WP.19
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)
(Geneva, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

VARIABLES DOCUMENTATION SYSTEM IN STATISTICS NORWAY

Contributed Paper

Submitted by Statistics Norway[1]

I. INTRODUCTION

1. This paper will focus on the development of a variables documentation system in Statistics Norway (SSB). The overall purpose of the system is to document variables in a central location, accessible by all, and to function as a tool for harmonizing names and definitions. The pilot system was intended to fill the needs of the 2001 population and housing census for documentation of variables. We will discuss user participation, stepwise development, resources and results. A requirement on the system has been that it must not be built in isolation, but must be linked to other relevant metadata systems.

2. Parallel to the development of the variables documentation system, SSB participates in the Neuchâtel group[2] where terminology and models connected to variables are discussed.

II. PURPOSE OF THE VARIABLES DOCUMENTATION SYSTEM

3. In Statistics Norway information about variables can be found in different documents and systems, which makes accessibility and harmonization more difficult. The variables documentation system (Vardok) is intended to be a central system for documenting variables in Statistics Norway (e.g. definition, validity periods, classifications used) and a tool for harmonization of names and definitions of variables.

4. At present Vardok can be accessed by everybody in SSB, but in the future external users will also, to some extent, have access to this information. As the variables are to be updated in Vardok but can be used in other systems, it is necessary to establish links to these other systems. To ensure this, one of the requirements on Vardok was that the system should not be established as a satellite system, but should find its place in SSB's network of metadata systems.

5. As a result of today's decentralized storage of metadata, one might find the same variable name defined in different ways in different parts of the organization, and one might also find the same variable definition named in different ways. In Vardok this lack of harmonization will be visible to all, and the system will therefore be a valuable tool for standardization and harmonization. Each variable in Vardok will have an owner (one of the subject matter divisions) that will be responsible for entering the variable into the system and updating it.

---

[1] Prepared by Anne Gro Hustoft (agt@ssb.no) and Jenny Linnerud (jal@ssb.no)

[2] The Neuchâtel group working with terminology models for classification databases was established in 1999 and consisted of Statistics Denmark, Statistics Sweden, Statistics Switzerland, Statistics Norway and run Software-Werkstatt. Statistics Netherlands has joined the group for the work on variables.

III. CONTEXTUAL DESIGN

6. Based on their experience in the European Commissions Information Society Framework V project FASTER (Flexible Access of Statistics, Tables and Electronic Resources) the project group decided to use contextual design for the development of the variables documentation system. The method of contextual design (Hugh Byer and Karen Holtzblatt, 1998) gives the designers and users the tools they need to enter a partnership in which the users are the experts in their domain and the developers are the apprentices. The role of the developers is then to help the users articulate their needs and to distinguish clearly between the users intent and how this should be implemented. The users can explain what they need in a dialogue, not a document, and the developers decide how this can be implemented.

7. The first step we carried out was to identify different groups of users who will need to be represented. The next step was to interview these representatives in their own offices. Two people conducted each interview. One had a dialogue with the user and the other observed and took detailed notes. During the interview we explained the purpose of the system, tried to create a common vision of the system, captured background information for the interviewee, discussed links to other metadata systems and as many details on content and functionality for the system as we could. Each interview lasted 1-2 hours.

8. Experiences from interviews:

- While many people were apprehensive about being interviewed they relaxed as soon as they realized that the developers were there to reach a deeper understanding of the daily work rather than as examiners! The developers enjoyed gaining a deeper insight into the work of their colleagues and the opportunity to collect informal feedback about other systems they had developed.
- In interviews one should use focus to steer the conversation but also remember that focus reveals detail but conceals the unexpected. The interviewer should always be willing to expand their focus to discover surprises and contradictions.
- It is important to be aware of how users say no – Huh?!, Ummm ... could be. Any of these reactions could mean that the user disagrees but is too polite to say so. The interviewer should backtrack and gather more information.

9. The interviewers then went back to the project team with the results and tried to reach a common interpretation of these. Questions that arose in the interpretation sessions could often be answered in interviews with other representatives.

10. The next step was to structure all the information that was gathered during the interviews by identifying duplicate issues and gathering related issues. In contextual design this is called making an affinity diagram. For the variables documentation system we ended up with a four level hierarchy where the first two levels are shown below.

General aim
- general need for documentation
- metadata for steering processes

Content and maintenance
- definitions for variables
- sources for variables
- changes in variables
- sensitive variables
- maintenance

User friendliness
- functionality
- flexible reports
- user support

Links to other systems
- classifications
- file descriptions
- other metadata systems and documentation.

