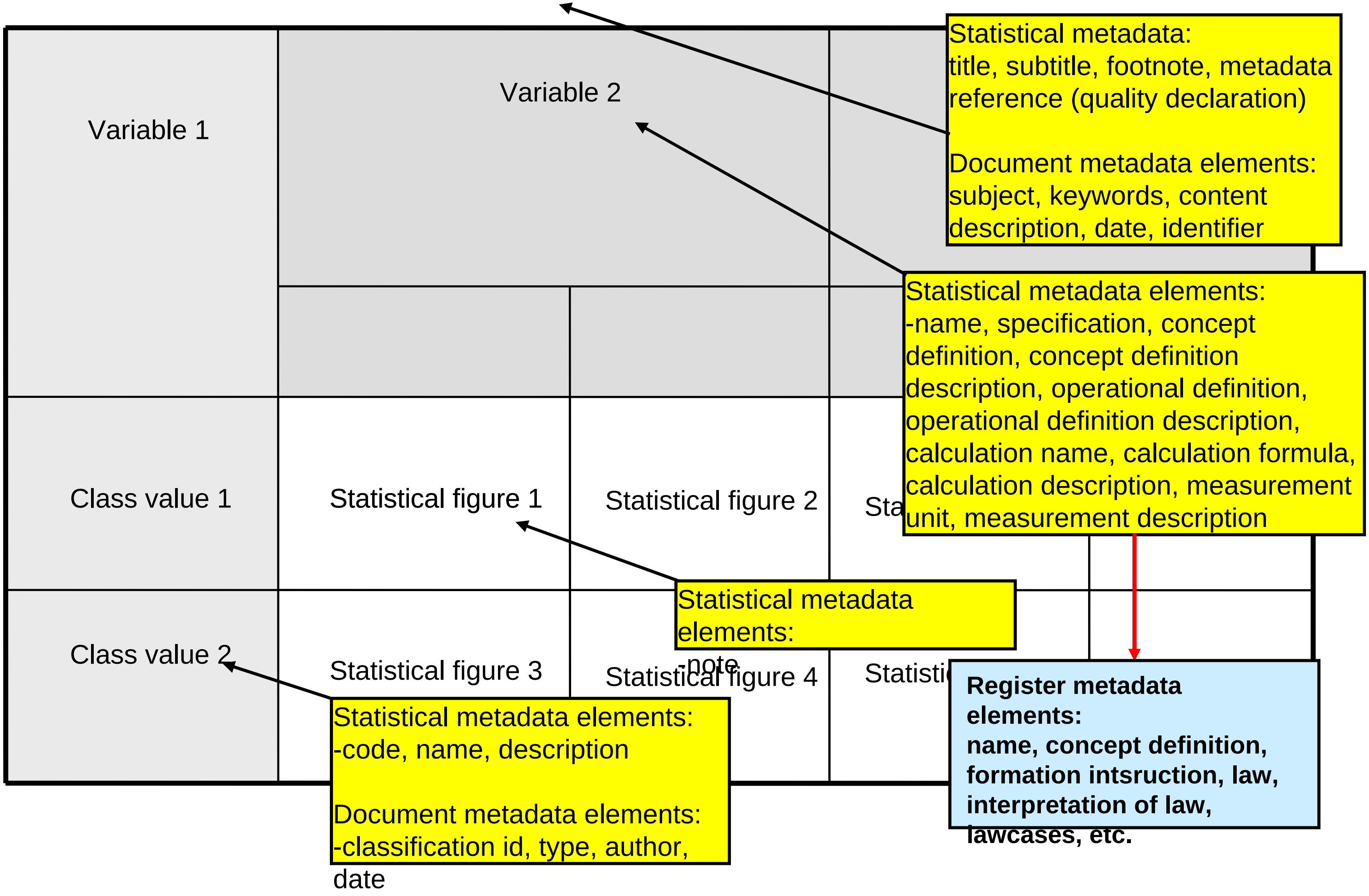


Statistical Metadata - Logical Concept Model (IV)



… and statistical metadata in tables

Table 1. Statistical Metadata in a informative statistical table (I)



Variable 1	Variable 2		
Class value 1	Statistical figure 1	Statistical figure 2	Sta
Class value 2	Statistical figure 3	Statistical figure 4	Statisti

Statistical metadata:
title, subtitle, footnote, metadata reference (quality declaration)

Document metadata elements:
subject, keywords, content description, date, identifier

Statistical metadata elements:
-name, specification, concept definition, concept definition description, operational definition, operational definition description, calculation name, calculation formula, calculation description, measurement unit, measurement description

