WP. 7
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Luxembourg, 9-11 April 2008)

Topic 2 (ii) Metadata concepts, standards, models and registries

CLASSIFICATIONS OF STATISTICAL METADATA

Submitted by Statistics Sweden[1]

Statistical metadata may be classified in a number of different dimensions, more or less orthogonal to each other. For example, statistical metadata may be classified according to

- **who** needs the metadata, and **why**, for which purposes (e.g. exploratory and explanatory)
- **what** the metadata inform about: metadata objects (attachment objects) and metadata variables
- **how** the metadata are structured and formalised (or not)
- **where** the metadata come from (source processes) and where they go (use processes)

The paper defines different classifications of statistical metadata, and it also discusses why and how it is useful to consider these classifications when statistical agencies are developing, operating, and evaluating statistical systems.

I. WHY CLASSIFY STATISTICAL METADATA?

1. Classifications play important roles both in everyday communication and in scientific work on different levels (description, analysis, explanation, prediction). From the day we are born we start categorising the perceptions we receive through our senses, in order to get an organised and meaningful understanding of the world around us, and to be able to communicate with our fellow human beings, guided by the classifications implied by the language used in the community where we live. The classification of plants and animals by Carl von Linné and the periodical system set up by Mendeleiev are famous example of classifications of great importance in science. Similarly classifications of statistical metadata may help us to communicate about metadata in a systematic and efficient way, and to develop theories, methods, and tools for the organisation and management of statistical metadata.

II. ONE-DIMENSIONAL VERSUS MULTIDIMENSIONAL CLASSIFICATIONS

2. Standard classifications used in scientific work are often one-dimensional in the sense that they consist of a single, multi-level hierarchy, where the concepts on a certain level in the hierarchy divide the domain of discourse into an exhaustive and mutually exclusive subdomains, or classes. The classification criteria used for mapping each object in the domain of discourse unambiguously into one and only class may be quite complex and may in themselves combine several different dimensions. In statistics production such classifications are quite common in connection with social and economic statistics.

[1] Prepared by Bo Sundgren (bo.sundgren@scb.se).

3. Another approach to classifications is to keep the classification criteria as simple and one-dimensional as possible and to achieve more complex classifications by explicitly combining a number of one-dimensional classifications into a Cartesian multidimensional classification. I will use this approach in this paper. The result will be a multidimensional classification of statistical metadata, where different projections of the multidimensional structure may be used for different purposes.

III. A MULTIDIMENSIONAL CLASSIFICATION OF STATISTICAL METADATA

4. I will use the following dimensions in my proposal for a multidimensional classification of statistical metadata:

(a) Statistical metadata classified by usage/purpose: **who** needs the metadata, and **why**, for which purposes?
(b) Statistical metadata classified by contents, or so-called attachment objects, the objects which they describe: **what** do the metadata inform about?
(c) Statistical metadata classified by sources: **where** do the metadata come from, in which processes are they born or created?
(d) Statistical metadata classified by form: **how** are the metadata represented and organised?

IV. STATISTICAL METADATA BY USAGE/PURPOSE – METADATA USE PROCESS

5. All processes in a statistical system[2] use (and produce) metadata. Thus we may use a model for the processes of a statistical system as a basis for a classification of statistical metadata in this dimension. Figures 1-4 illustrate such a process-oriented model from various perspectives.

