DEVELOPMENT OF A VARIABLES DOCUMENTATION SYSTEM IN STATISTICS NORWAY

Anne Gro Hustoft and Jenny Linnerud, Statistics Norway

Summary

This paper focuses 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 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, e.g. the file description database and classification server.

1 PURPOSE OF THE VARIABLES DOCUMENTATION SYSTEM

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.

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.

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 keeping it updated.

2 CONTEXTUAL DESIGN

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, instead of a document, and the developers can decide how this is to be implemented.

The first step we carried out was to identify different groups of users who needed 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.

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.

2.1 Advice 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 we used focus to steer the conversation remembering that focus reveals detail but can conceal the unexpected. The interviewers were always willing to expand their focus and discovered 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.

2.2 Structuring the information

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 thereof
- definitions for variables
- sources for variables
- changes in variables
- sensitive variables

- maintenance of content

User friendliness
- functionality
- flexible reports
- user support

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

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

After making this structure the project group had a vision of the system. This vision was shared with the user representatives in a brain storming session where the users were 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.

The project group used 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.

2.3 Paper prototyping and testing

The structured information from the previous section formed the basis of a user requirement specification that concentrated on the users intent and was written in the terminology of the users. Based on this the developers was able to 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

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

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 stopped when nothing new was uncovered.

3 STEPWISE DEVELOPMENT - PART 1 (2001-2002)

3.1 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 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

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.
The following user needs from the affinity diagram were given priority in the first step: Content + 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).

In Norway a lot of the Census data are collected from registers within Statistics Norway. The 5 subject matter divisions that delivered data to the Census, agreed to document their variables in Vardok.

3.2 Resources

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

Table 1 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 replaced 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

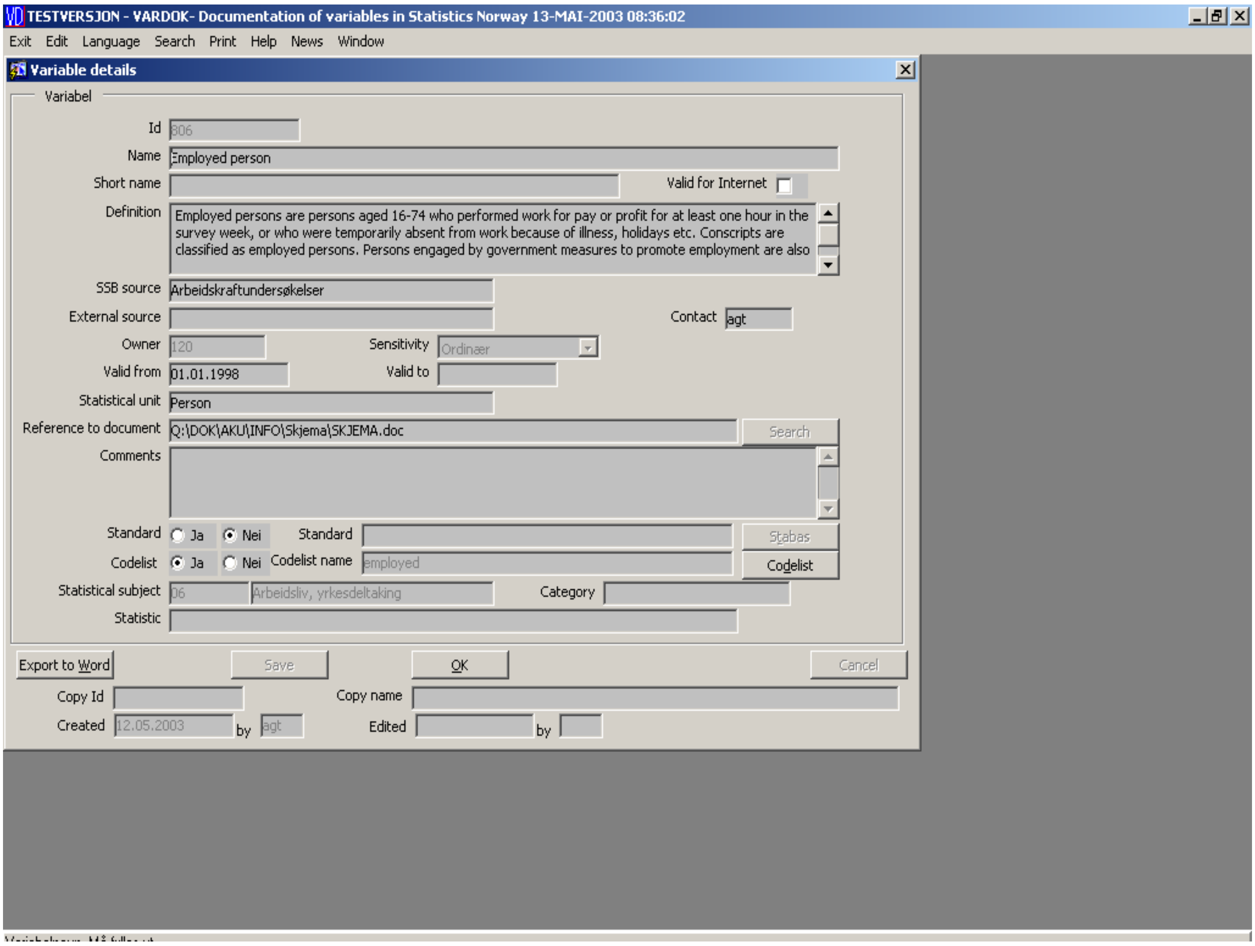
Figure 1 Number of people involved in system development - part 1