Statistical metadata elements:
-note

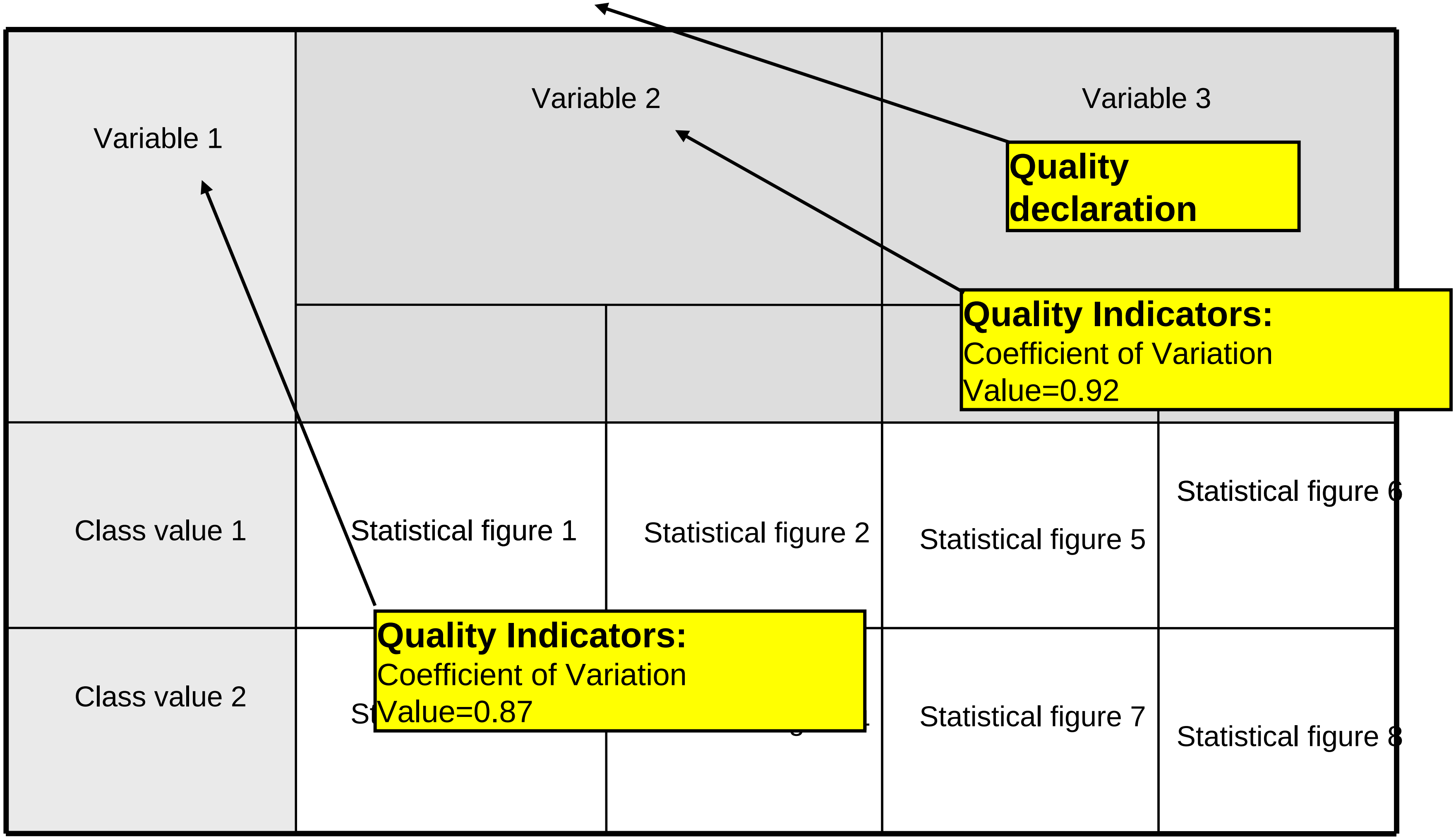
Register metadata elements:
name, concept definition, formation intsruction, law, interpretation of law, lawcases, etc.

Statistical metadata elements:
-code, name, description

Document metadata elements:
-classification id, type, author, date

Statistics Finland

Table 1. Statistical Metadata in a informative statistical table (II)



