

DDI-Publishing Made Easy- the Nesstar Way

Jostein Ryssevik
Nesstar Ltd.

The DDI in action – what do we know?

- The costs of migrating a data archive to the DDI are high
- The DDI is a “cathedral standard” (no data provider is buying the full package – they are all using the DDI building blocks to build their own more modest “parish churches”).
- The DDI is a very loose structure loaded with alternatives and ambiguities (a single study described by two different archives will probably look quite different)
- The DDI is only telling half the story (data providers will have to add their own local guidelines on top of the DDI (controlled vocabularies, mandatory elements etc) to secure internal standardization)
- The DDI is inflexible (there is no extension mechanism that allows a data provider to add local elements without breaking the standard)
-however, the gains of moving to the DDI are so high that it is all worth the efforts

Requirements to a DDI authoring/production system

- **Efficiency:** Metadata authoring is time-consuming. Methods to save time and increase efficiency are therefore important. This includes automation of processes and methods to avoid unnecessary repetition of tasks.
- **Data/metadata import:** Most data sources have some sort machine-readable documentation. An efficient metadata authoring tool should be able to capture this information and put it into the DDI structure
- **Write once – use many:** Once you have a metadata instance describing your data you should be able to use it for many purposes (like producing data files for different statistical packages)
- **Data/metadata integration:** An efficient and flexible metadata authoring tool should be able to handle the data as well as the metadata (integrity checking, automatic creation of summary statistics and frequencies etc.)
- **Simplicity:** Metadata authoring is often done by persons with limited technological skills. The tools should thus be designed for none-experts.
- **None-linearity:** The various sub-tasks of a metadata authoring and publishing process have no fixed ordering. As far as possible the tools should therefore allow sub-tasks to be carried out in freely chosen order.
- **Customisation:** Organizations are using the DDI differently but are aiming at internal standardization. The tool should support local customisation and organization-specific validation.

Nesstar DDI tools

- Nesstar is a complete software system providing distributed access to data over the Internet. Nesstar is building on and is fully compatible with the DDI.
- Nesstar supports searching on a variety of DDI elements.
- Nesstar uses the DDI to navigate data resources
- A first implementation of the new DDI model for aggregated data (cubes) has just been completed.
- Nesstar provides a set of DDI authoring tools that facilitates efficient production and publishing of DDI compliant data.
 - A java client for remote publishing of data to a Nesstar server
 - Various metadata entry and authoring tools based on technology from the statistical package NSDstat (C++), the last generation of this has been developed in cooperation with Health Canada.

Characteristics of the DDI-authoring tools

- Import of data from a variety of formats (SPSS, SAS, Statistica, Stata etc.)
- Data metadata integration facilitating integrity checking as well as automatic creation of summary statistics.
- Powerful metadata editing capabilities:
 - Tools for efficient entering of information for multiple variables.
 - Tools for efficient creation of the variable group hierarchies
 - A category set repository supporting efficient reuse of variable metadata within and across datasets
- Template driven interface supporting the development and freezing of local “best practices”, selection of obligatory elements, local controlled vocabularies etc)
- Export of data to a variety of formats

The variable info screen

Document Description | Study Description | File Description | **Variables** | Variable Groups | Data Entry | Other Material

Variables:

Number	Name	Label	StartCol	EndCol	Width
v1	REGION	Region or economic group	0	2	2
v2	COUNTRY	Country	2	4	2
v3	POPULATN	Population in thousands	4	6	2
v4	DENSITY	Number of people / sq. kilometer	6	8	2
v5	URBAN	People living in cities (%)	8	10	2
v6	RELIGION	Predominant religion	10	12	2
v7	LIFEEXPF	Average female life expectancy	12	14	2
v8	LIFEEXPM	Average male life expectancy	14	16	2
v9	LITERACY	People who read (%)	16	18	2
v10	POP_INCR	Population increase (% per year)	18	20	2
v11	BABYMORT	Infant mortality (deaths per 1000 live births)	20	22	2
v12	GDP_CAP	Gross domestic product / capita	22	24	2
v13	CALORIES	Daily calorie intake	24	26	2
v14	AIDS	Aids cases	26	28	2
v15	BIRTH_RT	Birth rate per 1000 people	28	30	2
v16	DEATH_RT	Death rate per 1000 people	30	32	2
v17	AIDS_RT	Number of aids cases / 100000 people	32	34	2
v18	LOG_GDP	Log (base 10) of GDP_CAP	34	36	2
v19	LG_AIDSR	Log (base 10) of AIDS_RT	36	38	2

Variable Information:

Value	Label	N	WN
1	OECD	100	32
2	East Euro	143	279
3	Pacific/As	13	298
4	Africa	106	45
5	Middle Ea:	161	91
6	Latn Amer	292	137

Category Text:

Documentation:

Weights	PreQuestion Text	Literal Question	PostQuestion Text
Interviewer Instructions	Response Unit	Analysis Unit	Universe
Variable Text	Security	Embargo	Standard Categories
Undocumented Codes	Imputation	Coder Instructions	Notes

Now a few questions about the European Union. On the whole do you think the European Union has been good for Scotland or, bad for Scotland?

