DDI and Data

Hans Jørgen Marker
Senior Researcher
Dansk Data Arkiv
hjm@dda.dk

Motivation

- Data and metadata belongs together
- Redundant metadater?
- Scope:
  - Tabular data
    - Data that will fit nicely into a table or a system of tables
      - Spreadsheet, Data base,Statistical data set

Pre DDI History

- OSIRIS
  - Dictionary
  - Data
- SSD
- Solutions based on OSIRIS

Standard Study Description

```
000001522 DK  DD941109Datamateriale DDA-1522:                        09
001   07*
002   10*
00301 folketingsvalg 2001, Enghave og Hellerup skoler;
      Enghave skole, folketingsvalg 2001;
      Hellerup skole, folketingsvalg 2001; partipr‘ference;*
007   Folketingsvalget 2001: Enghave Skole og Hellerup Skole*
101   Datamateriale DDA-11522:
     1Folketingsvalget 2001: Enghave Skole og Hellerup Skole.
     1Prim‘runders›gere: Bo Falsig, Thomas Hartvig, Michael Bucka og
      Gisle Thorsen.
     1DDA-11522, 1. udgave (ved Birgitte Gr›nlund Jensen og Bernhard
      Hansen).
     1Dansk Data Arkiv 2002.
     11 datafil (1131 respondenter, 15 variable) med tilh›rende
      maskinl‘sbar dokumentation (24 pp.).*
11105 11522*         … etc.
```

OSIRIS dictionary

```
T0010 PARTIER HAR IDEOLOGIER   001700010  010000009       V10
Q00100010 Spm. 5D: Jeg mener, at partierne stadig har ideologier (over-
K0010     ordnede visioner) om samfundet og Danmarks fremtid.
X0010     S‘t 1 kryds i hvert af f›lgende 4 udsagn
C0010      1021. Meget enig
C0010      6682. Enig
C0010      1673. Ved ikke
C0010      1574. Uenig
C0010       255. Meget uenig
C0010       129. Uoplyst
T0011 HVILKET PARTI STEMTE P□  001800020  010000099       V11
```

Archive file format

- Preservation strategies
- Loss free conversion?
- Storage of metadata and data
  - Database or archive file

Unsolved OSIRIS problems

- More than one table
- More than 9999 variables
- Codes on string variables
- Missing intervals

Is DDI the solution?

- The known issues with OSIRIS are solved in DDI
- Some structural issues in DDI 2.0 will be solved in 3.0
- But what about data and the archive file format?

Some central elements in DDI

<!ELEMENT codeBook (…,fileDscr*,dataDscr*,…)>

<!ELEMENT fileDscr >

<!ATTLIST fileDscr ID ID #IMPLIED …>

<!ELEMENT dataDscr (…,var*,…)>

<!ATTLIST dataDscr ID ID #IMPLIED>

<!ELEMENT var >

<!ATTLIST … files IDREFS #IMPLIED … >

Central elements explained

codeBook is the top level element of DDI 2.0

docDscr infomation on the DDI document it self

stdyInfo Study scope: Universe, methodology etc.

fileDscr Data file: structure, format etc.

dataDscr Variables: cubes, groups, questions, vars, codes

otherMat Other material: notes tables etc.