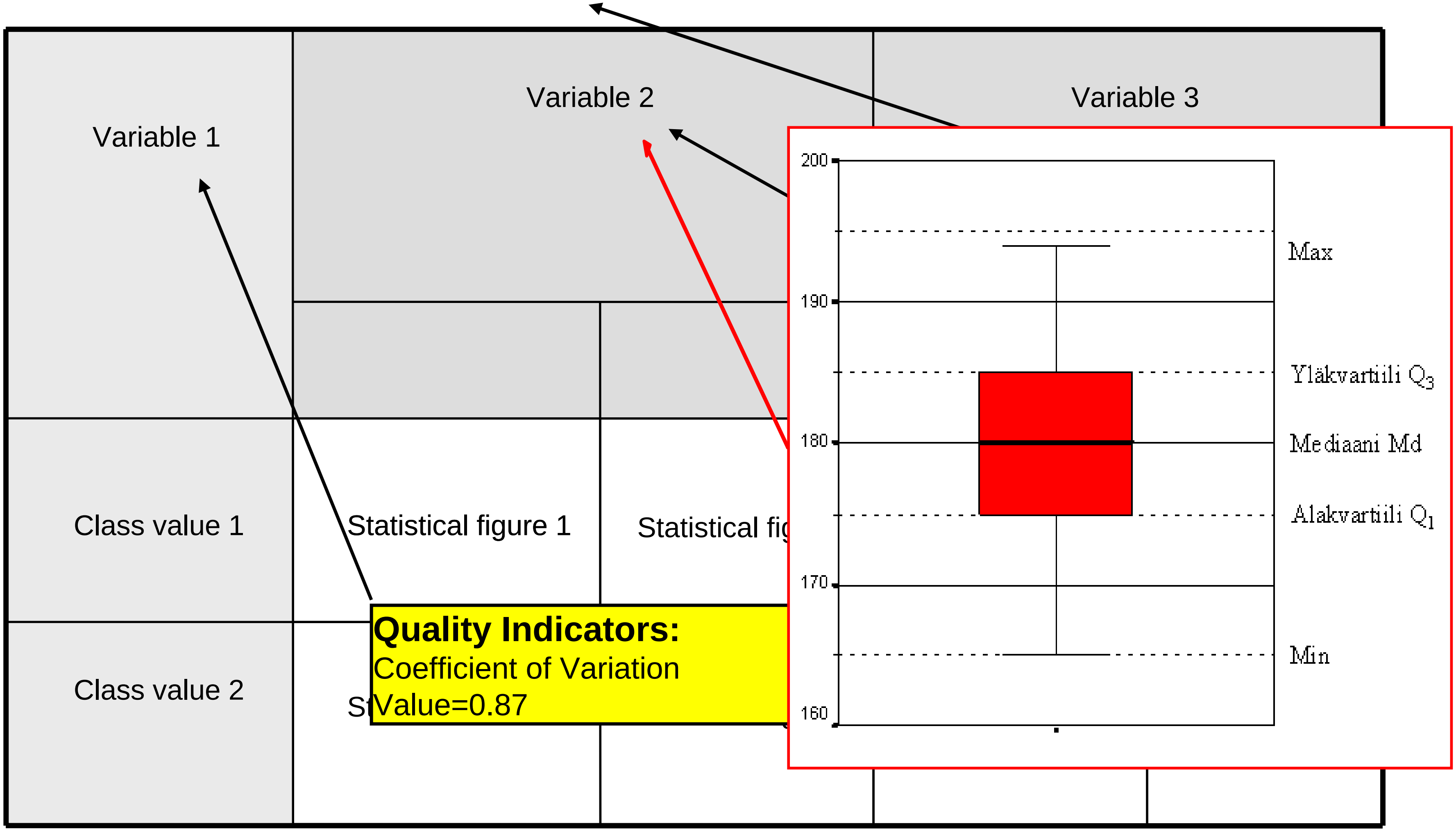
Variable 1	Variable 2		Variable 3	
Class value 1	Statistical figure 1	Statistical figure 2	Statistical figure 5	Statistical figure 6
Class value 2	St		Statistical figure 7	Statistical figure 8

Quality declaration

Quality Indicators:
Coefficient of Variation
Value=0.92

Quality Indicators:
Coefficient of Variation
Value=0.87

Table 1. Statistical Metadata in a informative statistical table (III)



