WP. 9
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Session 3: Topic III: National implementations of the GSBPM

How do end users of statistics want metadata?

Authors : Mogens Grosen Nielsen & Lars Thygesen, Statistics Denmark

1. Statistics Denmark is presently undertaking to integrate metadata systems, with a special emphasis on making metadata available to end users *with contents, and in a form that will facilitate end users in their business processes*, in situations where they might use or consider using official statistics. To this end, we have undertaken rather deep consultations with key user segments about their business processes. Rather than focusing on what we believe might be an ideal metadata model that would support all kinds of needs,, the emphasis is on trying to understand different kinds of users and how they can, or would be able to use statistical data, if we supplied them with metadata suited to their processes and purposes.

2. The approach is based on the following assumptions:

– We have moved from a relatively stable industrial society to a globalized knowledge society. Access to fast, relevant and reliable information plays a crucial role for many of our users. In the industrial society we used relatively fixed models to describe production, population, etc. This met most user needs. In the knowledge society, users face a growing internal and external complexity, where the traditional statistics such as yearbooks are not enough. Today we must for example provide information about complex interplay between economics, demography and geography, which is necessary in the organization of public services and investments. Our users must constantly increase their level of information to manage the increasing complexity. Solutions, models, etc., that we offer are therefore often dependent on the context. Understanding of users' business issues and processes must therefore go hand in hand with the development of metadata-models and information-retrieval-systems[1].

[1] Lars Thygesen & Bo Sundgren have explored new approaches to fulfill user needs. *Innovative approaches to turning statistics into knowledge*. Statistical Journal of the IAOS: Journal of the International Association for Official Statistics. Amsterdam 2009
Bo Sundgren provides a visualization of the statistical system as part of a giant feedback loop, with respondents and users etc. : *Towards a system of official statistics based on a Coherent combination of data sources.* Paper presented at the ESRA Conference in Lausanne 2011. The model illustrates that the organisation of statistical production can no longer be depicted within the traditional hierarchical

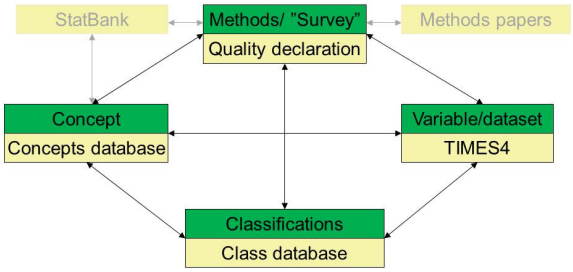
- Digital access to information, including Google's crucial role as the place where you get information, has led to a situation where knowledge is not restricted to a narrow group of scientists and experts. Fulfillment of user-needs from journalists, officials and citizens must be addressed as well. We therefore need to find methods that can support an effective and user-friendly information retrieval, not only for experts.
- Metadata must make users' development- and operational processes less complex. Therefore we need a system of metadata to facilitate external users' access to our statistical data.

3. The paper gives an account of consultations and our findings about the use of metadata It discusses how the needs of end users can be accommodated, taking into account existing metadata at Statistics Denmark as well as international recommendations and emerging standards such as GSIM. The paper will also deal with investigations into internal metadata needs in the organisation, but the authors feel that the needs of external users are key, and are not always taken sufficiently into account by statistical bureaus.

I. Background

4. Improving user oriented metadata is prominent in Statistics Denmark's Strategy 2015. Statistics Denmark and its services generally receive quite positive evaluations from end users, including paying customers, but during recent years end user feed-back on the performance of our services has repeatedly pointed to metadata as the weak point in our services., In our thinking about how to accomplish an improvement, we are making use the Common Metadata Framework as a guideline. We also take inspiration from best practice experiences from other countries, as we have learnt in METIS. One important point for improvement is the integration of the fragmented systems of metadata presently in operation, consisting of 4 elements: Quality declaration, concepts database, variables system and classifications database:

Figure 1:



5. Our hypothesis has been that we could improve the situation for users by making these elements inter-operate so that you could go from, e.g., one concept to all quality declarations of statistics where this concept is used. Users should be able to navigate across the whole space of metadata. We believed that this would be valuable even if it might not be possible to improve the basic metadata. We would also like a better integration between these metadata and our statistical database StatBank, holding all published statistics.

