

Statistical Metadata Panel e-SS Conference

Pasqualino “Titto” Assini
NCeSS (U.K. Data Archive)
University of Essex, U.K.



No Shortage of Statistical Metadata Standards

- # The Common Warehouse Metamodel (CWM) from OMG – data warehousing and business intelligence.
 - # ISO 11179 – data elements in a metadata repository.
 - # SDMX – multidimensional data and time-series.
 - # IQML, AskXML and Triple-S - questionnaire data
 - # The Data Documentation Initiative (DDI) – a general metadata standard for statistical data (micro as well as aggregated)
 - # And many other **related standards**. e-Social Science requires more than simple "data" metadata:
 - Thesauri, Classifications.
-

< ddi >

Metadata powered by the
Data Documentation Initiative

- # The **DDI** is particularly interesting from an e-SS perspective as it is widely adopted by social sciences data archives all over the world that provide many of the datasets used by social scientists for secondary analysis.
 - # Initiated and organised by the the Inter-University Consortium for Political and Social Research (USA) in 1995 to create a metadata standard for the social science community.
 - # Members coming from social science data archives and libraries in USA, Canada and Europe and from major producers of statistical data.
 - # First in SGML then in XML.
 - # DDI 1.0 published in 2000. Currently at version 2. Version 3 is being designed and it is scheduled for 2006.
-

The Structure of a DDI Codebook

Document Description

- Description of the codebook document itself (author, sources, etc).

Study Description

- Information about the entire study or data collection (content, collection methods, processing, sources, access conditions etc).

File Description

- Description of each single file of the data collection (formats, dimensions, processing information, etc.).

Data Description

- Description of each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.).

Other Study-related Materials

- References to reports and publications and other machine readable documentation.
-

Understanding Statistical Metadata

What is the current status of the DDI (and other statistical metadata standards) and what is the next step forward?

Different approaches to understanding:

What is it **for**?

- Statistical metadata has no value in itself, it is just a mean to an end. Its progress should be measured by the extent that it facilitates social research.

What is it **like**?

- Anything familiar we can relate it to? **Form of communication** might be a good choice.
-

What Do Social Researchers Want?

- # **Discover** available datasets (globally, not just in their own country) and related research literature.
 - # **Understand** in detail the origin, methodology and structure of datasets (social sciences datasets are modest in size but big in complexity).
 - # **Compare** and **Link** data from different sources.
 - # **Model** the social phenomena underlying the data.
 - # **Publish** their findings with all the **supporting evidence** (no 'iceberg' publishing) and **Reproduce** published results.
 - # **Connect** to other experts and **Share** informal comments and advice.
 - # **Enforce** confidentiality and intellectual property rights while maintaining accuracy and access to data sources.
 - # ... and more
-

The Simplest Form of Human Language: Pidgin.

Limited Vocabulary

“Proximity” Grammar

Simple Prose Only

Me Tarzan,
You Jane!



What Kind of Language is the DDI?

Could it be that a state-of-the-art metadata language as the DDI is really just a primitive pidgin language?

Actually it even worse than that:

- Pidgins have a very restricted grammar (grammar: “set of rules to create new meanings from existing ones”). The DDI has **none**.
 - Pidgins have a limited vocabulary. The DDI vocabulary is not only limited, it is **non-extendable**.
 - Pidgins are limited to simple prose. The DDI has a structure that it as rigid as a **tax form**.
-

Is the DDI Any Good for e-Social Science?

If the DDI is so extremely primitive, is it useful at all? It turns out that a little bit of metadata can go a long way (just ask Tarzan and Jane).

Certainly it can support at least some key use cases:

- **Discover** available datasets.
- **Understand** in detail the origin, methodology and structure of datasets.

What it does not (properly) support is pretty much all the rest: comparing, sharing informal advice, modelling, reproducing published results, etc.

Language Evolution: From Pidgin to Creole

Limited Vocabulary

“Proximity” Grammar

Simple Prose Only

Me Tarzan,
You Jane!

Shall I
Compare
Thee to A
Summer’s
Day?