Dansk Data Arkiv

Hans Jørgen Marker

Creating a home for the data

<!ELEMENT dataTable (record+)>

<!ELEMENT record (cell+)>

<!ELEMENT cell (#PCDATA)>

Dansk Data Arkiv

Hans Jørgen Marker

Linking Data and Documentation

<!ATTLIST cell var IDREF #REQUIRED>

and perhaps

<!ATTLIST dataTable dataDscr IDREF
#REQUIRED>

and on top of it

<!ELEMENT formArk (codeBook+,dataTable+)>

Dansk Data Arkiv

Hans Jørgen Marker

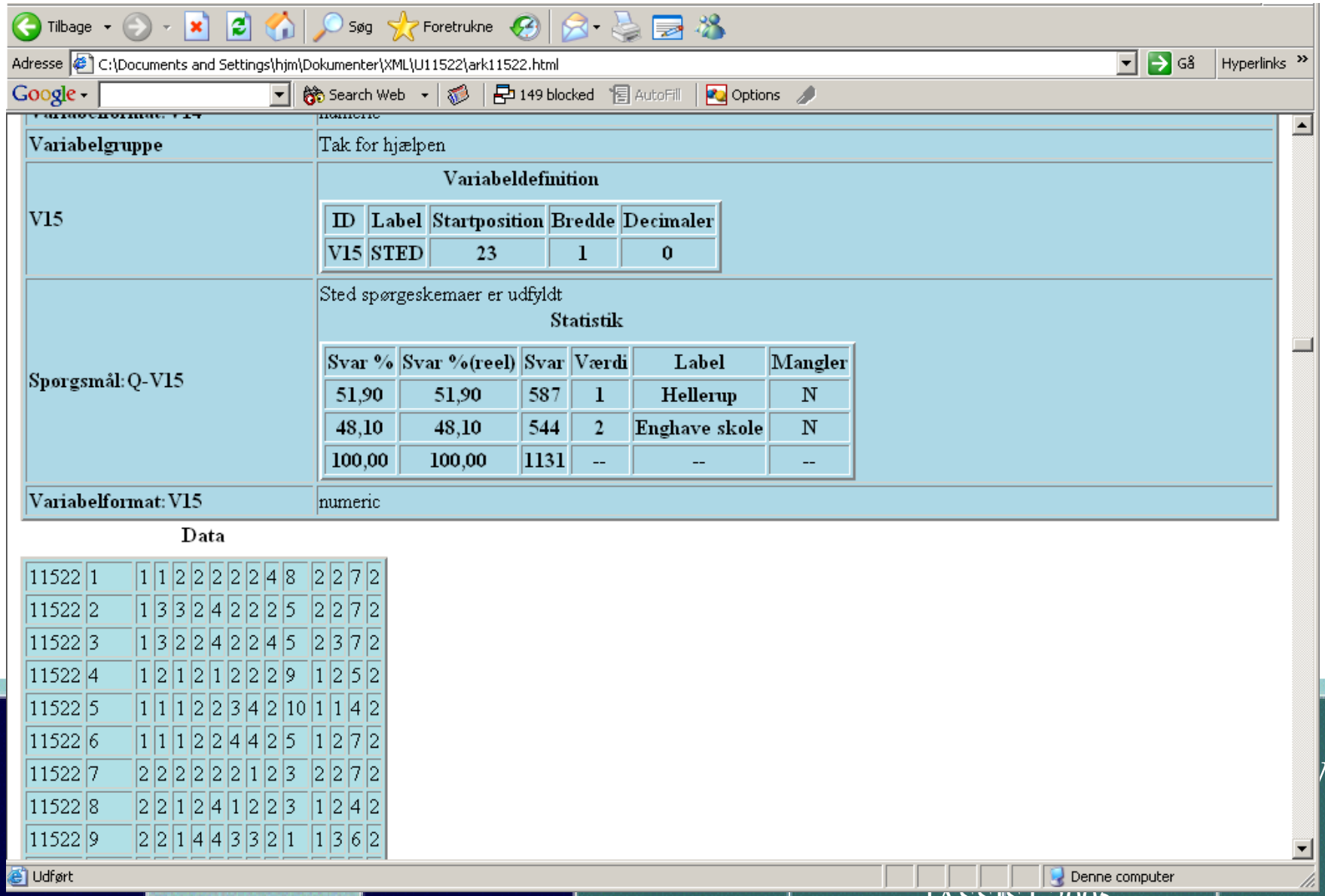
What about data structure?

- Primary key
- Index
- Foreign key

An example

```
        </dataDscr>
</codeBook>
<dataTable dataDscr="T1">
      <record>
       <cell var="V1">11522</cell>
       <cell var="V2">1</cell>
       <cell var="V3">1</cell>
       <cell var="V4">1</cell>
       <cell var="V5">2</cell>
       <cell var="V6">2</cell>
       <cell var="V7">2</cell>
       <cell var="V8">2</cell>
       <cell var="V9">2</cell>
       <cell var="V10">4</cell>
       <cell var="V11">8</cell>
       <cell var="V12">2</cell>
       <cell var="V13">2</cell>
       <cell var="V14">7</cell>
       <cell var="V15">2</cell>
      </record>
      <record>
       <cell var="V1">11522</cell>
```

Using a stylesheet



Potential usage

- Communication
- Processing