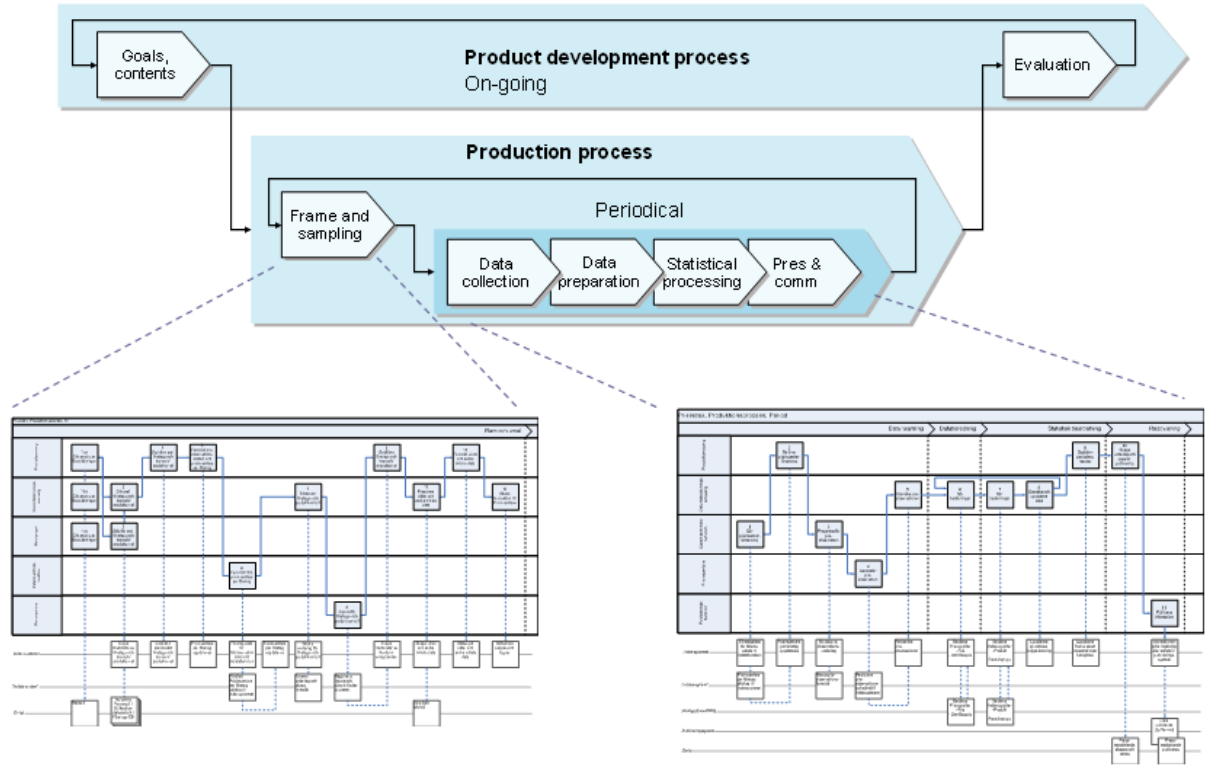

Figure 1. *A process-oriented model of statistics production.*

[2] "Statistical system" is a generic term. For example, it may be an international statistical system, a national statistical system, a subsystem or component of a statistical system, a standardised statistical system, or an individual application; it may be a traditional end-to-end survey, a register system, or an analytical system (index, accounts, etc).

