MSc in Official Statistics Statistical Computing: Statistical Metadata

Andrew Westlake

Survey & Statistical Computing

63 Ridge Road, London N8 9NP, UK

+44 (0) 20 8374 4723

AJW@SaSC.co.uk (E-Mail)

www.SaSC.co.uk

Outline

- What is Metadata?
- What is Statistical Metadata?
- What is it for, how is it used?
- Different ways of thinking about and classifying Meta-data
- Types and Applications of Metadata
- Examples of Standards for Metadata
- Where does it come from?

What is Metadata?

- Data *about* Data
  - » Not data about the subjects of the database (the instances in the design)
  - » Information about the design, or anything connected with it
  - » Can see this as a level of abstraction
- Broad Range
  - » Technical descriptions of files, through to
  - » Abstract specification of concepts
- Term used by Sundgren in 1973

What is Statistical Metadata?

- Any information that is needed by people or systems to make proper and correct use of the real statistical data when:
  - » Capturing
  - » Reading
  - » Processing
  - » Presenting
  - » Analysing
  - » Interpreting
  - » Exchanging
  - » Searching
  - » Browsing
- Broad definition
  - » Anything that might influence or control the way in which the core information is used by people or software

What is Statistical Metadata?

- Includes
  - » file descriptions
  - » codebooks
  - » processing details
  - » sample designs
  - » fieldwork reports
  - » conceptual motivations, terminology
- Use
  - » Can be used informally by people who read it (and use it to affect the way they work with or interpret information)
  - » And formally by software to guide and control the way information is processed
- Metadata is Data
  - » Has structure and associated functionality
- Processes can generate metadata

What is Metadata for?

- To supply human readable information that facilitates the finding and interpretation of electronic data in a complex environment
- To supply machine processable data that facilitates the exchange of information between systems and the processing of data within a system

Joanne Lamb, Metanet

Why is Metadata important?

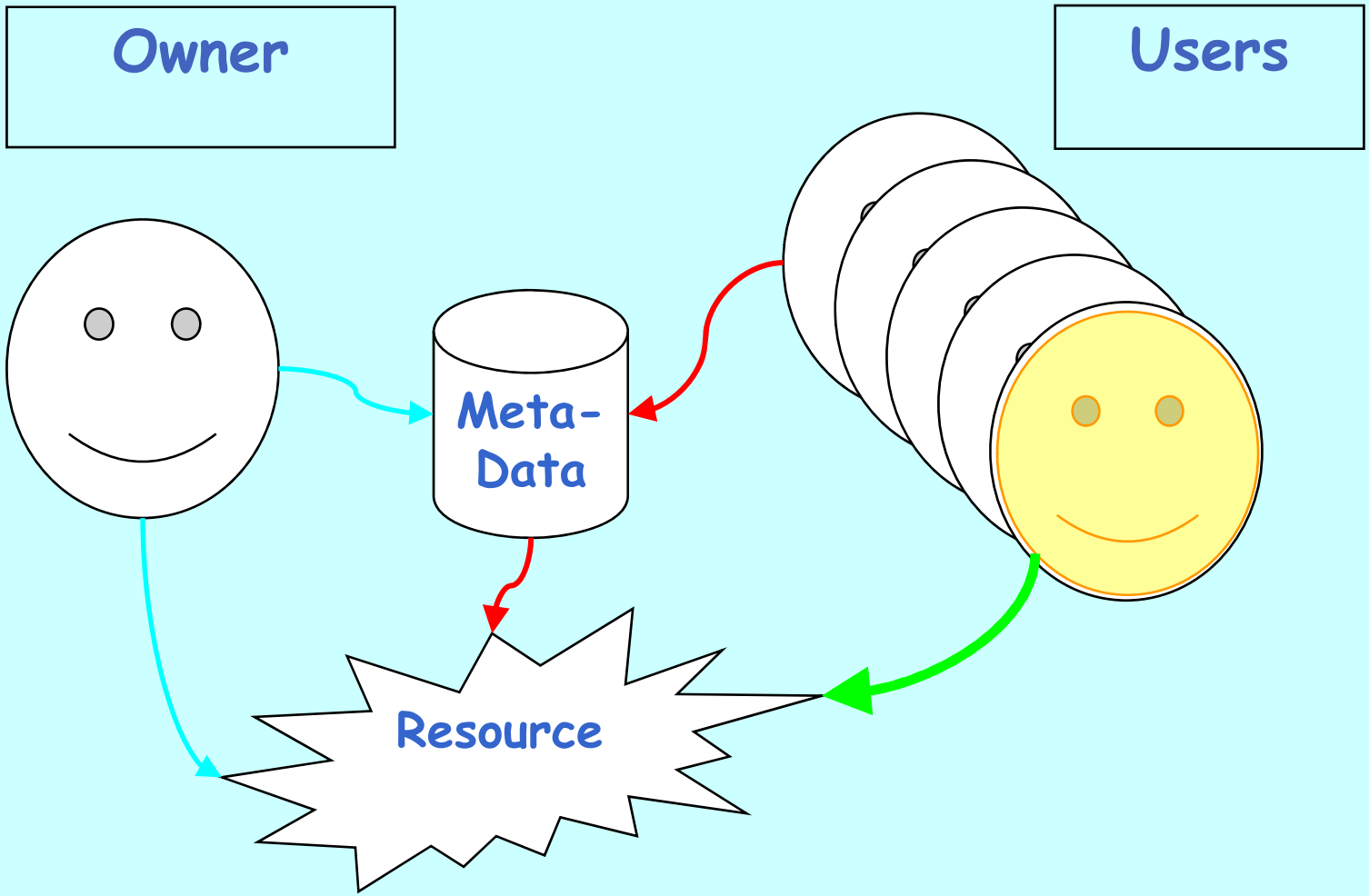
- Sharing data
  - » 'In my head' is not good enough
- Archiving
  - » Secondary users need good information
- Discovery
  - » Does data exist that can help me answer a problem?
- Automation
  - » Parameterisation of standardised processes