6. But how well do we know the preferences of our users? Statisticians and IT people may have very good ideas about what will benefit users, and there is a temptation to "just do it". Our idea was that it would be worthwhile to try to better understand how users wish to use statistics, and which role metadata could and should play in this process. This means trying to see metadata not just as a documentation that should be

---

administration model. We need more complex models in order to understand the complex interplay between statistical bureaus and their environments.

made because we want to have a good documentation, e.g. for archiving; but see them as part of a process that gives value to our users of today. Some documentation might prove to be of lesser interest to these users than we imagined. We want to give priority to developments that users would seem to really use in their business processes.

II. External users: Focus groups on documentation

7. In November 2010 Statistics Denmark carried out three focus groups on documentation with various groups users:
1. Intensive users, mostly government
2. Municipal and regional users
3. Education and the media

8. The aim was to identify the needs for documentation experienced by users when they used or tried to use statistics. This knowledge is input to Statistics Denmark's Steering Committee for Metadata and will be used to develop documentation in the coming years.

9. The focus groups each consisted of around 10-14 handpicked users, and were chaired by an external consultant, while a number of observers from Statistics Denmark were present but not allowed to speak unless they were asked (which was quite difficult). It gave a very lively and frank discussion, and many aspects of Statistics Denmark's services were constructively criticized. Observations were not limited to documentation but also included the statistics produced, revision policy, etc.

10. One objective was to test a prototype of coherent documentation that had been presented within Statistics Denmark: http://doktype.dst.dk/ . Would users see this as a step forward?

11. Some key conclusions regarding metadata were:

- The prototype and the underlying method with four interconnected components (quality declarations, concepts, variables and classifications) won strong support in all three groups. It was found that such a development of metadata would provide a good and logical approach to documentation.
- There is great need for metadata among intensive users, slightly less among municipalities, and the media say they have almost no needs of - they have no time to read anything
- There are essentially two ways that users search statistics: 1) Ad hoc, broad search, and 2) Deep search in a fixed subject area that you know well. By broad search, the documentation was important, but in all groups, there is a certain tendency to call a Statistics Denmark expert and ask.
- Documentation of data breaks and changes of definitions is insufficient. Each break should be mentioned and well explained in relation to the figures
- Statistics Denmark was encouraged to produce long time series which are corrected for breaks, otherwise each user must do it himself, which is not necessarily better, and in any event costs resources and gives different results.
- Revisions and revisions practices should be better documented
- Statistics Denmark must explain uncertainty and possible error sources, explaining what data can, respectively cannot be used for
- Documentation of comparability across domains is often very deficient and it is a problem that Statistics Denmark's employees often do not know enough about adjacent statistical areas
- Users are often not aware of the contents of quality declarations. They should be supplemented with 'pop up' messages in Statbank.dk, especially where there is a problematic figure.

- The variables documentation system was rated as very relevant, especially “high-quality documentation” was praised. There is a wish to be able to distinguish between variables used in register data (microdata) from aggregated data (StatBank). One should have a filter in the search, so you could ask only to see the desired type of variables
- All documentation now on paper should be made digitally accessible. There is some important documentation that exists only in books and Statistical Reports and is therefore hard to find. For example, it might be good if all the questionnaries were available, so you could see changes in the questions over time.

12. Somemessages of statistical production:

- Statistics Denmark is generally too sectoral. It is hard to use Statistics Denmark across internal organisational boundaries. Adjacent statistics don’t relate to each other in publication, and experts are not familiar with other domains of statistics.
- It's important for many users to compare Statistics Denmark’s numbers with international figures, typically from Eurostat or the OECD, but it is not easy. Where are the corresponding figures? Especially Eurostat has much less documentation than Statistics Denmark. Users often try to find an indicator in e.g. OECD.Stat where the number for Denmark is similar to the Statistics Denmark number. It would be very helpful if Statistics Denmark could link to relevant sites.
- Several expressed the desire for more development in the statistics production in relation to tasks that users have (relevance). The statistics must keep abreast of developments in society
- Users want to be clarified whether statistics are preliminary or definitive numbers (challenge with both wanting to have statistics quickly and in excellent quality)
- users would like to participate actively in development groups around new statistical domains

