4. Statistical Metadata Systems (German Federal Statistical Office)

3. Statistical Metadata in each phase of the Statistical Business Process (German Federal Statistical Office) | German Federal Statistical Office | 5. System and design issues (German Federal Statistical Office)

4.1 Metadata system(s)

4.1.1. GENESIS (in use)

GENESIS is a cube database used in the Verbund by many statistical offices. It is based on an extensive data and metadata model and handles its metadata internally. First drafts of the system date back to 1994. At that time, GENESIS was intended as a data warehousing solution mainly to store macro data for internal purposes. Although it is also used in this way, its main purpose has become to serve as a dissemination database to internet users (since 2002).

In many ways GENESIS overturned existing habits of disseminating data at Destatis and in the Verbund when it was introduced. The cube model along with the standardised metadata entry forced a new way of thinking onto subject matter statisticians. Constrained by organisational issues - especially coordination in the Verbund - and legacy IT-systems, it often stretched the resources of the subject matter departments and the central coordination unit. Despite the age of the design, it is only now that its full potential is being realised. Especially in combination with the centralised micro data storage build by SteP, it is possible to populate the cubes faster and hence build larger cubes and publish faster. GENESIS is integrated into the web pages of the offices in the Verbund. At Destatis it is linked to the press releases so that interested users can search for additional data.

GENESIS was implemented using a programming language called Natural and a pre-relational database technology named ADABAS. ADABAS was first introduced in 1971 and - with many updates - is still heavily used in legacy software at public institutions in Germany.

GENESIS has several clones, with each office having its own database. There is also a GENESIS clone with nationwide data at a regional level. The GENESIS model itself has over the years proven its worth as a data and metadata model for a dissemination database. One Land office (Rheinland-Pfalz) deployed a new dissemination database a few years ago which is based on the same GENESIS model while using relational technology.

4.1.2. RDC-Metadata system (being populated)

With the establishment of research data centres (RDCs), statistical offices in the Verbund began to realise the need for a database holding metadata that could explain the content of the research data files to the researchers. The decision was made to expand the metadata part of the existing GENESIS-system. The result was a metadata system that contains information especially on the level of individual data files, on the level of statistical activities and on the level of variables.

Each variable ought to be entered only once and is then tied to the data file and thereby to the statistical activity it is used in. To avoid the duplicated entry of variables with different names but similar content, an editorial team reviews each variable individually. This basically follows the same idea that was employed in the GENESIS database.

It is interesting to compare the (meta-) data model of the RDC-Metadata system (essentially an expanded GENESIS model) with the other models like the Neuchâtel model or ISO 11179. In some ways they are similar, but the idea of a conceptual variable or of an ISO 11179 data element scheme does not exist. Therefore, variable definitions have to be harmonised at a very low level. The variable is modelled as an object with a definition and a value domain. Categorical variables have their categories (called value domain items in Neuchâtel speak) as objects of their own. The value domain is not modelled as an object on its own. Therefore, variations in the value domain of a variable necessitate the entry of a new variable. As a result, the number of variables rises and the system today stores 5,600 variables for the micro data files of 33 statistical activities.

Nevertheless, the RDC-Metadata system has been successfully implemented and is popular with researchers using the data centres. Since it is not possible to access the research data files via the internet, any prior information about their content is welcome. The system is not yet fully populated as metadata exists only for 35 of the planned 60 statistical activities.

4.1.3. Output oriented metadata system (intended)

The success of the RDC-Metadata system quickly led to a decision to use the same system not only for the RDC-relevant statistical activities but to apply it across the board. The result is the idea of an "output oriented metadata system". Although a business case does not exist, it could be used to document the metadata of the finalized micro data files. A cost analysis of this project still has to be undertaken, but from what can be said today, it seems unlikely that the original idea of harmonising variables at such a very detailed level can be realized by way of an editorial team. With already 5,600 variables for the metadata of 33 statistical activities in the RDC-system, the figure is likely to rise significantly when variables for up to 390 statistical activities have to be stored. In any case, the number of variables stored in the system will most likely be too high to harmonise the variables by comparing them one to one.