See RSS Archive – Preserving and Sharing

General Metadata

- Can relate to any resource that is shared
  - » Impact of Dublin Core, for Web Discovery
- *Metadata is the information that the owner of a resource needs to make available to potential users of that resource, so that they can use it correctly*
  - » Not new information, but not accessible to users
  - » May be in heads, on paper, in proprietary software
- In statistics, can apply beyond micro datasets
  - » Aggregate data
  - » Survey motivation and design
  - » Derivations and transformations
  - » Data integration and statistical modelling
- Various standards and proposals, discussed later
- Focus on **using Metadata** to support the *use* of a resource

Summary: Metadata links Owner to User



Levels of Abstraction

- Helpful to think at different levels
  - » Depends on the purpose and objective
  - » Related to ‘top-down’ vs ‘bottom-up’ views of systems
  - » Applies to multiple aspects
    - Different dimensions, each with multiple levels
- Often choose 4 levels for modelling
  - » 0 The real thing, no abstraction
  - » 1 The specification (model) of the real thing
  - » 2 The specification of what is allowed in a model
  - » 3 The specification of what is allowed in a specification
    - (Level 3 is not used much)

Levels of Contextual Abstraction

- Actual files and databases
- Agreed standards for actual data
  - » NACE, ICD, GesMes, SDMX
- Conceptual structures
  - » Industry classification, Disease classification, Data Exchange structures
- Terminology
  - » In a Thesaurus, with Concepts, Terms and Relationships
    - Synonym, type of, broader than, etc