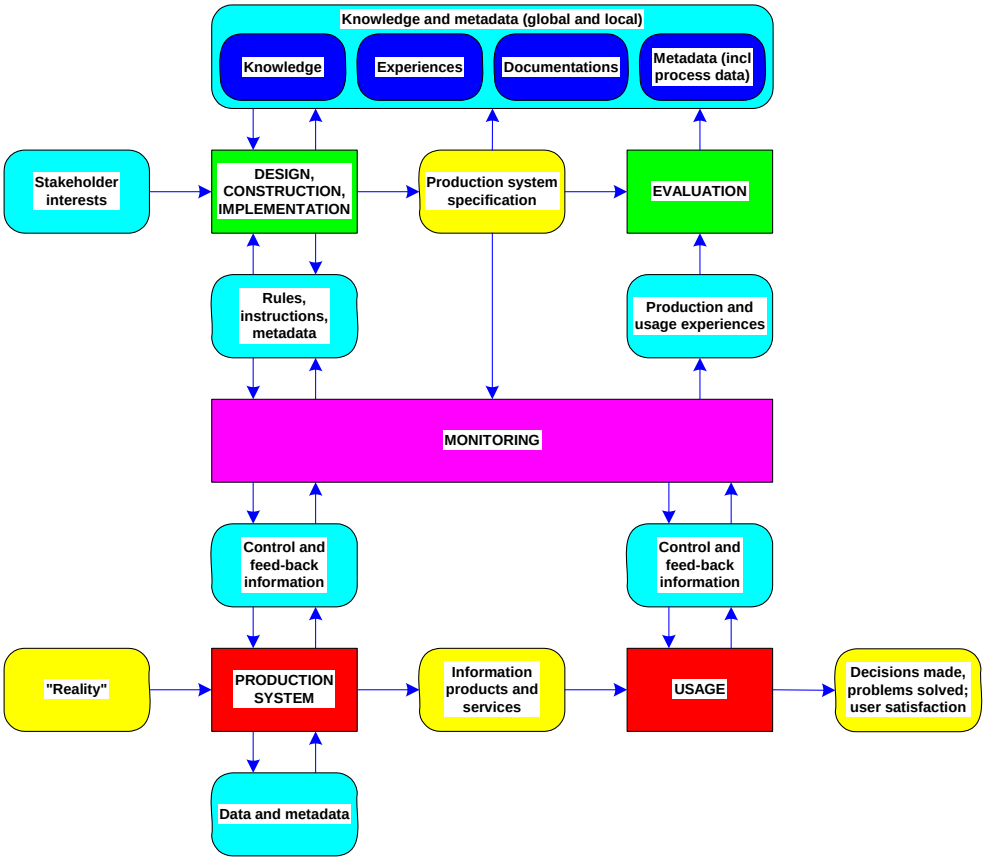

Figure 2. *Development, operation/monitoring, and evaluation of a statistical production system.*

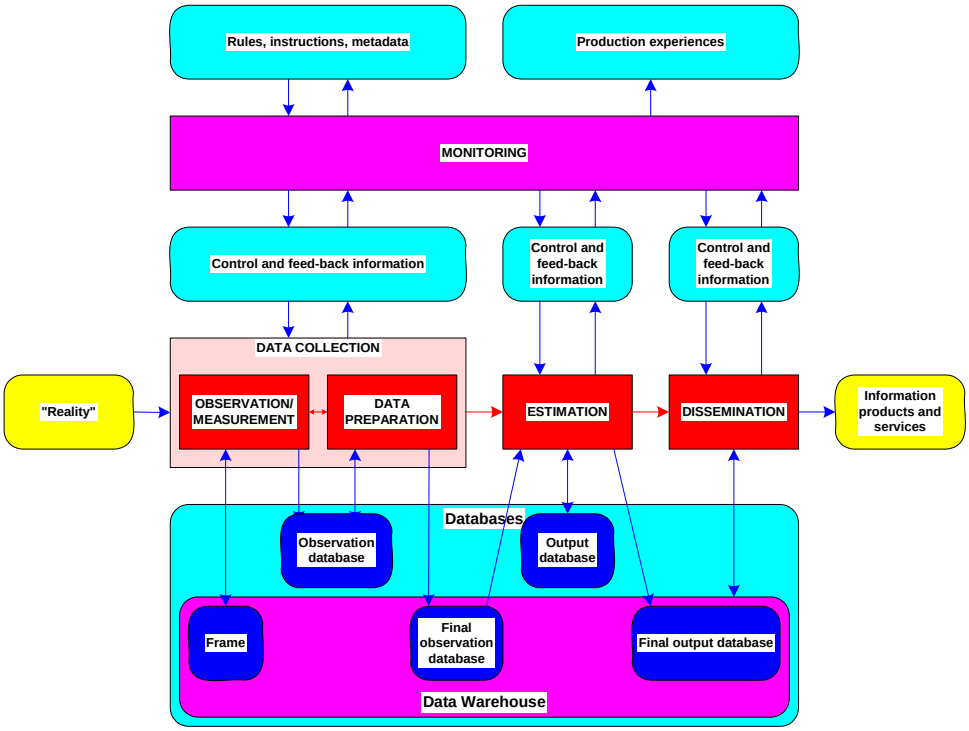

Figure 3. *More detailed view of the basic operations in a database-oriented statistical production system.*