11. It may seem a bit obvious to have uncovered that the users want a user-friendly system. The point is that under this level you have three groupings of related requests and under each of these there are more levels. Under these again are the individual requests. The interviews provide a wealth of information that can guide the decision making of the developers.

12. After making this structure the project group should have a vision of the system. This vision can be shared with the user representatives in a brain storming session where the users are asked to examine the hierarchy and come up with more ideas and/or point out potential problem areas. We experienced that this session increased the feeling of ownership for the users, which was an important motivation factor later in the process.

13. The project group can use the hierarchy to identify different development steps. The vision may well be a five-year plan that should be broken down into shorter steps. We chose to implement our variables system in one-year steps mainly because this fits Statistics Norway's annual planning and budgeting process. Within one year we usually had repetitive cycles with planning, developing, user testing, improving, retesting, approval and release. Approval was by the project group based on the user feedback.

14. The structured information forms the basis of a user requirement specification that concentrates on the users intent and is written in the terminology of the users. Based on this the developers can make a prototype of the system. We chose to make a paper prototype. We then arranged paper prototype interviews with our users where they could test out the planned functionality of the system. There were many reasons for choosing a paper prototype.

- We were creating a new system not improving or replacing an existing one
- A paper prototype does not create unrealistic expectations for the time needed to take the prototype into production
- The focus of the users was on the functionality and not the layout. The users were not limited by existing functionality - they came up with many valuable and surprising suggestions.
- A paper prototype is much quicker and cheaper to change, so developers don't mind doing so.
- Disagreements between the developers about what they thought the users really wanted, were very quickly resolved by presenting the alternatives to the users, using the paper prototype
- The developers thought it was fun
- The users found the whole approach non-threatening and fun

15. After paper prototyping, a functional requirement specification for the system, with stepwise development, can be written and handed over, with the paper prototype, to those who do the programming. User testing can be based on the user requirement specification. The system (including documentation) should then be improved according to the users feedback and retested until the users are satisfied. These phases (prototyping, testing) should be repeated until the entire system has been built to the satisfaction of the users. Printouts of the screens can be used for subsystems that are already released so that users focus on new steps of the system development. The updated user and functional requirement specifications can form the basis for the system documentation

16. How many people should be involved in the process and how do you know that you have identified all relevant user groups? In practice, the interviewees gave us the names of relevant colleagues to be interviewed if they felt it necessary. The interviewing can stop when nothing new is uncovered.

IV. STEPWISE DEVELOPMENT - PART 1 (2001-2002)

A. Pilot system - population and housing census

17. To make sure that the system developed took care of real user needs, we wanted a well-defined and motivated customer for our pilot system. The Census 2001 was chosen to be this customer because they urgently needed a system for documenting their metadata. In addition the census was considered a suitable first step because it had

- clearly defined variables
- limited number of variables
- no variable history needed (all variables related to 3rd of November, 2001)
- limited number of subject matter divisions involved

18. In cooperation with the Division for Population and Housing Census the aim of the pilot system was formulated like this: The pilot version of Vardok should be a database where all variables delivered to the census are stored together with their code lists and available documentation.

19. The following user needs from the affinity diagram were given priority in the first step: Content + Maintenance (Variable definitions, Sensitive variables, Variable sources), User friendliness (Functionality, User support) and Links

(Links between variable system and Datadok). Datadok is SSB's file documentation database (mostly technical information).

20. The 5 subject matter divisions that delivered data to the Census, agreed to document their variables in Vardok.

B. Resources

21. The Vardok project group started with specialists within standards (1), subject matter (2-4) and IT(2). There is also a reference group consisting of heads of relevant divisions and a steering group consisting of heads of relevant departments. These groups give advice on crucial subject matter questions and priorities, and assign resources.

22. The table shows the number of subject matter divisions and people involved in the different steps of the system development in part 1. Our aim was that the people chosen for documenting the variables in Vardok, should participate in the different development steps. In reality, all these people had not been appointed at the start of our project, and some of them were exchanged along the way. The disadvantage of this situation was that we did not have contact with users who would actually enter information into the system until rather late in the process, but the advantage was that we captured a wider range of requirements than we otherwise would have. This was very useful in guiding our further development and making the system known to a larger part of the organization.