Variables in Statistics

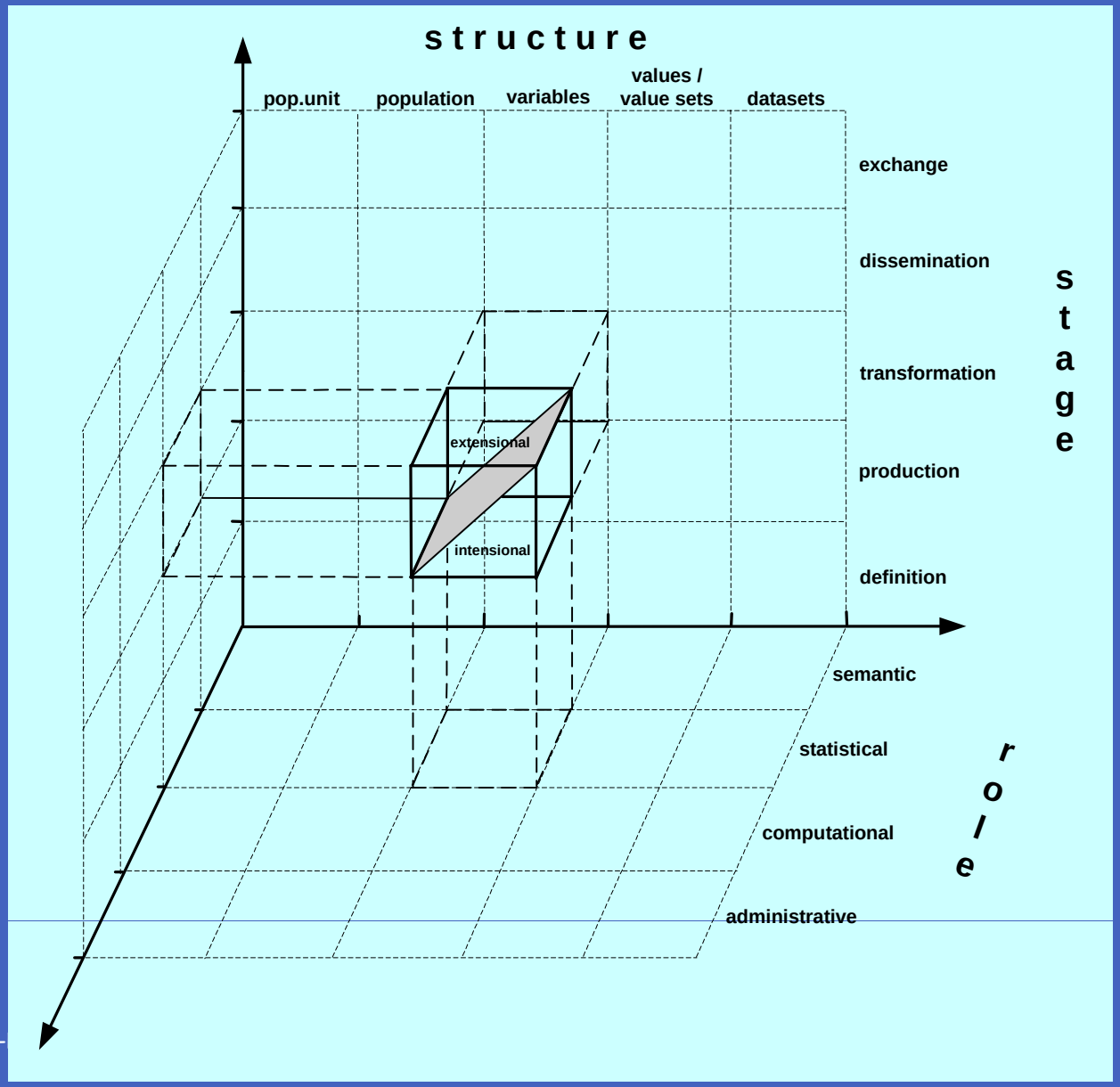
- E.g Employment Measurements
  - » 0 - Instances of employment for respondents within data files
  - » 1 - Agreed coding standards for Employment
  - » 2 - Concepts: what is Employment?
- Different purposes
  - » Concepts for Discovery and understanding
  - » Standards for exchange and comparability, presentation and understanding
  - » Codes in data for analysis
- Levels 1 and 2 are Metadata

Levels of Structural Abstraction

Level 4 **Tools and Models (Metamodel)** Structural scheme
Level 3 **Statistical Ontology** Structure model (Subject matter scheme)
Level 2 **Statistical Metadata (2nd Order Data)** Subject matter model (Data scheme)
Level 1 **(Statistical) data (1st Order Data)** Data model

Grossmann, Metanet

Four dimensions for classifying Metadata



Grossmann, Metanet

Statistical Structure

- What statistical structure is the metadata about?
  - » Population
  - » Population element
  - » Dataset
  - » Variable
  - » Value set
  - » Aggregation
  - » Conclusion
  - » ...