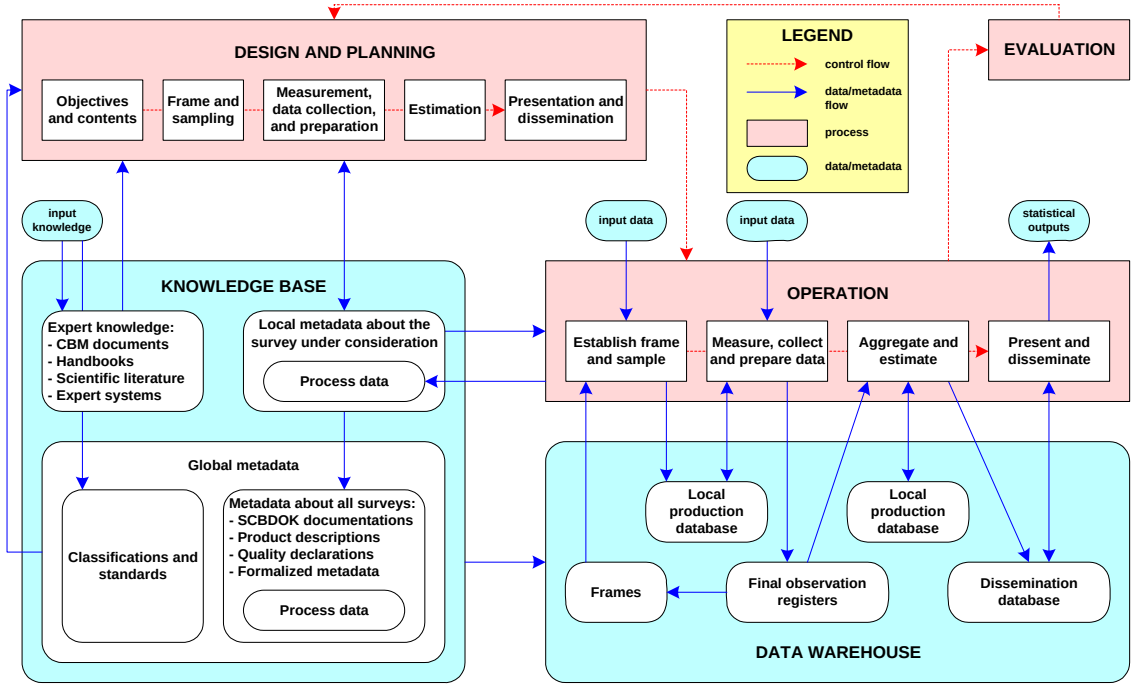

Figure 4. *Statistics production with integrated data and metadata management.*

6. Table 1 (with its three subtables 1a, 1b, 1c) gives a rough specification of the metadata needs of each one of the major processes and subprocesses in a statistical system. It also provides (in the third column) information about which metadata are typically produced by the respective process/subprocess. This information will be used later in the paper.

(RE)DEVELOPMENT	METADATA NEEDED	METADATA PRODUCED
(Re)develop a statistical system Provide information and knowledge helping developers:	Methodological knowledge, experiences and process data[3] from this and other systems	
- Plan: explore, specify, and reconcile the needs of customers and other stakeholders (e.g. respondents, staff, funders) as regards contents, qualities, functionalities, costs	- articulated requirements from stakeholders	- reconciled requirement specification
- Design: design a system that satisfies customers and other stakeholders in an optimal way	- reconciled requirement specification	- system specification
- Build, test, implement: build, test, and implement a working system according to the system specification	- approved system specification	- documentation of the working system, including instructions and training material - process-driving metadata

Table 1a. *Metadata needs (and metadata products) of the different subprocesses in the (re)development of a statistical system.*

[3] ”Process data” are metadata about a process, generated by the process itself during its execution, e.g. non-response rates generated by a data collection process.

OPERATION & MONITORING	METADATA NEEDED	METADATA PRODUCED
Operation and monitoring of a statistical system Assist and control the operation and monitoring of the system during the following phases:	- documentation of the working system, including instructions and training material - process-driving metadata - dynamically generated process data feedback	- process data signalling about performance and potential problems - metadata about derived and transformed data
- Data collection & preparation[4]	- questionnaire and instructions - data preparation rules - process data signalling suspicious errors and informing about their potential impact on important estimates	
- Aggregation and estimation	- aggregation and estimation rules - process data signalling problems	
- Presentation and communication, of statistical outputs[5], by producers	- publishing and communication rules - confidentiality protection rules - process data signalling problems	
- Use of statistical outputs: help users find, access, and use statistical data and services relevant for them … search for relevant outputs … access to selected outputs … interpretation and analysis	- overviews and explanations of available data and services: … graphs, hierarchies, and search systems[6] … more detailed explanations[7], invoked by the user when needed - links to requested data and services - warning signals - reports and experiences by other users	

Table 1b. *Metadata needs (and metadata products) of the different subprocesses in the operation and monitoring of a statistical system.*

[4] Including coding, data editing, computation of derived objects and variables, confidentiality protection.
[5] Products and services, e.g. microdata and documentation, statistical tables and explanations, analytical products, helpdesk.
[6] Exploratory metadata.
[7] Explanatory metadata.