Variable 1

Variable 2

Variable 3

Class value 1

Statistical figure 1

Statistical fig

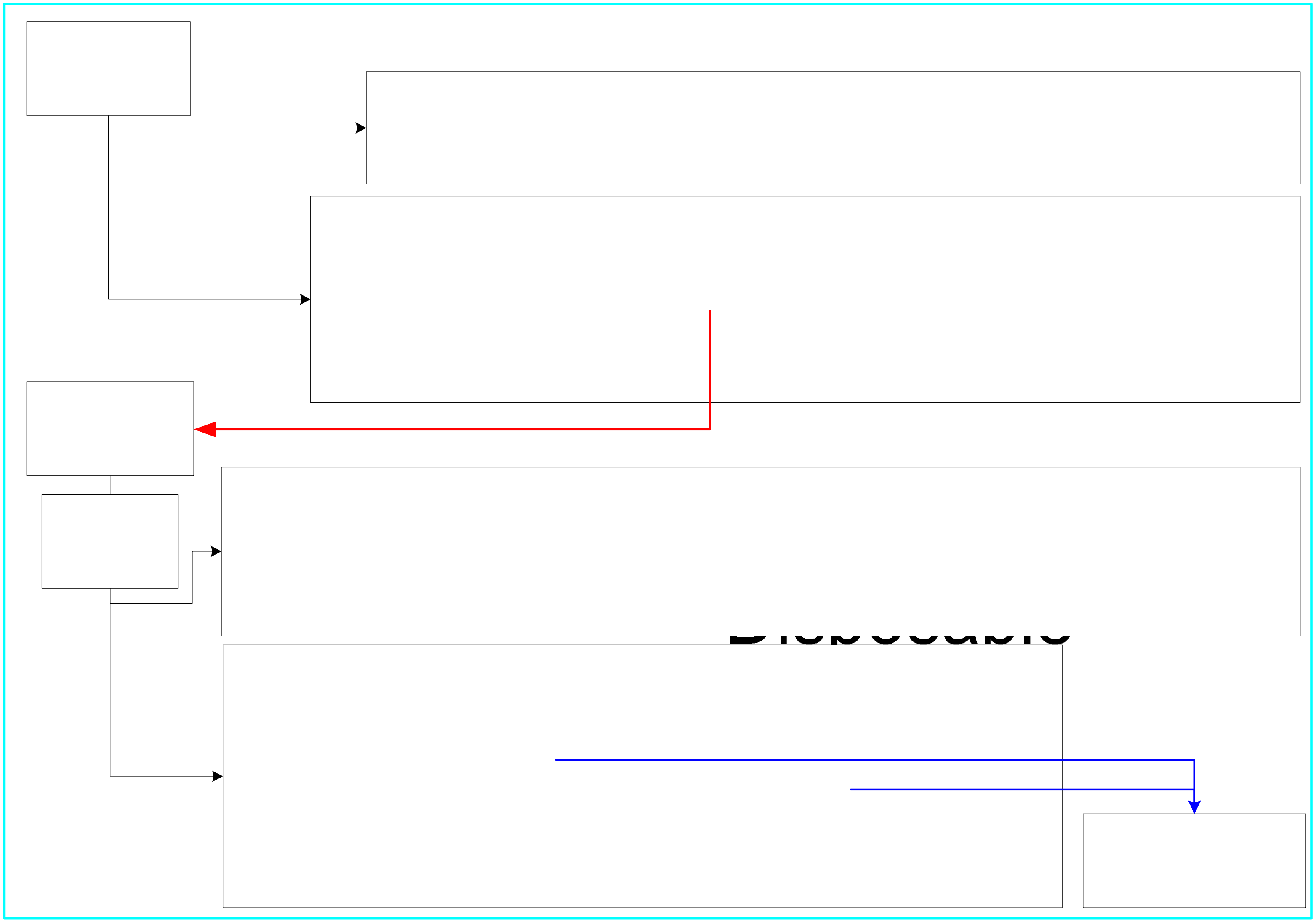
Class value 2

St

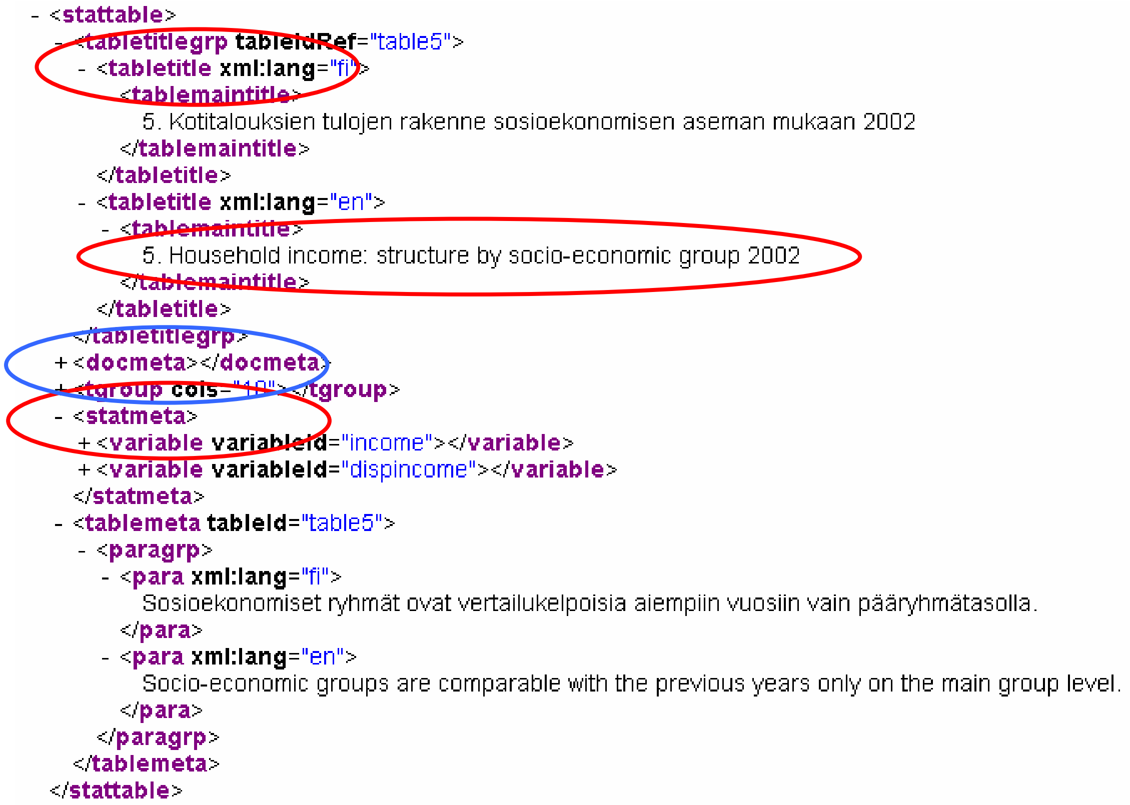
Quality Indicators:
Coefficient of Variation
Value=0.87