Processing Stage

- What stage does the metadata apply to?
  - » Design
  - » Data collection
  - » Data Processing
  - » Transformation and Analysis
  - » Dissemination
  - » Exchange
  - » ...

Role

- What is the metadata for?
  - » Semantics (understanding meaning and purpose)
  - » Statistical Validity (ensuring valid operations)
  - » Computation (ensuring and recording correct processing)
  - » Administration (recording ownership and responsibility)
  - » Discovery (providing access for secondary users of the statistical data)
  - » ...

Level of Operational Abstraction

- Intentional / Abstract
  - The purpose and intension for doing things a particular way
  - Explanations of why particular decisions were taken
  - Usually textual
  - Written by people, in advance of thing described
  - May come from (or be references to) collections of standards or guidelines
  - E.g. the reason or concept behind a particular variable
- Extensional / Concrete
  - Actual things done
  - Can flow from or be captured by systems/software that supports processes
  - E.g. the actual question, definition and coding of a variable

Metadata Functionality

- Access
  - » Display for people, read for systems
  - » Depends on Purpose, Structure and Context
- Linking
  - » To and from the thing it's about, including other metadata
- Languages
  - » Objects are independent of language, but their names and descriptions can be available in multiple languages
    - Not all concepts translate directly
- Versions
  - » Things change (e.g. classification revisions)
    - Data is coded according to a version
    - Versions must be accessible, and the **use** of an object must include the version

Discovery through Metadata

- Generic descriptions of subjects
  - » Population, Classifications, Measures
  - » Linked to concept definitions for searching
- Specific topics of a dataset or summary
  - » Formal definitions of standard components
    selection rules, standard classifications, measure types
  - » Specific descriptions of substantive content
    source variable definitions, questionnaire structure, etc
- Accessibility
  - » Metadata must be available to search engines and users
  - » Non-specialists do not approach through standard terms or structure

Metadata Standards

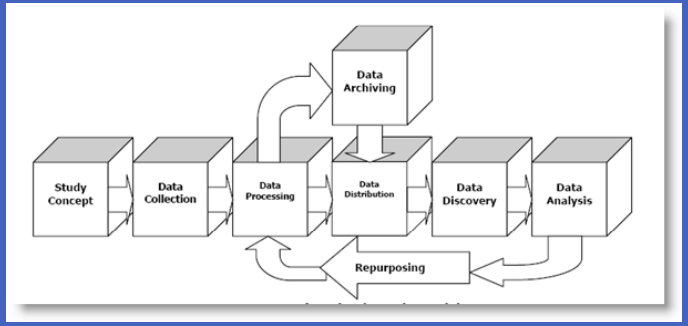
- Specification of structure and purpose (including semantics) of different types of metadata, of various types
- Lots of proposals
  - » Many omit purpose and semantics, just structure
- Generally have a limited view
- Useful to relate to the four dimensions

Triple-s

- Structure
  - » Describes a Dataset
- Stage
  - » Exchange
- Role
  - » Computation
- Level
  - » Extensional – describes fields and codes in a dataset

Data Documentation Initiative - DDI 3

- Structure
  - » Comprehensive - Describes a collection of Studies from conception to analysis
- Stage
  - » Design through to archiving and exchange
- Role
  - » Administration, semantics, computation
- Level
  - » Intentional and extensional



ISO 11179

- Structure
  - » Data Element
- Stage
  - » Design
- Role
  - » Computation, Discovery
    - Specification of data elements in a Repository
- Level
  - » Extensional

Neufchatel Terminological Model

- Structure
  - » Classifications
- Stage
  - » Collection, Dissemination
- Role
  - » Computation, Administration, Discovery
- Level
  - » Extensional (codes, labels and mappings) and Intentional (responsibilities, sources, case law)