EVALUATION	METADATA NEEDED	METADATA PRODUCED
Evaluation of a statistical system Managers on different levels to be alerted about improvement potentials	Methodological knowledge, experiences and process data from this system and other systems	- Evaluation reports
- Qualities, costs, response burdens	- Objectively observed or estimated measures of quality and cost components, including the costs of respondents[8]	
- Customer satisfaction and complaints	- Structured, but often more subjective estimates of customer satisfaction, verbal complaints and comments by customers, manifested customer behaviour (e.g. on a website)	
- Respondent satisfaction and complaints	- Structured, but often more subjective estimates of respondent satisfaction, verbal complaints and comments by respondents, manifested respondent behaviour (e.g. non-response, errors)	
- Feedback from other stakeholders	- Positive and negative reactions from funders, politicians, journalists, and the public at large	

Table 1c. *Metadata needs (and metadata products) of the different subprocesses in the evaluation of a statistical system.*

V. STATISTICAL METADATA BY CONTENTS – METADATA ATTACHMENT OBJECTS

7. Table 2 (with its four subtables 2a, 2b, 2c, 2d) gives a rough specification of the metadata contents of each one of the major types of metadata objects in a statistical system, the so-called metadata attachment objects, objects that are informed about by means of metadata. It also provides (in the third column) information about which metadata are typically produced by the respective process/subprocess. It also indicates (in the third column) how different metadata objects are typically linked to each other in the corporate metadata system of a statistical agency.

8. The major types of metadata attachment objects are, according to the proposed classification in this dimension:

- system/subsystem/process
- input/output data (statistical data – metadata – process data)
- conceptual object
    - observation characteristic
        - population
        - variable
        - value set
        - value
    - target characteristic
        - population
        - variable

[8] Ideally the qualities should be measured and structured according to some standard scheme for quality declaration of statistics, and the costs should be measured and structured according to some standard accounting system.

        - value set
        - value
        - statistical measure
- instrumental resource
    - data collection – Metadata collection – Process data collection
    - software
    - documentation
    - methods
    - knowledge & experiences
    - persons & organisations
    - equipment & localities
    - financial resources

SYSTEM/SUBSYSTEM/PROCESS	METADATA CONTENTS	LINKS TO
Statistical system/subsystem - Standard system (e.g. SNA) - Application (of standard system) - Generic system (e.g. repetitive) - Execution (of generic system) - Census, (sample) survey - Event-based system (e.g. register)	- Administrative information - Frequency (if any) - Time series (if any) - Description - Experiences - Evaluation data	- More detailed description - Evaluation reports - Generic system(s), if any - Execution(s), if any - Subsystems and supersystems - Processes and subprocesses - Data sets (input, intermediary, output) … populations/subpopulations … classification variables and value sets … summation variables and value sets … estimates (if any) and value sets - Metadata sets - Process data sets - Instrumental resources, including … data/metadata infrastructure … methods and software … technical tools and equipment … organisation and staff
Process/subprocess - Standard process - Application (of standard process) - Generic process (e.g. repetitive) - Execution (of generic process) - (Re)development of a statistical system … Plan … Design … Build, test, implement - Data collection - Data preparation - Aggregation and estimation - Publishing and communication	- Administrative information - Description - Experiences - Evaluation data	- More detailed description - Evaluation reports - Generic system(s), if any - Execution(s), if any - Subsystems and supersystems - Processes and subprocesses - Data sets (input, intermediary, output) - Metadata sets, e.g. … system/process driving parameters … derived, generated - Process data sets, e.g. … control parameters (feed-back) … generated and stored - Instrumental resources, including … data/metadata infrastructure … methods and software … technical tools and equipment … organisation and staff

Table 2a. *Statistical metadata by attachment object and contents.*
Part a: metadata about systems, subsystems, and processes.

INPUT/OUTPUT DATA	METADATA CONTENTS	LINKS TO
Statistical data collection (input/output) - Data collection instance - Data collection series[9] - Event-driven, continuously updated data collection (reflecting current situation) - Event-driven, continuously updated data collection (reflecting current situation and complete history) Production stage - RawData - FinalMicro - FinalMacro - EndProduct	- Administrative information - General description - Contents … observation characteristics … statistical target characteristics - Quality … relevance … accuracy … timeliness … comparability & coherence … availability - Technical information	- Systems and processes - Conceptual objects - Instrumental resources
Metadata (input/output) - System/process driving metadata - Passive metadata	- Administrative information - General description - Contents - Technical information	- Systems and processes - Conceptual objects - Instrumental resources
Process data (input/output) - Control parameters - Generated process data	- Administrative information - General description - Contents - Technical information	- Systems and processes - Conceptual objects - Instrumental resources

Table 2b. *Statistical metadata by attachment object and contents.*
Part b: metadata about statistical data, metadata, and process data.