4.1.4. Statistikdatenbank (in use, being reengineered)

The Statistikdatenbank stores metadata for all statistical activities at a very high level. It exists currently in the form of two MS-Access databases. One is used to maintain the central catalogue of all statistical activities (called EVAS - Einheitliches Verzeichnis aller Statistiken) of the Verbund.

The second one is used for management purposes, containing basic information on methodology, legal background, etc. The reengineering will combine this information in a single application that will allow accessing and querying the information via the internal web portal of the Verbund. As a result, general information on all statistical activities will be visible to all users in the Verbund.

In the course of its further development, the Statistikdatenbank will become a central hub for the management of statistical activities at Destatis and in the Verbund. Every new statistical activity will first have to be registered in the Statistikdatenbank and is then identifiable by its unique EVAS-code (registration meant as a business process, not necessarily in a strict IT-sense). The Statistikdatenbank can easily be amended and combined with other metadata storages at Destatis that use the same EVAS-catalogue. For example, it is planned to integrate the quality reports directly into this application. Quality reports contain partially overlapping information but are currently stored as single text files written according to a given template. It is conceivable that other EVAS-based systems - like the database used to compile Destatis' Strategy and Programme Plan or internal accounting databases - will also be loosely attached or linked to the Statistikdatenbank.

4.1.5. KlassService (in use, being reengineered)

KlassService is a tool developed by the Bavarian State Office for Statistics and Data Processing. It is used to classify and code answers entered in free text fields in questionnaires. It currently houses only two classifications (the German NACE and PRODCOM versions). Since the administration of standard classifications is under the responsibility of Destatis rather than the Länder, the classifications and the additional thesaurus are maintained by Destatis using a web interface. KlassService has also been declared a standard IT-tool under the SteP guidelines. As such, it is used to support the classifying and coding of responses in many offices of the Verbund.

KlassService was built using an ADABAS database and is now considered a legacy system. Because of rising maintenance costs, the Bavarian State Office expressed the wish to move to relational technology. At the same time, Destatis' classification department was making plans to build a comprehensive classification server. The classification department had previously advised the Turkish National Statistical Institute on the design of such a system.

As a result of these initiatives, a business case was drafted that involved a redesign of the old KlassService in three successive stages. The first stage basically consists of the database itself and basic import functionalities. The succeeding stages will focus amongst other things on the user interfaces. The first stage is being carried out by the Bavarian State Office for Statistics and Data Processing. The later stages will be put out to tender in the Verbund.

The new KlassService will bear little resemblance with the old system. It will be based on the Neuchâtel Terminology, Part I, which will only be slightly altered to fit the relational technology employed. Web service functionalities enable connections to other databases and IT-tools (namely to other metadata systems). The system will also be designed to support multiple language versions of the classifications.

4.1.6. Census metadata system (being drafted)

According to a decision made by the heads of the offices of the Verbund, a separate metadata system has to be developed for the Census 2011. The system will be of modular design and so several drafts for individual business cases have to be written. Some of the applications could possibly remain in use after the census has been completed and - if applicable - be employed in other statistical activities as well. A decision on the implementation will be made in collaboration with the IT-working group for the Census.
Issues of census metadata management include:

4.1.6.a Database for methodology documents and other documentation

It is standard practice among statisticians to deliver most of the documentation in written form. A sophisticated methodology, the need for coordination between many parties involved and a very intense preparation phase lead to an enormous amount of text files being written for the Census 2011. However, there is currently no tool to store such documentation in a structured way in the Verbund. In order not to change existing work practices, the first measure to be taken for the Census 2011 will be a relatively simple document management system. It will be structured according to the Census process model (fig. 3); requiring statisticians to deliver documentation in the form of a text file for each applicable phase on level 2 of the model (see also 3.2.). The documents provided will largely be documentation already existing but structured according to the process model.

4.1.6.b Database for variables, value domains and statistical units

To move the documentation of variables for the Census 2011 from written text files to a more accessible and regularly updated form, a database for variables will have to be realised. The draft for this application is currently being written. It will be based on the Neuchâtel Terminology Model (Part II, Variables and related objects).