SDMX – Statistical Data & Meta-data Exchange

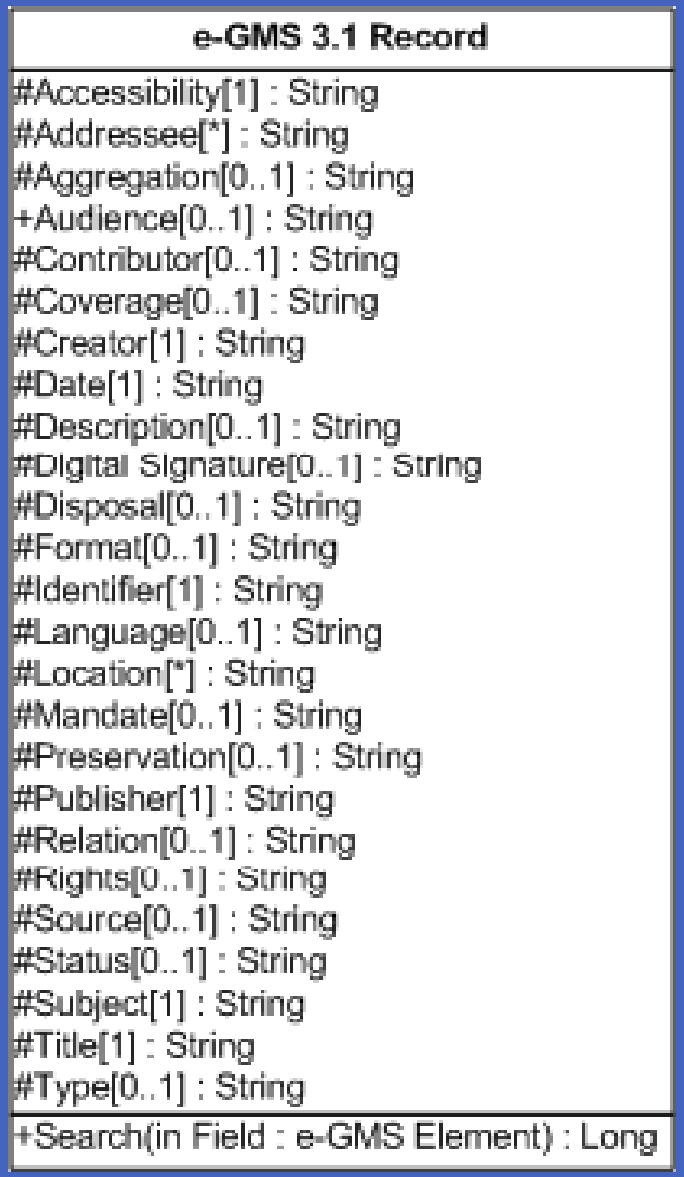
- Structure
  - » Aggregate data and time series
- Stage
  - » Exchange, Dissemination
- Role
  - » Administration, Discovery, Transfer
- Level
  - » Mostly Extensional

Well-resourced with big players (IMF, World Bank, Eurostat), limited objectives, links to DDI 3

e-GMS - Government Metadata Standard

- Structure
  - » Any resource (not just statistics)
- Stage
  - » Dissemination
- Role
  - » Discovery
- Level
  - » Intentional

Derived from the Dublin Core standard for discovery metadata



Unified Metainformation Architecture in Statistics (UMAS)

- Structure
  - » All statistical objects and processes
- Stage
  - » All
- Role
  - » All
- Level
  - » Both

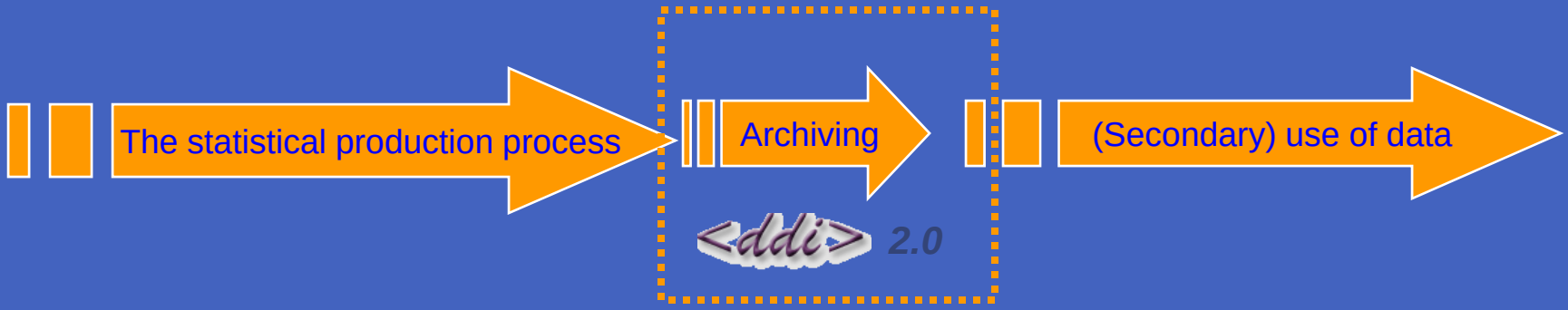
This is a generic model of meta-information for all statistical processes