13. Some messages regarding dissemination:
- Users suggest that there is a thematic approach to large domains, such as statistics on municipalities
- Users would like to have one entrance to all documentation : from the statistics there should be access / link to the right spot within the integrated documentation model
- Some users have difficulty downloading the statistics - it may be technically more simple?
- Be honest in the announcement if Statistics Denmark estimates that data quality is poor
- Definitions, comments to tables and figures should appear when the mouse is moved over the cells (in StatBank); likewise, warnings if there are breaks or concerns about data quality

III. Internal users web survey on the use of documentation in the Statistics Denmark

14. The internal survey aims to show how staff of Statistics Denmark use documentation to find or understand statistics, particularly statistics from areas other than their own. All employees were asked. Response rate was 51%. The answers contain many comments that show great interest among respondents for this topic.
15. Among those who responded, 88% knew about our documentation systems and 73% said that they used (one or more of) them; only these users were asked more questions about the systems.

Table 1: *Knowledge and use of documentation systems collectively distributed as follows in groups:*

	Total	Staff group			
		Academic	Manager	Clerical	IT
	per cent				
Know about documentation systems	88	94	96	79	86
Have used documentation systems	73	87	92	55	57

Table 2: *Knowledge and use of each documentation system, as a percentage of all respondents*

	Know of	Use
	per cent	
All documentation systems	88	73
Quality declarations	71	53
Concepts	47	15
Variables	56	24
”High quality documentation”	37	12
Classifications	58	35

16. Thirty answered yes to the question "Do you miss documentation of statistics?"

17. Some key messages that recur in many of staff comments to all questions are:
    1. There is no single, comprehensive documentation strategy, including how the different documentation systems complement each other
    2. Documentation should be more coherent and systematically maintained. This is mainly an organizational problem. A small departmental team should be established, clearly responsible of documentation systems and processes, supporting agencies using those systems and processes.
    3. Lack of guidelines for minimum documentation. We lack standardized metadata that follow data through the statistics production process. Metadata should be born where data are born.
    4. Many of the really important metadata do not exist in the four metadata systems. Documentation in the Statistical Bulletin or thematic publications is usually much better than quality declarations. There is a need for reference to where more in-depth documentation is available
    5. There is a wish for consistent documentation in terms of overall documentation pages, and links between different parts of the documentation
    6. Quality: The documentation is often incomplete (parts missing) and not up-to-date. There must be processes to ensure update. Furthermore, the quality of content varies a lot, something is really good.
    7. History: In connection with revisions, documentation of previous versions is often missing (but it is a historical problem, we make sure to do better going forward)
    8. There is also a need for process documentation and “cookbooks” for how to conduct a statistics.

IV. How can the findings be put to use?

18. In order to benefit users, it is important but not sufficient to find out what their unfulfilled needs are: We must also do something to move in the right direction.

IV a. External users

19. Our investigations confirmed the feeling that the fragmented pieces of metadata we already have are useful but far from ideal. Integration is important and must be addressed urgently. Next we have to look at completeness, the quality and better ways to present metadata, in connection with the data as well as a help to users to find the right data.
We should continue trying to understand important aspects of users' business processes in order to continually adapt our services to their needs. This is a never ending task.

IV b. Internal users and producers

20. In relation to our own staff there is also work to do. We wish to link this work to the GSBPM, which has been adopted as the process model for Statistics Denmark.
Our local version of GSBPM will be extended in order to explain how different groups of staff should make use of existing metadata in the different phases of the statistical production process. Guidelines for the production of metadata should be included in the GSBPM, explaining the processes in which staff must compile different types of metadata.

IV c. The metadata project

21. We have to admit that much is still unclear in this initial phase – even though we have been working purposefully with metadata for a few decades.

22. We have started by setting a *vision*: Where do we wish to go; how should users be able to get to statistics, find out if they are any good, and use them intelligently, assisted by good reference metadata?

23. We also set out a *strategy*, explaining how we intend to go in the direction of the vision. Knowing that we know too little to foresee what will be important in a couple of years' time, we break up the project in waves of around a year each, with less and less certainty about the exact content as time goes on.

24. The first wave of the project will largely concentrate on integration, keeping the separate systems more or less as they are but adding an integration layer. This wave will also include a thesaurus, defining preferred terms and relating all terms and concepts.

25. We are all the time trying to define a metadata model that would link all aspects of metadata. We would prefer this to be an international model that would allow us to work closely with NSIs from other countries and increasingly share solutions.