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.

3.3 Results

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

Figure 2 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 mandatory 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.



Id 806
Name Employed person
Short name
Valid for Internet ☐
Definition Employed persons are persons aged 16-74 who performed work for pay or profit for at least one hour in the survey week, or who were temporarily absent from work because of illness, holidays etc. Conscripts are classified as employed persons. Persons engaged by government measures to promote employment are also
SSB source Arbeidskraftundersøkelser
External source
Contact agt
Owner 120
Sensitivity Ordinær
Valid from 01.01.1998
Valid to
Statistical unit Person
Reference to document Q:\DOK\AKU\INFO\Skjema\SKJEMA.doc
Comments
Standard ○ Ja ● Nei
Standard
Codelist ● Ja ○ Nei
Codelist name employed
Statistical subject 06 Arbeidsliv, yrkesdeltaking
Category
Statistic
Copy Id
Copy name
Created 12.05.2003 by agt
Edited by

Figur 2 Variable information - 2002

In addition to the fields mentioned, it was possible to document the source of the variable (where the variable first was defined) both inside (e.g. surveys) and outside (e.g. registers owned by different authorities) SSB 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 not.

4 STEPWISE DEVELOPMENT - PART 2 (2003)

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 was the linking of Vardok to other metadata systems.

4.1 Linking to other systems

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 this 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.

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.