… the result from a statistics standpoint …

Income distribution statistics – conceptual model

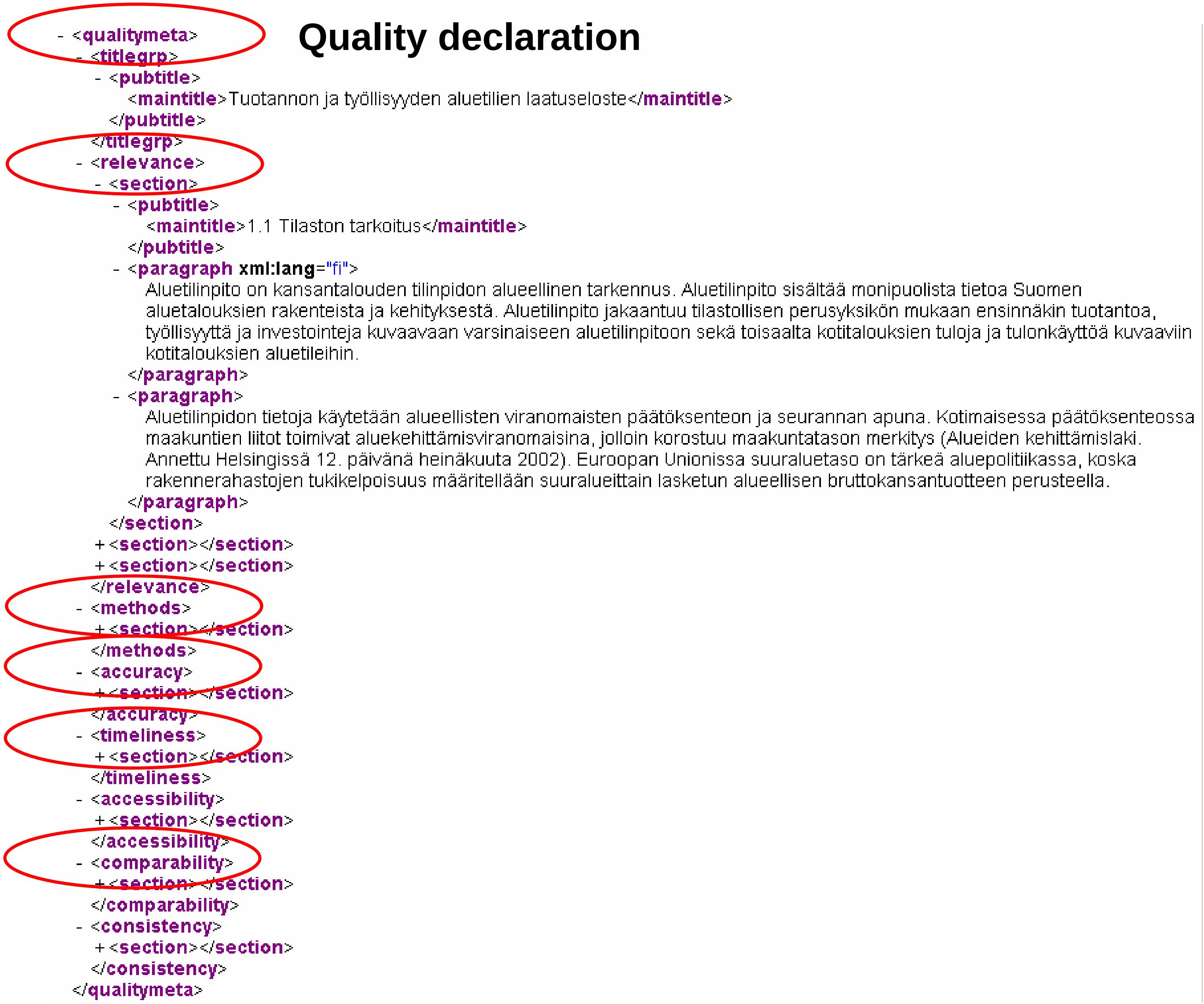


Statistical Table – parts and entity



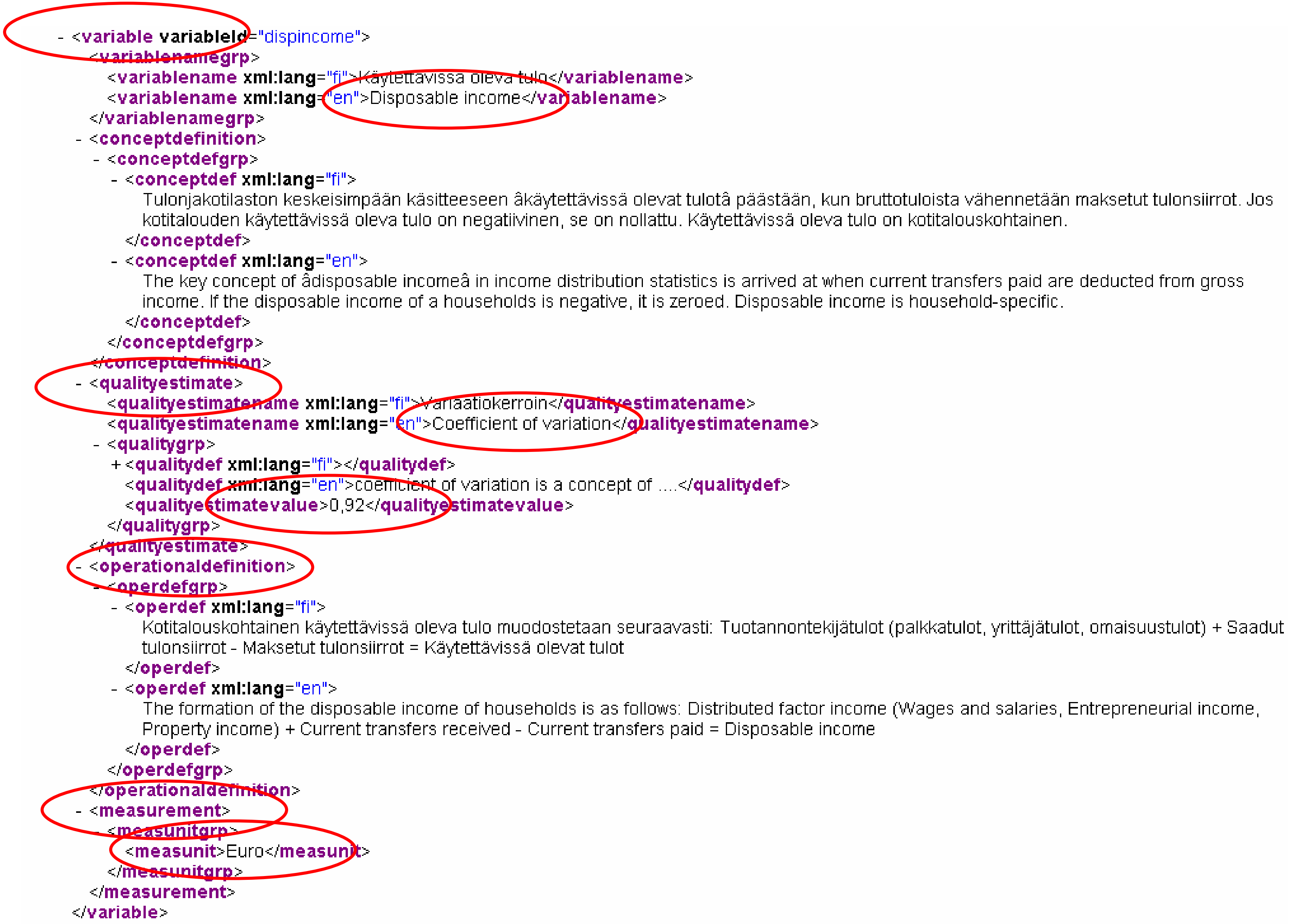
```
- <stattable>
  - <tabletitlegrp tableIdRef="table5">
    - <tabletitle xml:lang="fi">
      - <tablemaintitle>
          5. Kotitalouksien tulojen rakenne sosioekonomisen aseman mukaan 2002
        </tablemaintitle>
      </tabletitle>
    - <tabletitle xml:lang="en">
      - <tablemaintitle>
          5. Household income: structure by socio-economic group 2002
        </tablemaintitle>
      </tabletitle>
    </tabletitlegrp>
  + <docmeta></docmeta>
  + <tgroup cols="10"></tgroup>
  - <statmeta>
    + <variable variableId="income"></variable>
    + <variable variableId="dispincome"></variable>
    </statmeta>
  - <tablemeta tableId="table5">
    - <paragrp>
      - <para xml:lang="fi">
          Sosioekonomiset ryhmät ovat vertailukelpoisia aiempiin vuosiin vain pääryhmätasolla.
        </para>
      - <para xml:lang="en">
          Socio-economic groups are comparable with the previous years only on the main group level.
        </para>
      </paragrp>
    </tablemeta>
  </stattable>
```