Predefined valuesets:

Variable information:

Data Type: Numeric

Measure: Nominal

Minimum: 1 Maximum: 6 Decimals: 0

Implicit decimals:

Missing data:

Creation of the variable group hierarchies

The screenshot displays a software interface with a menu bar at the top containing: Document Description, Study Description, File Description, Variables, Variable Groups, Data Entry, and Other Material. Below the menu bar, there are two main sections: 'Groups:' on the left and 'Group parameters:' on the right.

The 'Groups:' section shows a tree view with the following structure:

- Health indicators
 - Background variables
 - Smoking habits
- Politics
 - Background variables

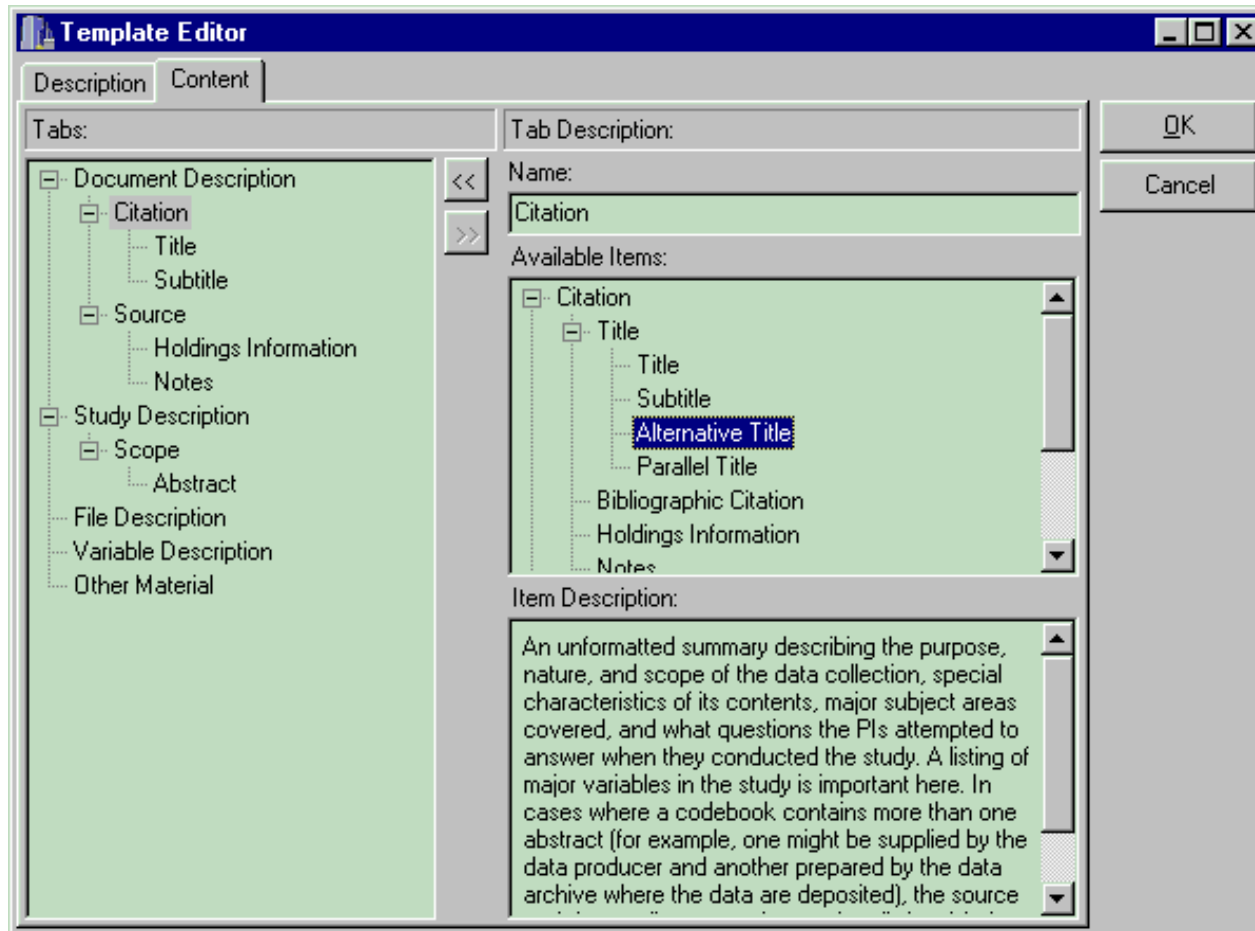
The 'Group parameters:' section has two tabs: 'Description' and 'Variables'. The 'Variables' tab is active, showing a table with the following data:

Dataset	Number	Name	Label
hse97.ai	v1	H SERIAL	(D) household serial number
hse97.ai	v3	ADULTS	Number of adults in HH
hse97.ai	v4	CHILDREN	Number of children in HH
hse97.ai	v5	INFANTS	Number of infants in HH
hse97.ai	v6	HHSIZE	(D) Household Size
hse97.ai	v7	HHDTYPB	(D) Household Type
hse97.ai	v8	TENURE	Household Tenure
hse97.ai	v9	FURN	Is accommodation furnished?
hse97.ai	v10	BEDROOMS	No. of bedrooms in household
hse97.ai	v11	CAR	Car available
hse97.ai	v12	NUMCARS	Number of cars available
hse97.ai	v13	P SERIAL	(D) person serial number
hse97.ai	v15	SEX	Sex
hse97.ai	v16	AGE	Age last birthday

Entering of study-level metadata

Document Description	Study Description	File Description	Variables	Variable Groups	Data Entry	Other Material
Study Scope	Data Access					
Abstract:						
The National Child Development Study (NCDS) originated in the 'Perinatal Mortality Survey', which examined social and obstetric factors associated with still birth and infant mortality among over 17,000 babies born in Britain in the week 3-9 March 1958. Surviving members of this birth cohort have been surveyed on five further occasions in order to monitor their changing health, education, social and economic circumstances - in 1965 (age 7), 1969						
Keywords:						
ABORTION ACCIDENTS ADOPTED CHILDREN ADVANCED LEVEL EXAMINATIONS ADVANCED SUPPLEMENTARY LEVEL EXAMINATIONS AGE						
Topic Classification:						
1970 British Cohort Study (BCS70) - Major studies National Child Development Study - Major studies Social attitudes - Social issues, attitudes and behaviour Family life and marriage - Social issues, attitudes and behaviour Child development and child rearing - Social issues, attitudes and behaviour General studies - Employment and labour						
Universe:						
Population: Previous NCDS and BCS70 respondents living in Great Britain and outlying islands (Channel Islands, Isle of Man, Orkney, Shetland and the Western Isles), during 1999-2000, who were not known to have died, emigrated or refused.						

Creating a template



Looking at the data matrix

The screenshot displays a software interface for viewing a data matrix. The main window is titled 'Data Entry' and contains a table with 15 rows and 5 columns. The columns are labeled 'v1 - REGION', 'v2 - COUNTRY', 'v3 - POPULATN', 'v4 - DENSITY', and 'v!'. The rows are numbered 1 through 15. The fifth row is highlighted in blue. To the right of the table is a 'Value labels' panel with a table of 'Value' and 'Label' pairs. Below this panel is a 'Variable information' section with fields for 'Minimum', 'Maximum', and 'Decimals'. At the bottom is a 'Case documentation' section with a large empty text area and a 'Missing data' section with three input fields.

	v1 - REGION	v2 - COUNTRY	v3 - POPULATN	v4 - DENSITY	v!
1	Pacific/Asia	Afghanistan	20500	25.0	
2	Latn America	Argentina	33900	12.0	
3	Middle East	Armenia	3700	126.0	
4	OECD	Australia	17800	2.3	
5	OECD	Austria	8000	94.0	
6	Middle East	Azerbaijan	7400	86.0	
7	Middle East	Bahrain	600	828.0	
8	Pacific/Asia	Bangladesh	125000	800.0	
9	Latn America	Barbados	256	605.0	
10	East Europe	Belarus	10300	50.0	
11	OECD	Belgium	10100	329.0	
12	Latn America	Bolivia	7900	6.9	
13	East Europe	Bosnia	4600	87.0	
14	Africa	Botswana	1359	2.4	
15	Latn America	Brazil	156600	18.0	

Value labels:

Value	Label
1	OECD
2	East Europe
3	Pacific/Asia
4	Africa
5	Middle East
6	Latn America

Variable information

Minimum:	Maximum:	Decimals:
1	6	0

Missing data:

	*	

Variable documentation