CONCEPTUAL OBJECTS	METADATA CONTENTS	LINKS TO
Observation characteristic - generic/instance • Population • Variable • Value set • Value	- Name - Description - Definition - Code - Quality information	- Data sets, metadata, process data - Systems and processes - Registers - Questionnaires and questions - Classifications - More detailed documentation - Feedback information - Evaluation reports
Statistical target characteristic - generic/instance • Population • Variable • Value set • Value • Statistical measure	- Name - Description - Definition - Code - Quality information	- Data sets, metadata, process data - Systems and processes - Registers - Questionnaires and questions - Classifications - More detailed documentation - Feedback information - Evaluation reports

Table 2c. *Statistical metadata by attachment object and contents.*
Part c: metadata about conceptual objects.

[9] A data collection series is a (time) series of data collection instance. The instances in a series have many properties in common, but typically there are minor differences between different instances in the same series, caused by minor changes in the design of the survey(s) behind the data collection.

INSTRUMENTAL RESOURCE	METADATA CONTENTS	LINKS TO
Instrumental resource (corporate infrastructure) instance/collective[10]	- Name - Description - Availability - Cost data - Use data - Evaluation data	- Owner - Operator - Authorized users - Systems using this resource - Documentation - Experience reports - Evaluation reports
Data collection - Register/frame - System of registers - Database - Data warehouse - Archive		
Metadata collection - Questionnaire library - Catalogue of variables - Classification - Classification database		
Process data collection - Transaction logs - Summarised process data		
Software - Software product - Software library		- Methods supported
Documentation - System/process documentation - Data/metadata documentation - Software documentation - Documentation library		- Documented object
Methods - Method - Methods library		- Knowledge item - Knowledge library - Supporting software
Knowledge & experiences - Knowledge item (article, website, etc) - Knowledge library		
Persons & organisations - Staff member - Customer - Supplier/provider - Person database	- Roles & responsibilities - Person/organisation data	- Administrative systems
Equipment & localities		- Administrative systems
Financial resources		- Administrative systems

Table 2d. *Statistical metadata by attachment object and contents.*
Part d: metadata about instrumental resources.

[10] A shared corporate resource may be a collective resource, e.g. an archive, a library, a repository, or a resource instance, e.g. a single register/frame.

VI. STATISTICAL METADATA BY SOURCE AND USAGE (FROM WHERE – TO WHERE)

9. Table 3 combines information from Table 1 and Table 2 so as to show the flow of different kinds of metadata between the different processes in a statistical system. The rows show which kinds of metadata are produced by a certain process[11], the source process of the respective kinds of metadata, and which processes are the users of the respective kinds of metadata. The columns in the table show which kinds of metadata are used by a certain process, and which are the respective source processes of these metadata.

10. As it stands here, table 3 is just a superficial sketch. The processes may be broken down into subprocesses on several levels, and the metadata contents in each cell may also be specified in much greater detail, and with a more clear indication of the metadata objects to which they are attached (like in Table 2).

From where – To where	**R&D**	**(RE)DEVELOP: Plan-Design-Build**	**OPERATION & MONITORING**	**EVALUATION**
R&D	General knowledge	General knowledge		General knowledge
(RE)DEVELOP: Plan-Design-Build	System doc	Requirement spec System spec System doc+instruc	System doc+instruc	System doc
OP & MONITORING	Process data	Process data	Process data	Process data
EVALUATION	Evaluation reports	Evaluation reports		Evaluation reports

Table 3. *Statistical metadata by source and usage: where do they come from, and where do they go?*

VII. STATISTICAL METADATA BY FORM – DEGREE OF STRUCTURING AND FORMALISATION

11. Statistical metadata may also be classified by form, e.g.

- coded – formatted – free text – other
- structured – semistructured – unstructured
- organisation: tightly integrated with data – linked to data – separate metadata repositories

VIII. REFERENCES

Sundgren, B. (2004a) *"Statistical systems – some fundamentals"*, Statistics Sweden.
Sundgren, B. (2004b) *"Designing and managing infrastructures"*, Statistics Sweden.

[11] We have added Research & Development (R&D) as a major process, the source of different types of "general knowledge", e.g. methodological knowledge, documented in textbooks, articles, websites, etc.