5. System and design issues (German Federal Statistical Office)

4. Statistical Metadata Systems (German Federal Statistical Office) | German Federal Statistical Office | 6. Organizational and workplace culture issues (German Federal Statistical Office)

5.1 IT Architecture

Given that the Verbund - explicitly and implicitly - follows a step by step approach, no single, all-encompassing metadata system exists or will exist in the future. Instead the metadata architecture will consist of different independent systems. Each system will be developed on its own and therefore have its own IT-architecture both in terms of data model and business architecture.

To allow internal and (at a later stage) external users to access the metadata stored in the existing metadata repositories (including various applications that in one way or another store metadata) a web portal will be set up (see 1.2 "metadata portal"). In order to connect to the portal, each of the participating applications will need to have web service functions.

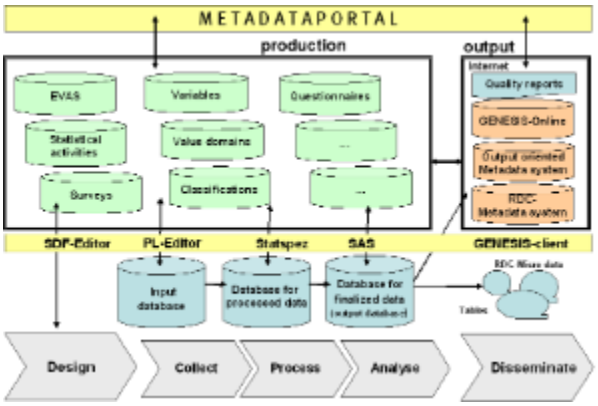

Figure 5

However, when an assessment of prospective underlying systems began, it soon became clear that establishing a technical connection between the systems and the portal was not the decisive issue. Instead it turned out that the systems had different ideas on how to structure metadata. Not only were the formats and data models unique, each system also had its own terminology. Sometimes, the same term could mean different things in different systems. While this was acceptable or even desired with respect to the different tasks of the systems, it surely complicates interaction. Hence, it soon became clear that no meaningful presentation of the content could be given without a shared metadata model and a common terminology to name and identify the metadata.

In order to arrive at such an overarching model several international metadata models were reviewed. However, with the Census 2011 having become the most important issue, the insights gained will first find application in the metadata system for the census. Within step 12 - metadata (see 1.), a possible overarching model as a solution for the interaction of the different systems still has to be discussed. Meanwhile, a first step towards a metadata portal will be to make Statistikdatenbank accessible via web services to internal users of the Verbund.

With respect to the census, the work began with choosing a metadata model that would fit the requirements of the census. The census is based on existing register information that will be combined with multiple surveys for data not covered by the registers. As a result, there are different data sources to be described, which often contain slightly different variables. Therefore, the task of a comprehensive metadata management is to track slight changes - for example in the meaning of terms or in the coding of enumerated variables - while reusing existing information to a large extent.
After reviewing the Neuchâtel-model, the responsible working group decided that the Neuchâtel model was the model that best fitted the requirements. It seems that the object types in the model represent nearly all conceivable meta-information needed to describe data in different settings. The Neuchâtel model therefore offers the possibility to integrate our existing systems (mainly .BASE and GENESIS) as well as systems that are currently being realised (KlassService and Statistikdatenbank). A high level overview exists that explains the conceptual connections between the systems (see below). It is based on the METASTAT@FSO design developed by the Swiss Federal Statistical Office which implements the Neuchâtel terminology to a large extent.
KlassService and Statistikdatenbank are not part of the census funded systems but will play a role in the management of the census metadata and are therefore part of the model. Not represented is a tool for document management that is being developed as part of the census. The centrepiece of the model is a variable server that is currently being drafted (see also 4.1).

