

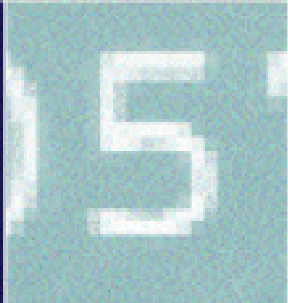
DDI and Data

Hans Jørgen Marker

Senior Researcher

Dansk Data Arkiv

hjm@dda.dk



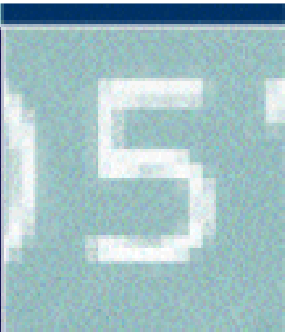
Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

Motivation

- Data and metadata belongs together
- Redundant metadata?
- 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

15

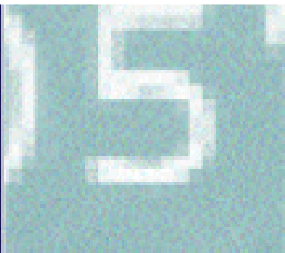
Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

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.



Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

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

15

Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

Archive file format

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

15

Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

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?

15

Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005

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 information on the DDI document itself

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.

Creating a home for the data

<!ELEMENT dataTable (record+)>

<!ELEMENT record (cell+)>

<!ELEMENT cell (#PCDATA)>

Linking Data and Documentation

<!ATTLIST cell var IDREF #REQUIRED>

and perhaps

<!ATTLIST dataTable dataDscr IDREF
#REQUIRED>

and on top of it

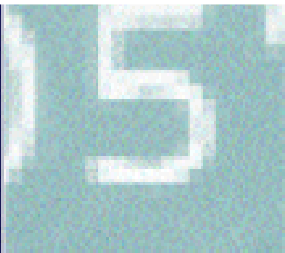
<!ELEMENT formArk (codeBook+,dataTable+)>

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

Tilbage Søg Foretrukne

Adresse C:\Documents and Settings\hjm\Dokumenter\XML\U11522\ark11522.html Gå Hyperlinks

Google Search Web 149 blocked AutoFill Options

Variabelgruppe	Tak for hjælpen																								
V15	<p>Variabeldefinition</p> <table border="1"> <thead> <tr> <th>ID</th> <th>Label</th> <th>Startposition</th> <th>Bredde</th> <th>Decimaler</th> </tr> </thead> <tbody> <tr> <td>V15</td> <td>STED</td> <td>23</td> <td>1</td> <td>0</td> </tr> </tbody> </table>	ID	Label	Startposition	Bredde	Decimaler	V15	STED	23	1	0														
ID	Label	Startposition	Bredde	Decimaler																					
V15	STED	23	1	0																					
Spørgsmål: Q-V15	<p>Sted spørgeskemaer er udfyldt</p> <p>Statistik</p> <table border="1"> <thead> <tr> <th>Svar %</th> <th>Svar %(reel)</th> <th>Svar</th> <th>Værdi</th> <th>Label</th> <th>Mangler</th> </tr> </thead> <tbody> <tr> <td>51,90</td> <td>51,90</td> <td>587</td> <td>1</td> <td>Hellerup</td> <td>N</td> </tr> <tr> <td>48,10</td> <td>48,10</td> <td>544</td> <td>2</td> <td>Enghave skole</td> <td>N</td> </tr> <tr> <td>100,00</td> <td>100,00</td> <td>1131</td> <td>--</td> <td>--</td> <td>--</td> </tr> </tbody> </table>	Svar %	Svar %(reel)	Svar	Værdi	Label	Mangler	51,90	51,90	587	1	Hellerup	N	48,10	48,10	544	2	Enghave skole	N	100,00	100,00	1131	--	--	--
Svar %	Svar %(reel)	Svar	Værdi	Label	Mangler																				
51,90	51,90	587	1	Hellerup	N																				
48,10	48,10	544	2	Enghave skole	N																				
100,00	100,00	1131	--	--	--																				
Variabelformat: V15	numeric																								

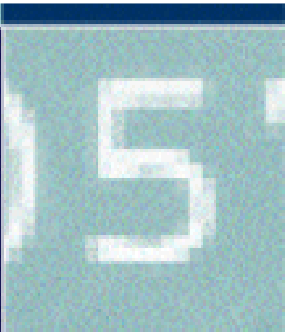
Data

11522 1	1	1	1	2	2	2	2	2	4	8	2	2	7	2
11522 2	1	3	3	2	4	2	2	2	5	2	2	7	2	
11522 3	1	3	2	2	4	2	2	4	5	2	3	7	2	
11522 4	1	2	1	2	1	2	2	9	1	2	5	2		
11522 5	1	1	1	2	2	3	4	2	10	1	1	4	2	
11522 6	1	1	1	2	2	4	4	2	5	1	2	7	2	
11522 7	2	2	2	2	2	1	2	3	2	2	7	2		
11522 8	2	2	1	2	4	1	2	2	3	1	2	4	2	
11522 9	2	2	1	4	4	3	3	2	1	1	3	6	2	

Udført Denne computer

Potential usage

- Communication
- Processing



Dansk Data Arkiv

Hans Jørgen Marker

IASSIST 2005