	No. of divisions	No. of people (new*)
Interviews	15	26 (26)
Walking the affinity diagram - brainstorming	8	9 (2)
Paper prototype 1	8	10 (4)
Testing 1	11	15 (5)
Production	5	9 (5)
Paper prototype 2	4	5 (0)
Testing 2	4	5 (0)

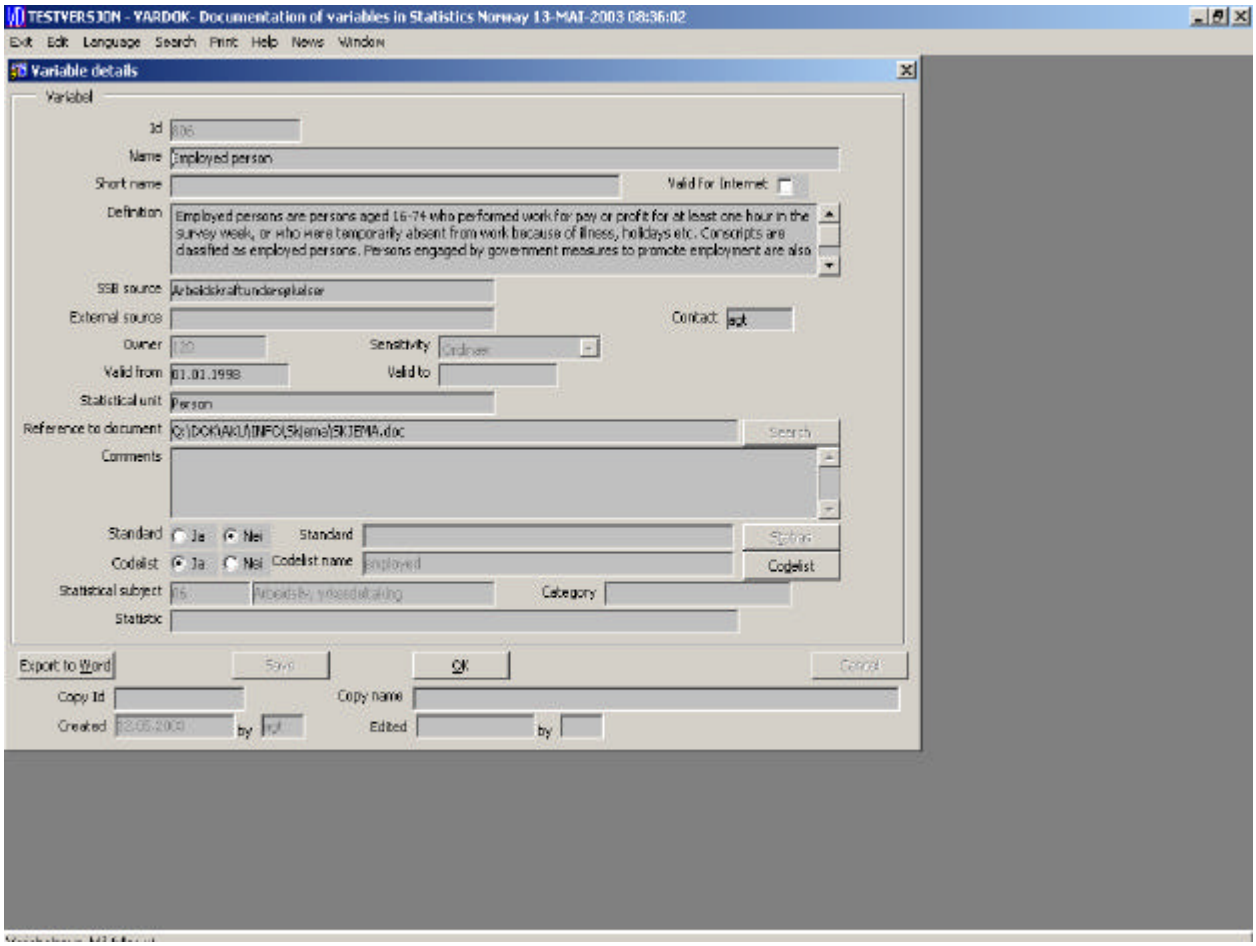
Total no. of people – 42 new* – not involved in any previous step

23. In total the resources spent on developing the pilot system in part 1 was a bit more than 2 man-years. The IT developers used about 70% of this. Note that this total does not include the amount of time spent actually entering the variables.

C. Results

24. 158 Census variables and 36 codelists were documented by the end of 2002.