StatModel from OPUS

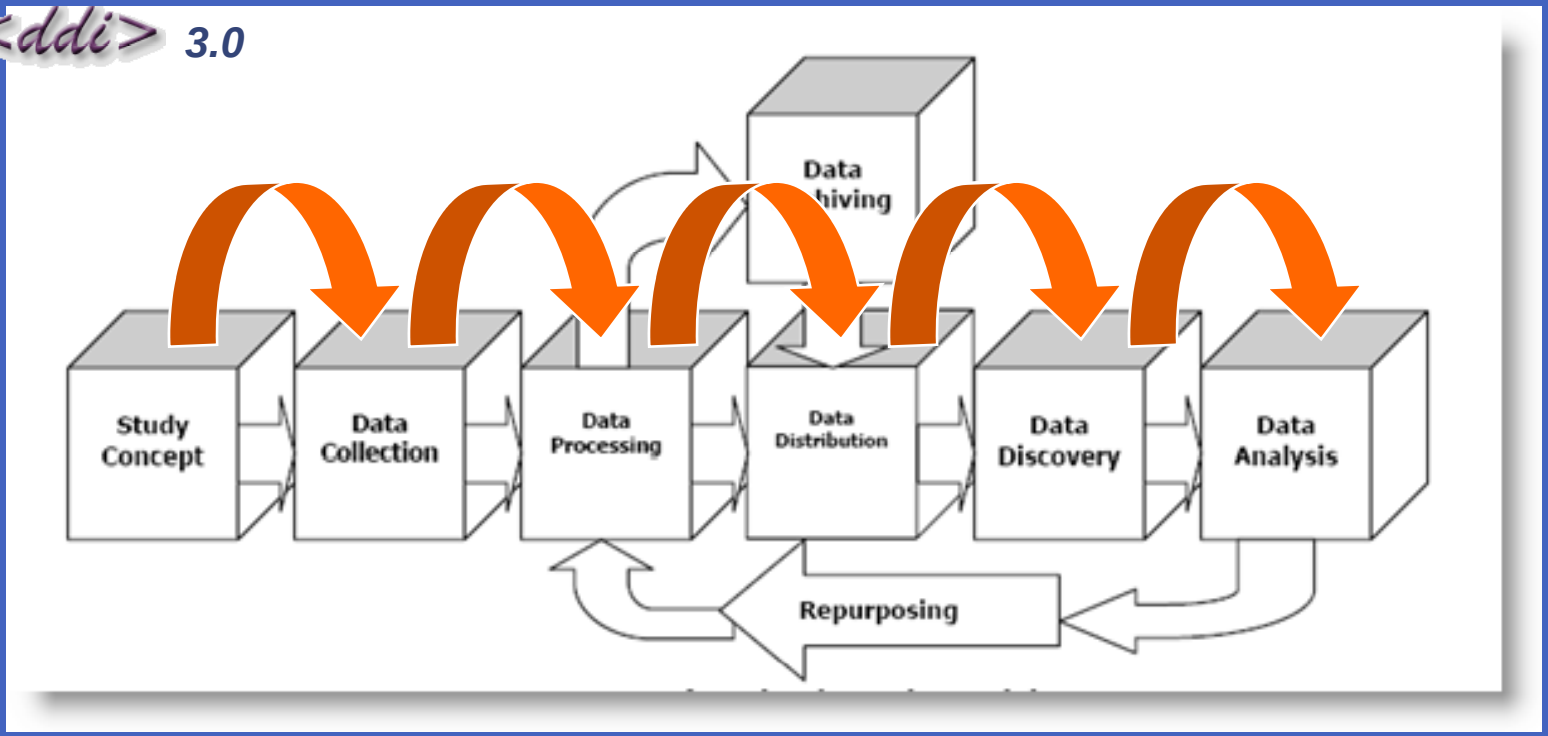
- Structure
  - » Statistical models and related Variables, Parameters and Data
- Stage
  - » Specification, Fitting (including use of Data), Results
- Role
  - » Discovery, Presentation, Validation, Exchange (of models)
- Level
  - » Intentional (Definitions, Motivations, Sources) and Extensional (Specifications, Processes)

