Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki

Documentation of the Social Statistical Database

Mogens Grosen Nielsen
Statistics Denmark
mgn@dst.dk

Chapter 1. Introduction

This paper describes the experiences regarding the introduction of a new approach to documentation that supports the implementation of the strategies on organisation and storing of data in the Social Statistical Database.

Chapter 1 (Background) - tells shortly about the Register Strategy, the use of new technologies and other issues that formed the background for the new documentation approach.

Chapter 2 (Objectives) - tells about visions and short term objectives that guided the development of the documentation system.

Chapter 3 (The solution) tells about how we addressed the objectives and how we handled (and are handling) the implementation of the documentation system in the Social Statistical Department.

Chapter 4 (Lessons learned and the way forward) tells about pro and cons of the selected approach and in addition shortly about considerations about the way forward.

Chapter 2. Background

The revised documentation system is part of the general work on reorganising data prepared for social statistic. In short the strategy for this reorganisation implied modelling and implementation of a connected set of statistical registers. And in contrast to the old organisation of data the new organisation of data should be based on common definitions of concepts, reduction of data redundancy and a common documentation to support the use of data across statistical domains.

In parallel to this development it was decided to migrate from a mainframe platform to a PC/LAN based platform. This decision gave a technological push towards a better organisation of data via the use of database technologies.

The starting points for the documentations system were the two above mentioned aspects. The system should support the revised strategy on registers and the use of database technologies. And because of lack of resources it was emphasised that there should be as much automatisation as possible. In addition we looked for international experience and included the following recommendation as an important starting point.

> *"It should be a challenge for every statistical service to organize its metadata flows in such a way that*
> *1) as many metadata as possible can be obtained from existing metadata holdings, whenever they are needed by a certain statistical system*

> *2) as few metadata as possible have to be produced for its own sake, rather than as a side-effect of other (necessary) activities of the statistical systems monitored by the statistical office"*[1]

On basis of the above mentioned aspect we emphasised that as much input as possible should be extracted from external sources. These sources were to begin with database systems like Oracle and SAS. In the future we expect to include other kinds of applications. In other words: we want the documentation system to work as an "umbrella metadatasystem".

Chapter 3. The objectives for the documentation system

In continuation of aspects mentioned above we set up the following generel objectives:

- It must be easy for subject-matter staff, IT-staff og operation staff to learn about and get acquainted with any statistical domain. For subject-matter staff it must be easy to get acquinted with objectives, concepts and the relevant processing of data from input to final data at the Internet and in publications. IT-staff must have sufficient knowledge in order to develop and maintain the system. The operational staff must have adequate knowledge on how to operate the applications.
- Data used across statistical domains must only be described once and it must be easy to give access to and use information about data and processes accross statistical domains. This objective becomes especially relevant, since we want to reduce data redundancy.
- External users should have easy access to generel descriptions via declarations of data and via targeted decriptions of data.
- Besides the traditional metadata descriptoins it must be possible for staff at a statistical domain to store / remember day-to-day knowledge. E.g. "the data editing of inputvariable on income showed extreme variations in 2006." In order to differentiate it must be possible to store information on three levels: 1. Staff employed at a statistical domain, 2. All staff in Stastics Denmark and 3. All external users.

As mentioned in the background an important additional condition was to ensure that the building of a the new documentation system should be implemented in a situation with few available ressources. Therefore we had the following additional objectives:

- The information in data-dictionaries in products like SAS and Oracle should be used directly. The system should work as "an umbrella metadatasystem".
- Data on variables and value-sets should be automatically converted from the old documentation system to the new documentation system.
- The documentation system should be used actively during the development proces. Documentation should not be an additional task to carry out after the application is in production. Documenations must be a by-product of the development proces.

Chapter 4. The solution

The application was developed at the IT-department of Statistics Denmark. The set-up of the requirements and testing the solution were done in a user-group where the four departments in Statistics Denmark were represented. This group discussed various problems and solutions in parallel with the development process in the IT-department.