25. The next figure shows the information documented for one variable in the pilot system. The fields are filled in by the subject matter specialists, and there are 5 compulsory fields (name, definition, contact, statistical unit and statistical subject). Owner is filled out automatically by the system - based on a division specific password, which the user needs to put variables into Vardok.



26. In addition to the fields mentioned, it is possible to document the source of the variable (where the variable first was defined) both inside and outside SSB (e.g. registers, surveys) and the validity period of the variable, give a reference to a relevant document (which can be immediately accessed by "double clicking" in the field) and link the variable to a codelist (e.g. in Datadok if it had already been documented there). One should also indicate if the data belonging to the metadata were sensitive (confidential) or ordinary.

V. STEPWISE DEVELOPMENT - PART 2 (2003)

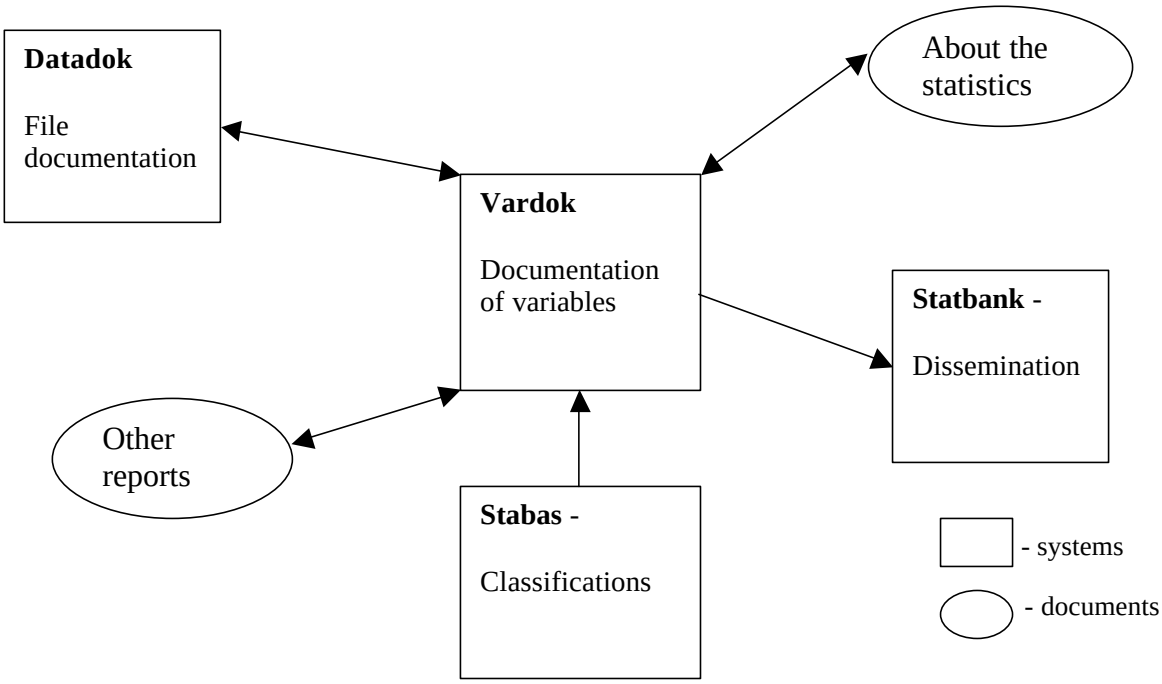
27. In 2003 we involved two more subject matter divisions as customers in order to gain experience with a wider spectrum of variables: one division responsible for social statistics and the other for industry statistics. These divisions would share their experience from statistics production, define new demands for functionality, test this functionality and document variables. One important effort in 2003 has been the linking of Vardok to other metadata systems.

D. Linking to other systems

28. As mentioned earlier, the pilot version of Vardok had a link to Datadok, the file documentation database, but only at the level of codelists. In 2003 the relation was extended to variables so that variables and their definitions in Vardok can be linked to variables in Datadok. In 2002 we also had the possibility to link variables in Vardok to documents both on our documentation server and on our intranet. In addition it was possible to copy and paste text between Vardok and About the statistics (a description related to all statistics disseminated on the web and one of the other places where one can find variable definitions in SSB). In the future our aim is to link About the statistics to Vardok and collect all its variable definitions directly from Vardok. In 2003 a link between Vardok and our classification server (Stabas) was established. Now you can link a variable to a classification in Stabas, and get direct access to the classification from Vardok.