Quality declaration



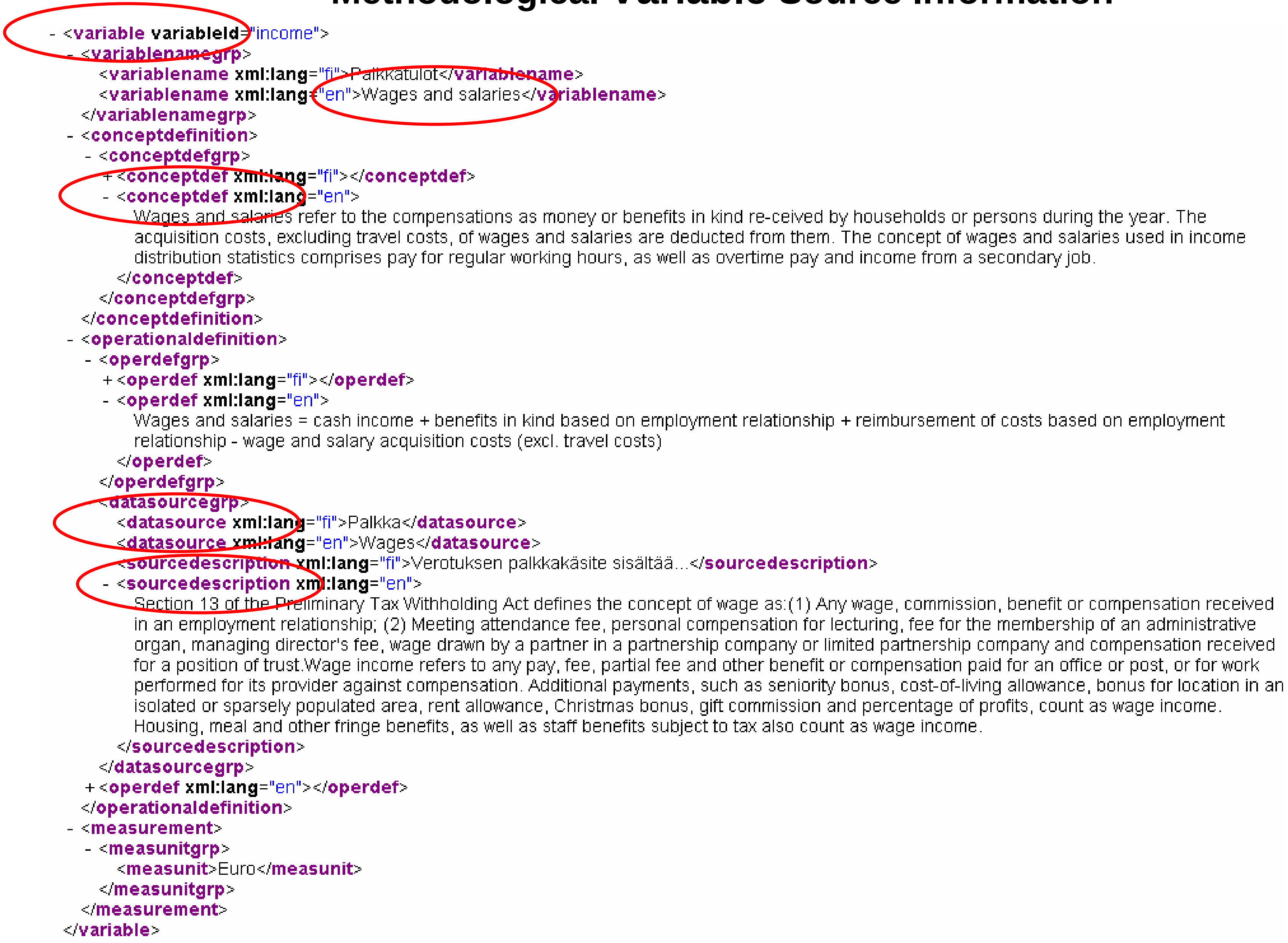
```
- <qualitymeta>
  - <titlegrp>
    - <pubtitle>
        <maintitle>Tuotannon ja työllisyyden aluetilien laatuseloste</maintitle>
      </pubtitle>
    </titlegrp>
  - <relevance>
    - <section>
      - <pubtitle>
          <maintitle>1.1 Tilaston tarkoitus</maintitle>
        </pubtitle>
      - <paragraph xml:lang="fi">
          Aluetilinpito on kansantalouden tilinpidon alueellinen tarkennus. Aluetilinpito sisältää monipuolista tietoa Suomen
          aluetalouksien rakenteista ja kehityksestä. Aluetilinpito jakaantuu tilastollisen perusyksikön mukaan ensinnäkin tuotantoa,
          työllisyyttä ja investointeja kuvaavaan varsinaiseen aluetilinpitoon sekä toisaalta kotitalouksien tuloja ja tulonkäyttöä kuvaaviin
          kotitalouksien aluetileihin.
        </paragraph>
      - <paragraph>
          Aluetilinpidon tietoja käytetään alueellisten viranomaisten päätöksenteon ja seurannan apuna. Kotimaisessa päätöksenteossa
          maakuntien liitot toimivat aluekehittämisviranomaisina, jolloin korostuu maakuntatason merkitys (Alueiden kehittämislaki.
          Annettu Helsingissä 12. päivänä heinäkuuta 2002). Euroopan Unionissa suuraluetaso on tärkeä aluepolitiikassa, koska
          rakennerahastojen tukikelpoisuus määritellään suuralueittain lasketun alueellisen bruttokansantuotteen perusteella.
        </paragraph>
      </section>
    + <section></section>
    + <section></section>
    </relevance>
  - <methods>
    + <section></section>
    </methods>
  - <accuracy>
    + <section></section>
    </accuracy>
  - <timeliness>
    + <section></section>
    </timeliness>
  - <accessibility>
    + <section></section>
    </accessibility>
  - <comparability>
    + <section></section>
    </comparability>
  - <consistency>
    + <section></section>
    </consistency>
  </qualitymeta>
```