[1] *"Guidelines for the modelling of statistical data and metadata"* United Nations, Geneva, 1995. and *"Documentation and Quality in Official Statistics"* Bo Sundgren, Statistics Sweden 2001.

The main characteristics of the application are as follows:

- In order to ensure the access for all stakeholders the first page for a statical domain was designed to give an overview and to give direct acces to relevant information for all. This implies that you from the first page have direct acces to user-manual, various decriptions of variables, processes etc, access to user-manual and access to maintenance manuals for IT-staff.
- Variables in the Social Statistical Database should only be described once. Therefore we introduced the concept task-group (group of domains). Inside each task-group a variable can only occur once. This implies that the statistical area that owns the variable are responsible for the documentation of the variable. And if that particular variable occurs on other domains the documentation system automatically sets a pointer to the owner's variable.
- In order to ensure targeted description to relevant stakeholders it is possible to direct internal information either to staff working on a statistical domain or to all staff at Statistics Denmark. Externally it is possible to differentiate between descriptions to the general user and detailed descriptions to e.g. researchers.
- Documentation should be easy to insert from other metadatarepositories. Therefore we built the application so that all basic information about dataset and variables are imported and updated directly from meta-data repositories in Oracle and SAS. Inserting a description of a dataset is just a matter of point and click.
- The objective regarding integration with the development proces was among other things handled via storing of diagrams and descriptions of conceptual models used at the statistical domains[2]. The development proces was further supported via a 'draft-mark' on all descriptions. If the 'draft-mark' is set, then it is only the staff in the development-group that has access to the descriptions. The draft-mark is typically set during the development phase and when the descriptions are finalised the draft-mark is removed so that all staff members of Statistics Denmark have access to the descriptions.

The main aspects of the implementation of the documentation system in the Social Statistics department are as follows:

- In order to ensure discussion and legitimate decisions on the use and objectives of the new documentation system we developed a set of polices on what to include in the documentation of a Statistical domain. The policies included among other things descriptions on which elements in the documentation that should be public and guidelines on how to describe processes and variables.
- Seen in relation to the old documentation system the requirement on how and what to document were changed in many ways. The roll-out of the solution in the Social Statistics department therefore took place in three steps. First we focused on description of variable to be used across subject-matter areas. This enabled us to reach the objectives for the strategy on Registers. Secondly we focused on the description of all variables. And finally we focused on finalising both the more IT-related matters and also general descriptions of processes.

[2] The concepts used are similar to standards in ISO/IEC 11179. See introduction to concepts in the paper "Metadata standards and their support of data management needs" Statistics Canada, Bureau of Labor Statistics, United States. Invited paper at the Joint UNECE/Eurostat/OECD work session on Statistical metadata (METIS).

Chapter 5. Lessons learned and the future work

The introduction of a new documentation system gave rise to discussions as expected. They were both positive and negative. On the positive side the users expressed, that the link to database systems implied that the times used on documentation was reduced dramatically.

The application was build, so that subject-matter statisticians, IT-staff and operational staff all needed to contribute and access data. This involved more cooperation between subject-matter statisticians and IT-staff. After some discussions the result is now, that sharing of knowledge between various groups has increased.

It is not all the objectives that have yet been transformed into real life. The integration of the documentation system into the development process still needs attention. And due to prioritising there are still areas that need to finalise the documentation task. In order to ensure the way forward the process should be followed closely at the management level.

The future work consists in further development of the documentation system including additional requirement from other departments. We also need to coordinate description of variables with other departments. If a variable has its source in another department then the description should be collected or referred to in the Social Statistics department.

On particular challenge in Social Statistics department is how we at the same time a) ensure reduction of data redundancy and b) give easy access to data across statistical domains. We are for the moment discussing and testing various solutions. When we reach a final solution to this problem we will adjust the way we create and use the documentation.