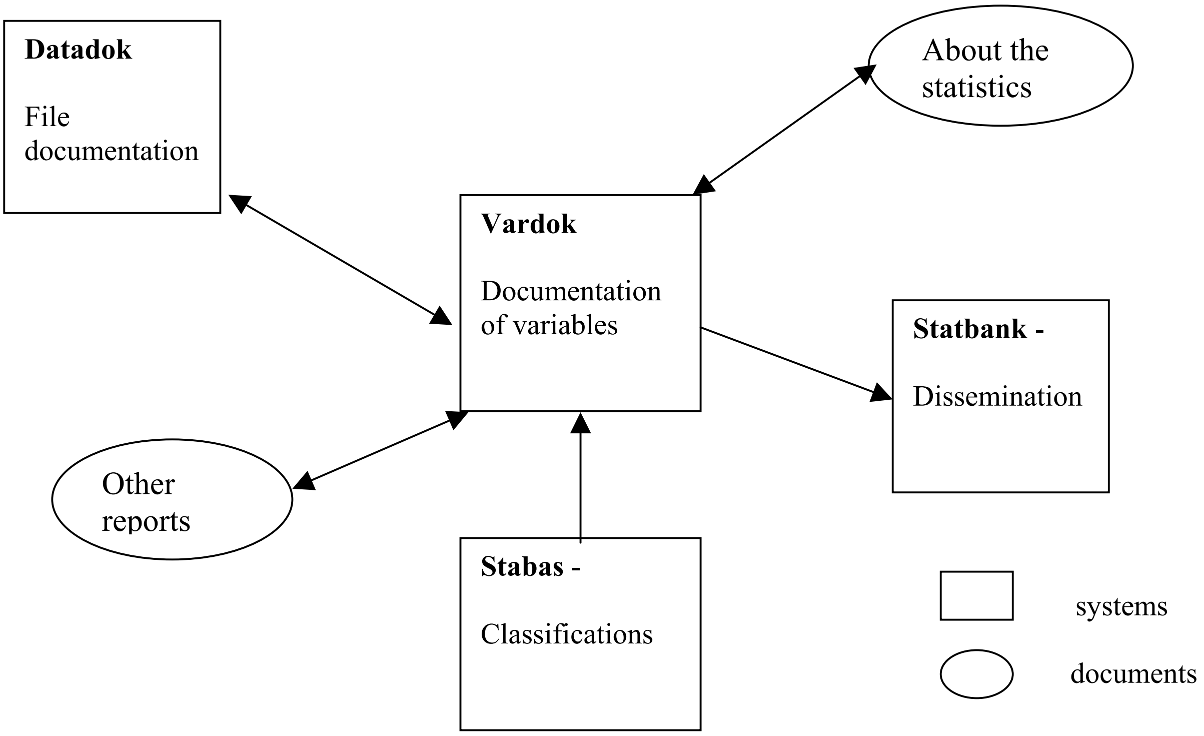


Figure 3 Metadata systems and documents linked to Vardok

4.2 Results and resources used

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 statistic
- 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 external users. As long as the variables are not approved for dissemination, they will only be visible to people within the division.

The resources spent in 2003 were 1. 2 man-years. The IT developers again spent about 70 % of the resources. 509 variables were documented in Vardok at the end of 2003 (but not all of them are approved for internal dissemination yet).