Quality information of Variable



```
- <variable variableId="dispincome">
   <variablenamegrp>
     <variablename xml:lang="fi">Käytettävissä oleva tulo</variablename>
     <variablename xml:lang="en">Disposable income</variablename>
   </variablenamegrp>
 - <conceptdefinition>
   - <conceptdefgrp>
     - <conceptdef xml:lang="fi">
         Tulonjakotilaston keskeisimpään käsitteeseen âkäytettävissä olevat tulotâ päästään, kun bruttotuloista vähennetään maksetut tulonsiirrot. Jos
         kotitalouden käytettävissä oleva tulo on negatiivinen, se on nollattu. Käytettävissä oleva tulo on kotitalouskohtainen.
       </conceptdef>
     - <conceptdef xml:lang="en">
         The key concept of âdisposable incomeâ in income distribution statistics is arrived at when current transfers paid are deducted from gross
         income. If the disposable income of a households is negative, it is zeroed. Disposable income is household-specific.
       </conceptdef>
     </conceptdefgrp>
   </conceptdefinition>
 - <qualityestimate>
     <qualityestimatename xml:lang="fi">Variaatiokerroin</qualityestimatename>
     <qualityestimatename xml:lang="en">Coefficient of variation</qualityestimatename>
   - <qualitygrp>
     + <qualitydef xml:lang="fi"></qualitydef>
       <qualitydef xml:lang="en">coefficient of variation is a concept of ....</qualitydef>
       <qualityestimatevalue>0,92</qualityestimatevalue>
     </qualitygrp>
   </qualityestimate>
 - <operationaldefinition>
   - <operdefgrp>
     - <operdef xml:lang="fi">
         Kotitalouskohtainen käytettävissä oleva tulo muodostetaan seuraavasti: Tuotannontekijätulot (palkkatulot, yrittäjätulot, omaisuustulot) + Saadut
         tulonsiirrot - Maksetut tulonsiirrot = Käytettävissä olevat tulot
       </operdef>
     - <operdef xml:lang="en">
         The formation of the disposable income of households is as follows: Distributed factor income (Wages and salaries, Entrepreneurial income,
         Property income) + Current transfers received - Current transfers paid = Disposable income
       </operdef>
     </operdefgrp>
   </operationaldefinition>
 - <measurement>
   - <measunitgrp>
       <measunit>Euro</measunit>
     </measunitgrp>
   </measurement>
 </variable>
```

Methodological Variable Source Information



```
- <variable variableId="income">
  - <variablenamegrp>
      <variablename xml:lang="fi">Palkkatulot</variablename>
      <variablename xml:lang="en">Wages and salaries</variablename>
    </variablenamegrp>
  - <conceptdefinition>
    - <conceptdefgrp>
      + <conceptdef xml:lang="fi"></conceptdef>
      - <conceptdef xml:lang="en">
          Wages and salaries refer to the compensations as money or benefits in kind re-ceived by households or persons during the year. The
          acquisition costs, excluding travel costs, of wages and salaries are deducted from them. The concept of wages and salaries used in income
          distribution statistics comprises pay for regular working hours, as well as overtime pay and income from a secondary job.
        </conceptdef>
      </conceptdefgrp>
    </conceptdefinition>
  - <operationaldefinition>
    - <operdefgrp>
      + <operdef xml:lang="fi"></operdef>
      - <operdef xml:lang="en">
          Wages and salaries = cash income + benefits in kind based on employment relationship + reimbursement of costs based on employment
          relationship - wage and salary acquisition costs (excl. travel costs)
        </operdef>
      </operdefgrp>
    - <datasourcegrp>
        <datasource xml:lang="fi">Palkka</datasource>
        <datasource xml:lang="en">Wages</datasource>
        <sourcedescription xml:lang="fi">Verotuksen palkkakäsite sisältää...</sourcedescription>
      - <sourcedescription xml:lang="en">
          Section 13 of the Preliminary Tax Withholding Act defines the concept of wage as:(1) Any wage, commission, benefit or compensation received
          in an employment relationship; (2) Meeting attendance fee, personal compensation for lecturing, fee for the membership of an administrative
          organ, managing director's fee, wage drawn by a partner in a partnership company or limited partnership company and compensation received
          for a position of trust.Wage income refers to any pay, fee, partial fee and other benefit or compensation paid for an office or post, or for work
          performed for its provider against compensation. Additional payments, such as seniority bonus, cost-of-living allowance, bonus for location in an
          isolated or sparsely populated area, rent allowance, Christmas bonus, gift commission and percentage of profits, count as wage income.
          Housing, meal and other fringe benefits, as well as staff benefits subject to tax also count as wage income.
        </sourcedescription>
      </datasourcegrp>
    + <operdef xml:lang="en"></operdef>
    </operationaldefinition>
  - <measurement>
    - <measunitgrp>
        <measunit>Euro</measunit>
      </measunitgrp>
    </measurement>
  </variable>
```

Thank you for your attention!