

# Case study: German Federal Statistical Office

Case study: German Federal Statistical Office German Federal Statistical Office: Use of GSBPM German Federal Statistical Office: use of GSIM

|                        |  |
|------------------------|--|
| <b>Contact person*</b> | Dr. Sven Claußen   |
| <b>Job title</b>       | Assistant Head of section "Cross-Cutting IT Processes Relating to Metadata and Data Quality" |
| <b>Email</b>           | metadata@destatis.de   |
| <b>Telephone</b>       | 0049 (0) 611 / 75 - 1  |

## Summary\*

In Germany the Federal Statistical Office and several Länder Offices are working on building up a metadata management system. The metadata management system shall merge several existing metadata systems and tools already involved in the statistical production. In doing so Germany takes account of international standards like GSBPM and GSIM.

## Metadata strategy

## Organization structure (Germany)

Official statistics in Germany is characterized by a decentralized political system that divides the functions of government between the central government and 16 regional (Länder) governments. As a result, each region (Land) may have a statistical office of its own. Because some Länder merged their statistical offices, there are now 14 Länder offices in total. These offices are organizationally entirely independent institutions.

For the purpose of compiling most federal statistics (those ordered by a federal law), there is a predefined work sharing scheme between the offices. The Federal Statistical Office (Destatis) as the central authority is responsible for the methodological and technical preparation of the surveys and for the compilation and publication of nationwide results. The statistical offices of the Länder are responsible for the core processes of statistical production. Despite that, there are also several cases in which Destatis collects and processes data itself, e.g. in the area of foreign trade.

It is an important feature of the statistical system in Germany that for every major statistical activity a law has first to be passed by parliament. For those statistical activities that involve federal and Länder offices, both the federal parliament and the parliaments of the Länder have to be consulted. The laws usually specify who is legally obliged to provide the information and what information is sought.

Apart from the statistics compiled by the statistical offices, there are also some statistics that are produced by other official agencies. For example, some of the data on unemployment comes from the Federal Employment Agency.

To organize the work in the "German Statistical Verbund" - as the network of federal and Länder offices is known in German - a number of working groups exist with several governing committees on top. Senior governing committees are staffed with the heads of offices. Work sharing also extends to the area of IT-development. IT-systems that are to be used by all offices are put out to tender among the offices in the German Statistical Verbund on the basis of an agreed business case (dubbed the "one for all"-policy).

Destatis itself is mostly organized according to subject areas and therefore resembles the classic case of a stovepipe organization. There are currently five subject matter divisions and three central divisions. The central divisions deal with administration and legal questions (Division A), management, dissemination and coordination issues (Division B) and information technology (Division C). Methodological issues are split between Division B (research and development) and Division C (mathematical methods). More details can be found in the organizational chart.

## Metadata Strategy

Metadata management has been an issue in the statistical system in Germany for many years. Maybe typical for a federal system, solutions have been found and implemented in isolated areas but they have not been coordinated through a common strategy. The current situation therefore resembles a "bottom-up" approach rather than a unified "top-down" solution.

The experience at Destatis and in the German Statistical Verbund, however, shows that there is a strong need for a more coherent approach to handling metadata in the future. Several key projects in the German Statistical Verbund - like standardization of production or quality management - depend on standardized structures and concepts to understand the content of the different statistical activities in a coherent and uniform way. A metadata strategy would also help to provide a framework for the different projects.

Any future metadata strategy would need to be formulated in accordance with at least the most important stakeholders and it would need to be approved by the responsible committees. Therefore, it is not likely to take shape and become formally adopted in the near future. In the near past there were several projects - independently planned and implemented - that involved a centralized metadata management. The task is to combine the projects in a way that at least the outline of a common metadata strategy starts to emerge.

For the purpose of standardization the German Statistical Verbund has instituted a committee of standardization experts which also steers the further progression of metadata projects and activities. The focus was directed at the development of a centralized metadata management system, which interconnect all existing metadata systems and make the contents accessible to production tools.