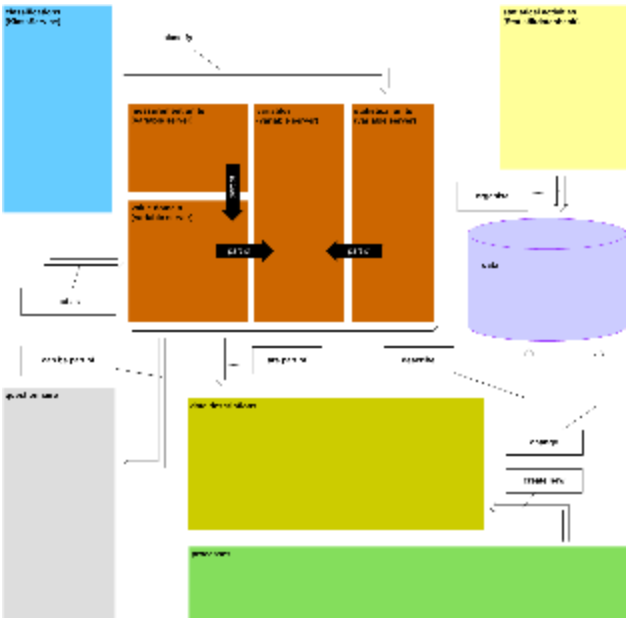

Figure 6

5.2 Metadata Management Tools

Links between data and metadata exist in the .BASE-system. As part of this system, there is a tool for defining data editing rules on the basis of pre-defined metadata. These variables and their value domains are stored in a separate repository (survey database). Reuse of metadata is encouraged by allowing users to share their variables. Before a user can delete or change a variable, all other users of this variable are asked to agree.

The GENESIS cube database also features a metadata repository. In contrast to the .BASE-system, GENESIS stores its metadata internally. Before creating a cube, the cube variables first have to be defined. Reuse of existing metadata is facilitated by an editorial team that checks each variable individually. There are currently about 1.300 active cube variables in the system at Destatis (for a total of 180 statistical activities) - a figure that has proven to be manageable.

5.3 Standards and formats

So far, mainly internal and national standards are applied. The .BASE-system runs on several nationally designed XML-formats. DatML/SDF (Survey Definition Format) describes the survey (esp. the variables). DatML/EDT holds the metadata that defines the data editing rules. DatML/ASK is metadata to set up electronic questionnaires.

GENESIS has its own database model, which can be seen as a national standard since most Länder offices either use the GENESIS database or a database based on the model. GENESIS is used to send data and metadata to Eurostat with the SDMX-standard.
The reengineered KlassService will be based on the Neuchâtel model, Part I.

There is no standard format used in the Statistikdatenbank. Nevertheless, the understanding of the term statistical activity ("Statistik" in German) is much the same as in the Neuchâtel model. Therefore, Statistikdatenbank can interact with other systems within a distributed metadata systems that allocates different functions to different systems according to the Neuchâtel model.

5.4 Version control and revisions

In theory, there seem to be two different ways to versionise metadata. One is to attach validity periods (valid from, valid until) to metadata objects. This is done nearly in all databases. The other seems to be to create additional object types for the versions of a metadata object type. In this way, there exists an object type for general information on an object plus another object type that captures a list of versions which record the changes to an instance of the first object type over time.

Validity periods are used in the .BASE-system, for example to identify active surveys. Inactive surveys are disposed of, if not archived (archiving functionality planned).
Instantiations are used in the KlassService where - following the Neuchâtel model - the classification versions are an object type of their own. General information (that does not change over time) about classifications is captured by the object type classification.

Instantiations have also been introduced in the RDC-metadata system where each statistical activity has a list of statistical activity instances capturing the individual features of each successive survey.

5.5 Outsourcing versus in-house development

There is a combination of in-house development, Verbund development and outsourcing (see list below).

- GENESIS has been developed as a Verbund project, with the programming work being shared by several offices.
- The RDC-metadata system and the projected output oriented metadata system are spin-offs from GENESIS.
- .BASE was an outsourced development with substantial input to the business case by Destatis' IT-department.
- KlassService redesign is a Verbund project. It is being developed by the Bavarian State Office for Statistics and Data Processing as was the original KlassService.
- Statistikdatenbank is a Destatis project carried out as an in-house development.
- All metadata systems to be developed for the census will be outsourced under the general rules laid out for the IT-development of the census.

5.6 Sharing software components of tools

A major problem in software sharing is language. In most systems the user interfaces are in German only. Maybe more important, few of our systems allow content to be stored in more than one language. Two exceptions are the redesigned KlassService, which supports n-languages, and GENESIS, which supports English content as well as German.

As of yet, there has not been any case of software sharing between Destatis or the Verbund and any external partner. It is not impossible, however. Most of the IT-systems are either owned by Destatis or the Verbund. Any prospective effort to share IT-technology will have to be reviewed by the responsible committees.

5.7 Additional materials

On request (mostly in German).