4.1.6.c Database for matrix and processes

To fully document the statistical data collected, processed, analysed and disseminated in the 2011 census, the different data holdings ought to be documented. According to the current plan, this will either be an extension of the variable database or an independent system. A separate draft will be written for this part, but it will also be based on the Neuchâtel Terminology Model (Part II).

4.1.6.d Connection to production tools

To realize the potential of metadata and reduce duplicated entries, the metadata system ought to be connected to production tools. A draft will be written to explain the connections between the systems and how a coupling of the different tools can be realized.

4.1.6.e Classifications

Several standard classifications will be used in the census. Since these classifications should be stored in the KlassService database, avoiding duplicated entry, these systems must be linked in some way. A draft will be prepared to explain the connections between the systems.

4.1.7. .BASE - Common IT-Applications for Statistical Surveys (in use)

.BASE (Basis Anwendungen fuer Statistische Erhebungen) is the umbrella name for several IT-tools - developed for the Verbund - to support a standardised e-workflow and forms an important part of the SteP-project. Some of the .BASE tools - notably a data editing tool - are metadata driven and load their metadata from a central storage named "survey database".

The survey database registers every survey in the Verbund. A statistical activity may consist of one or more individual surveys. For every survey, several resources can be uploaded and accessed in the survey data base. Apart from text files and other documentation, several XML-files containing metadata to drive production processes can be stored. These XML-files contain for example registered variables and executable code to drive data editing processes in different IT-environments.

The metadata in the survey database is clearly on a technical level. In the terminology of the Neuchâtel model the variables are on a level lower than the conceptual level. It is obvious that the survey database would therefore provide an almost perfect vehicle to transport conceptual metadata (being stored in classification servers or variable databases) into the production process. This would link conceptual and production metadata and - together with a powerful data warehouse at the end of the statistical value chain (see GENESIS above) - would almost finalise a metadata driven production process (provided other IT-tools would use the survey database as well).

To that end, however, several steps will have to be taken beforehand. The survey database was not designed with international metadata standards in mind. For obvious reasons, the focus of the designers was to connect production tools to the database. Given the myriad ways to design a statistical activity, it is unsurprising that the definition of the term "survey" remains somewhat ambiguous. This is currently not a problem for the existing .BASE tools but it will become more of a problem when the survey database is connected to more production tools and other metadata storages (like classification servers or variable databases). Therefore, an overarching metadata model and a standardised terminology is needed to integrate additional production and metadata systems, to facilitate the interoperability of the SteP-tools and thus to ensure the overall success of the SteP project.

4.2 Costs and Benefits

Metadata management contributes directly to the realisation of major objectives in Destatis' corporate strategy. It enables the further standardization of processes, the harmonisation of statistics and the monitoring of data quality (see corporate strategy). Metadata systems help with the documentation of surveys. To ensure that the public trusts in the data which Destatis and the Verbund produce and to be able to claim that the data has been compiled according to an appropriate methodology, a good documentation is indispensable. With central metadata systems in place, duplicated entry of metadata becomes unnecessary, it will be possible to share information easily, to drive production systems and to keep internal and external users informed about the statistical activities. A metadata model that allows for the correct representation of the metadata of all statistical activities can itself be a powerful tool in the standardisation of business processes and IT-systems because it represents a common structure for all statistical activities.

4.3 Implementation strategy

Since several IT-systems that run on metadata are already in place and given the complexity of the issue, we have decided in favour of a stepwise implementation strategy. The Statistikdatenbank and the new KlassService are the systems that will become operational first while a variable database and a tool for managing textual documentation (both part of the Census 2011-project) are next in line. A detailed project management is in place in the Census project. The development of KlassService is managed by the Bavarian State Office for Statistics and Data Processing. The progress of the Statistikdatenbank is dependent on the resources of Destatis' IT-department.

The major design work on the Census 2011-related systems will have to be finished in the first half of 2010. As the census will be conducted in May 2011, maintenance and helping users with the systems could have become a major task by then. After 2011 the attention might turn to generalising the lessons learned and broaden the metadata management with involvement in SteP gaining in importance.