## Current situation

There is currently one major project that involves centralized metadata management: "SteP - Standardization of Production". SteP is a joint initiative of the German Statistical Verbund to standardize production. A major objective of SteP is the design and deployment of generic IT-tools as building blocks of a standardized IT-landscape. Although SteP currently deals predominantly with IT-issues, a stronger involvement of subject matter experts should strengthen its outreach in the future.

SteP is organized around a simple process model that names the basic processes mainly in the collection and processing stages (see Section 2.3, fig. 2). There are sub working groups (called "steps") dealing with individual aspects of the statistical value chain (see [here](#)). A sub working group for metadata - called "step 12-metadata" has been established in 2008.

In general, SteP has so far been a successful project for Destatis and the German Statistical Verbund. In several of the most urgent areas, production was streamlined and economies of scale could be exploited. Apart from the metadata portal, important ongoing projects within SteP include a database for incoming data.

The task for step 12-metadata is to find a way to harmonize the different IT-systems in a way that the metadata stored can first be accessed and understood by users and secondly be shared by all IT-systems along the value chain. The major project is therefore to develop a centralized metadata management system which interconnects all existing metadata systems and makes the contents in a standardized manner accessible. Possible clients could be a metadata portal for internal users and a metadata portal for external users.

Apart from this, there are several other activities that involve centralized metadata management issues at Destatis. There is for example a close cooperation between quality reporting and metadata management since they overlap in many ways. Production tools already collect quality information during the sub-processes. The metadata management system shall combine these information in the future to assemble quality reports and provide quality indicators.

## Metadata Classification

The RDC-metadata system uses a classification that distinguishes between semantic, technical and administrative metadata. Semantic metadata include definitions of variables and other definitions as well as all kinds of methodological documentation. Technical metadata define metadata on the level of record types. Administrative metadata is mainly information about the responsible persons and institutions.

The RDC-classification reflects the need to classify metadata **according to different levels of abstraction**. Although the RDC-system does not have a separate conceptual level in the sense of the Neuchâtel-model, the term semantic metadata can be seen as synonymous with conceptual metadata. In the future we might need to supplement this classification with a **contextual** level functioning as a mediating level between the **conceptual** and **technical** levels.

With the broadening of Destatis' approach to metadata and the involvement in the Census 2011, it soon became clear that additional classifications had to be introduced to reflect a stronger focus on the statistical process chain. However, it also became clear that there were endless possibilities to structure and classify metadata. Initial experiments with the proposed CMF-classification were made before we realized that Sundgren (2008)\* was right to state that multiple linear classifications of metadata exist and that each of them serves a purpose. Roughly following Sundgren, we decided to classify the metadata **by form** into **structured**, **semi-structured** and **unstructured** metadata. Structured metadata is metadata that exists in metadata systems being structured according to some information or metadata model. Semi-structured metadata exists in the form of written text in a linear order where each text file is the instance of some given template. Typical examples of this kind of metadata are quality reports (or this case study). Unstructured metadata basically consists of text files (methodological documents, etc.) that are structured only on the basis of the author's needs and taste. This classification works fairly well in the census, where most of the metadata is of the unstructured kind.

In addition to using proper classifications\*\* we also distinguish metadata according to user groups and according to attachment objects (like statistical activity or statistical activity instance). These distinctions do not constitute classifications in the strict sense, because - as of yet - we have neither an exhaustive list of user groups nor an exhaustive list of attachment objects (the latter being the same as an overarching exhaustive metadata model). However, we do classify metadata **according to the processes that use or produce these metadata** thereby using the process model as a classification.

Apart from classification, the terms quality metadata and production metadata are used in the office. Quality metadata refers to after the fact interpretation of metadata and applies to all metadata that is deemed important for evaluating data quality. Since such an evaluation must be based on existing metadata (frequently called documentation in this context), the degree to which such metadata exists is itself an important quality indicator and part of quality metadata. Production metadata is a term often heard in connection with software development indicating metadata used to execute and control (sub-)processes in the production of statistical data.