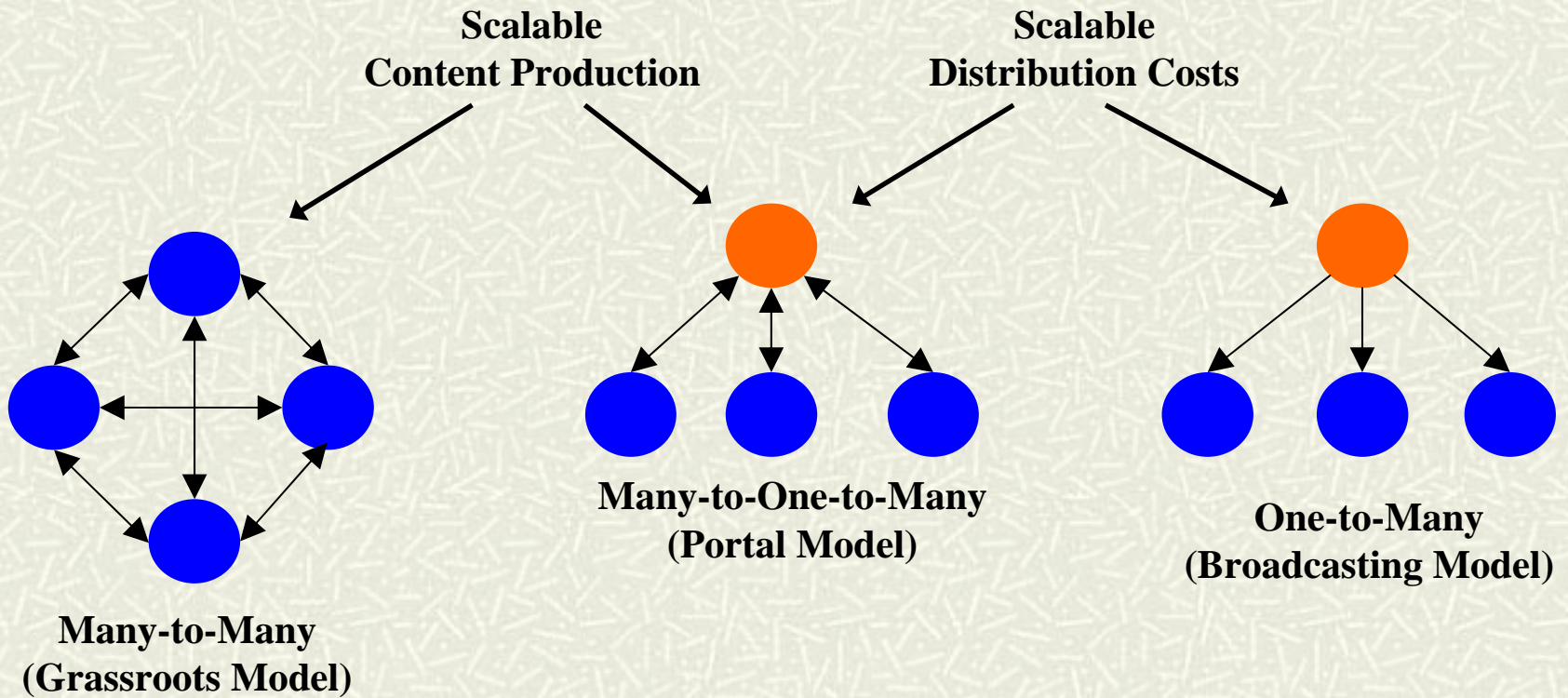
All Literary Forms

Grammar

Large Vocabulary



Models of Communication Or Where Will Statistical Metadata Come From?



Something to Debate ...

Statistical metadata is here and it is already changing the way people **locate** and **make sense** of data but it does not yet support most use cases of interest to social scientist. What we will need to move forward is:

- ✦ **Grammar**, a standard Semantic infrastructure (e.g. as provided by the Semantic Web):
 - Semantic extendibility
 - Ability of integrating (merging and overriding) descriptions from different sources
 - ✦ Large **Vocabulary**, by integrating different flavours of metadata:
 - Unique identifiers for data and research literature
 - Statistical data metadata (full life cycle)
 - Ontologies, Thesauri and Classifications (and mappings among them).
 - Statistical processing metadata
 - “Secondary metadata”: annotations, quality assessment, links to research literature.
 - Experts metadata (FOAF)
 - ✦ **Scalable** model of statistical metadata production
 - From archives to portals: integrate and republish both primary and secondary metadata.
-

“A good speech should be like a woman’s skirt: short enough to arouse interest but long enough to cover the essentials.” R. Knox