5 STEPWISE DEVELOPMENT - PART 3 (2004)

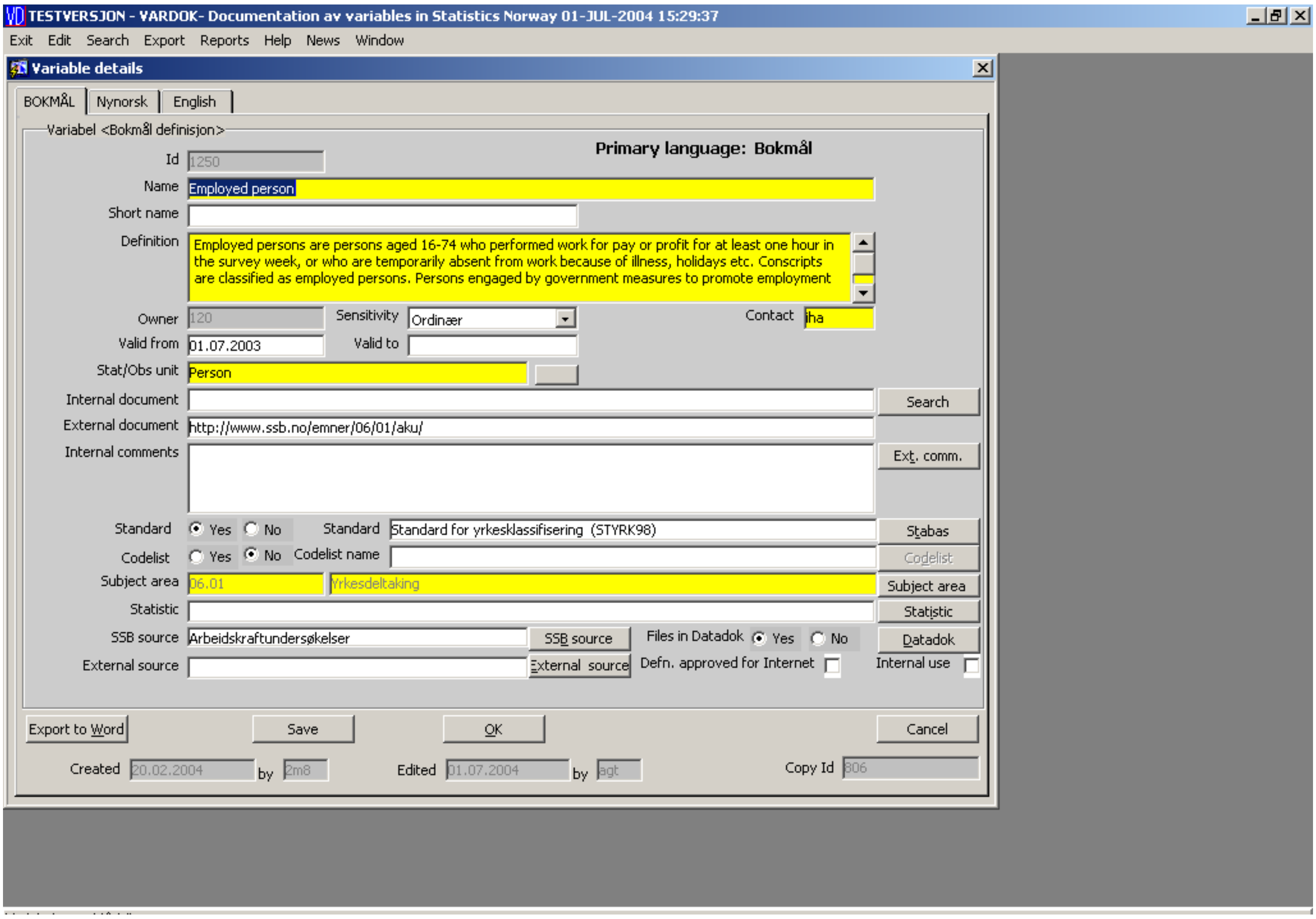
5.1 Plans and results

The 2004-project group has three members - two from IT development and the project leader from the Division for Statistical Methods and Standards. 6 divisions are documenting their variables in the system. In order to test out ownership and harmonization, 3 other divisions are following the documentation of the variables of particular interest to their subject areas. One contact person for each division is responsible for coordinating communication between the project group and users in their division. The heads 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 (longitudinal) databases will also be involved in planned linking to these systems.

Planned resources for the project group in 2004 are 1500 hours from IT development and 500 hours from standards. The project group has used a little less than 1000 hours by June 2004. The IT-developers have spent about 75% of the resources. 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. About 750 variables are contained in the system by June 2004.

The main effort so far in 2004 has been the implementation of multilingual functionality. 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 is necessary because the dissemination database (Statbank) publishes all statistics both in Norwegian and English.

The implementation of multilingual functionality is now completed, and the information window for a variable is shown in figure 4. This picture is more complex than the one in figure 2, because the new functionality of stepwise development parts 2 and 3 has been implemented; e.g. the user can choose if he/she wants to link the variable to a codelist in Datadok or to a classification in Stabas (all classifications are not contained in Stabas yet). The users choose the language in which they want to document their variable, by clicking at the appropriate fan. When the variable is translated into another version of Norwegian or into English, the user doesn't need to fill in all the fields once more. The only mandatory fields in the translated version are name and definition. External document and external comment may be filled in if relevant. Other fields will be translated into the relevant language by the system.



Primary language: Bokmål

Id: 1250

Name: Employed person

Short name:

Definition: Employed persons are persons aged 16-74 who performed work for pay or profit for at least one hour in the survey week, or who are temporarily absent from work because of illness, holidays etc. Conscripts are classified as employed persons. Persons engaged by government measures to promote employment

Owner: 120

Sensitivity: Ordinær

Contact: iha

Valid from: 01.07.2003

Valid to:

Stat/Obs unit: Person

Internal document:

External document: http://www.ssb.no/emner/06/01/aku/

Internal comments:

Standard: ● Yes ○ No

Standard: Standard for yrkesklassifisering (STYRK98)

Codelist: ○ Yes ● No

Codelist name:

Subject area: 06.01 Yrkesdeltaking

Statistic:

SSB source: Arbeidskraftundersøkelser

Files in Datadok: ● Yes ○ No

External source:

Defn. approved for Internet ☐

Internal use ☐

Created: 20.02.2004 by 2m8

Edited: 01.07.2004 by agt

Copy Id: 806

Figure 4 Variable documentation - present version

5.2 Future challenges

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[1] group. Some challenges in the future development are as follows.
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

---

[1] *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.*

should be a by-product of the whole production process and not an extra burden on the last link in the process.

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.

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.

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.

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.

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, and the project group is following this work.

Acknowledgement

The authors want to thank Jan Rønningen who has been responsible for the IT-development in Vardok.

6 REFERENCES

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