This proposal goes far beyond usual concepts of Statistical Metadata

DDI 3.0 design goals – life-cycle model



ddi 3.0



Benefits of the DDI Approach

- **Interoperability**
  - » Codebooks marked up using the DDI specification can be exchanged and transported seamlessly, and applications can be written to work with these homogeneous documents.
- **Richer content**
  - » The DDI was designed to encourage the use of a comprehensive set of elements to describe social science datasets as completely and as thoroughly as possible, thereby providing the potential data analyst with broader knowledge about a given collection.
- **Single document - multiple purposes**
  - » A DDI codebook contains all of the information necessary to produce several different types of output, including, for example, a traditional social science codebook, a bibliographic record, or SAS/SPSS/Stata data definition statements. Thus, the document may be repurposed for different needs and applications. Changes made to the core document will be passed along to any output generated.
- **On-line subsetting and analysis**
  - » Because the DDI markup extends down to the variable level and provides a standard uniform structure and content for variables, DDI documents are easily imported into on-line analysis systems, rendering datasets more readily usable for a wider audience.
- **Precision in searching**
  - » Since each of the elements in a DDI-compliant codebook is tagged in a specific way, field-specific searches across documents and studies are enabled. For example, a library of DDI codebooks could be searched to identify datasets covering protest demonstrations during the 1960s in specific states or countries.

Capturing Metadata

- It's boring
- If it is not done well it is not used, so is not worthwhile
- Experience from SCB Doc
- Wherever possible, capture from other processes
  - » Intentional metadata from project justifications, standard designs, …
  - » Extensional metadata from metadata-aware systems, for design, data entry, manipulation, …
- Contextual linking will generally need to be manual
- DDI and SDMX developing Tools and Components

What's New?

- Internet
  - » Vastly improved ease and scope of accessibility
  - » Need to focus discovery processes and provide functionality and access
- XML (eXtended Markup Language)
  - » Representation and exchange of complex data structures
  - » But what are the structures and semantics?
- Money
  - » Data Warehouses, OLAP
  - » Commercial pressures for Standardisation

What is happening?

- Commercial developments
  - » Big initiatives all about structure, e.g. CWM, eb-XML
    - Valuable, but not enough for statistics
  - » Survey data systems
    - Triple-S, SPSS MR Dimensions, … Smart data entry software (QEDML, Askia, …)
- Statistical Office Initiatives
  - » Very bottom-up
    - Coding, exchange formats, Often focussed on aggregates (SDMX)
  - » Standardised documentation (Dublin Core, e-GMS)
    - Little structure
  - » Neufchatel and related initiatives for Classifications, …
  - » ONS Modernisation, Repository including Meta-data …
- Research Projects
  - » Survey data
    - **DDI Codebook**, StatModel, …
  - » Eurostat (plus partners) – all dead!
    - Statistical systems using metadata – Nesstar, Metaware, Tadeq
    - Exploration of functionality – Idaresa, Addsia, IMIM , Codacmos
    - Development of Standards and Structures – IQML, MetaNet

What next?

- We need standards about structure and functionality for statistical metadata
- Build metadata into system designs, as part of statistical data structures, at whatever levels are feasible
- Capture metadata wherever possible, with as much structure as possible
- Think about the other people who will need to know about the data, or what you did

Summary

- Statistical Meta-data is important
  - » Especially for secondary or subsequent users
  - » Simplify processing by using prior specifications
  - » Improving quality through precise specification and explanations
- Can apply to all aspects of statistical processing
  - » Most current use focussed on data description and exchange
- Lots of proposals for standards
  - » Few actually used
  - » Triple-s, DDI, SDMX all have merit
- Small market (cf. RDBMS, for example) so little investment
  - » Encourage more initiatives and use! Build into other projects!