29. Statbank is SSBs dissemination database. In the future we also plan to establish a link between Statbank and Vardok. This will enable us to show variable definitions from Vardok connected to the relevant variables in Statbank. So

far a first demo of such a link has been made. Statbank will be one of the ways in which external users will have access to information stored in Vardok.



E. Results and resources used

30. Besides some changes in the layout of the user interface, the following functionality was introduced:
- Extended link to Datadok and a link to Stabas
- The possibility to link documentation to be displayed on the web, and write comments that can be shown outside SSB (made to prepare for external use of Vardok)
- The possibility to link the variable to a specific statistics
- The possibility to mark if the variable is approved for dissemination internally or externally. This is done to ensure that the subject matter divisions can put their variables into the system, and use them as a basis for internal discussions, before they are released for use by the rest of SSB or to external users. As long as the variables are not approved for dissemination, they will only be visible to people within the division.

31. The resources spent by 31. October 2003 was 1 man-year. The IT developers again spent about 70 % of the resources. 507 variables are documented in Vardok (but not all of them are approved for internal dissemination yet).

F. Multilingual functionality

32. There are two versions of Norwegian with equal status; bokmål ("book language") and nynorsk ("new Norwegian"). SSB disseminates statistics in both languages. Vardok must therefore offer the possibility to document the variables in both languages. In addition, it must be possible to provide variable documentation in English. This multilingual functionality is our priority for the end of 2003.

VI. STEPWISE DEVELOPMENT - PART 3 (2004)

33. The project group in 2004 will have three members - two from IT development and the project leader from the Division for Statistical Methods and Standards. 6 divisions will document their variables in the system. In order to test out ownership and harmonization, 2 other divisions will follow the documentation of the variables of particular interest to their subject areas. One contact person for each division will be responsible for coordinating communication between the project group and users in their division. The leaders of those divisions that are entering variables for the first time, have been added to the reference group. The steering group is unchanged. Those responsible for event-history databases will also be involved in the linking to these systems. A journalist will read the definitions and provide feedback intended to improve the quality.

34. Planned resources for the project group in 2004 are 1500 hours from IT development and 500 hours from standards. Contact persons plan to use 75 hours and the person responsible for the event history databases plans 100

hours. In addition each division entering variables plans to use 1 week for proposing new functionality, testing and giving feedback.

35. There are many conceptual and practical challenges in the future development of Vardok. We will continue to work on these challenges in the project group, on a larger scale within SSB, and internationally within the Neuchâtel group. Some challenges in the future development are as follows.

36. Vardok's place in the production process - Vardok must find its place in the production process so that documentation is produced throughout the production cycle rather than being left until after the numbers have been published and there are no time or resources left before the next production cycle begins. Variables which, due to lack of time and resources, are documented after a production cycle, can be reused in the next cycle. Ideally documentation should be a by-product of the whole production process and not an extra burden on the last link in the process.

37. Units - Input variables should be connected to collection and reporting units. However, end-users would like to see output variables connected to statistical units. In our present version of the system we have not yet distinguished between different types of units - these must be deduced from the context.

38. Linked variables - The definitions of many variables refer to other variables e.g. net income = gross income - tax. Ideally the system should provide an easy way to see variables that are linked. Circular definitions should of course be avoided.

39. The role of time - Variable definitions change with time for many reasons. There is a constant struggle between the need to change a definition and the need for unbroken time series. It is important that those receiving the documentation are alerted to any changes. The changes may be at the conceptual level or at the production level. Validity time points and reference times should be available. The role of time in event-history databases is also important.

40. Types of variables - Numerical input variables can become quantifying or classifying output variables. Categorical input variables become classifying output variables. The handling of these types of variables throughout the production process will require care.

41. Users - Different users have different needs. Some users require access to very detailed information, while other users do not. Internal users need easy access to internal information. Requirements for comparability will be at different levels for different users. Investigating and meeting the needs of external users is a task for the years to come.

42. Definitions - The making of good definitions is not an easy task. The definitions are intended for different users with different needs. The definitions need first to be harmonized within a division/subject area and then harmonized across these. At all times we need to be aware of external definitions. International and national attempts to collect and update definitions in one central, accessible place are also taking place, e.g. Statistical Data and Metadata exchange (SDMX) vocabulary.

REFERENCES

Hugh Byer and Karen Holtzblatt, Contextual Design - Defining Customer-Centred Systems, Morgan Kaufmann Publishers, 1998.