With the development of GSIM there is a big chance to get an international standardization of metadata objects and their classification. The German Statistical Verbund started in 2012 to develop a metadata model which is oriented towards GSIM and adopted the five groups of information objects specifically. Since there are already a lot of metadata tools and systems bringing their own models with them, which are to be considered in general, it is a long way to agree to a standardized metadata model in Germany.

\* Sundgren, Bo (2008): Classifications of Statistical Metadata. Paper presented at the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Luxembourg, 9-11 April 2008.

\*\* Classification meant as a list of mutually exclusive categories that exhaustively classifies each object within its scope according to some explicit or implicit criteria

## Metadata system(s)

### GENESIS (in use)

GENESIS is a cube database used in the German Statistical 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 German Statistical Verbund when it was introduced. The cube model along with the standardized metadata entry forced a new way of thinking onto subject matter statisticians. Constrained by organizational issues - especially coordination in the German Statistical 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 realized. Especially in combination with the centralized 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 German Statistical Verbund. At Destatis it is linked to the press releases so that interested users can search for additional data.

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.

### RDC-Metadata system (in use)

With the establishment of research data centers (RDCs), statistical offices in the German Statistical Verbund began to realize 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 harmonized at a very low level. The variable is modeled 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 modeled 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 about 9,000 variables for the micro data files of 48 statistical activities.

Nevertheless, the RDC-Metadata system has been successfully implemented and is popular with researchers using the data centers. 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 48 of the planned 60 statistical activities.

## **Statistical Portal (in redesign)**

For external users the German Statistical Verbund offers several web sites with specific information (tables, regional outputs and even metadata as well). For example for researchers the RDC Metadata system has been build up. Destatis and the Länder Offices has their own web sites with tables and further information about methodology etc.. The idea behind an "Output oriented metadata system" is to centralize in one single system all offers about metadata which is important for the understanding of statistical products to the public. Since the German Statistical Verbund has already a similar system (including selected outputs and releases) which has to be redesigned in the near future for different reasons Germany plans to combine them in the next release.

The Statistical Portal will contain links to other public systems (e.g. to the dissemination system and the classification server) but also present certain information from internal systems like the Database for Statistical Activities or the Variable Server which is not confidential (e.g. the variable names and their use in statistical activities). A global search will combine the results of several systems which are requested in parallel.

## **Database for Statistical Activities (in use)**

The Database for Statistical Activities stores metadata for all statistical activities at a very high level. It maintains the central catalogue of all statistical activities ("EVAS") of the German Statistical Verbund and is used mainly for management purposes, containing basic information on methodology, legal background, etc.. The system architecture consists of a relational database management system and a single application that will allow accessing and querying the information via the internal web portal of the German Statistical Verbund. As a result, general information on all statistical activities will be visible to all users in the German Statistical Verbund.

Every new statistical activity will first have to be registered in the Database for Statistical Activities and is then identifiable by its unique EVAS-code (registration meant as a business process, not necessarily in a strict IT-sense). The Database for Statistical Activities can easily be amended and combined with other metadata storages at Destatis that use the same EVAS-catalogue like the database used to compile Destatis' Strategy and Programme Plan or internal accounting databases.

## **KlassService (in use)**

KlassService is a tool to browse the classifications and to classify and code answers entered in free text fields in questionnaires. It currently houses five classification series (the German NACE and PRODCOM versions, the Classifications of Occupations, the Classifications of Constructions and the Waste Classifications). 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 German Statistical Verbund.

The KlassService is 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 is also designed to support multiple language versions of the classifications.

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

.BASE ("Base Applications for Statistical Activities") is the umbrella name for several IT-tools - developed for the German Statistical Verbund - to support a standardized 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 surveys in the German Statistical 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. To realize a metadata driven production process by using the survey database, conceptual metadata being stored in classification server or variable databases have to be transported into the production process, and other IT-tools should 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 standardized 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.

## **Census metadata system (finished)**

According to a decision made by the heads of the offices of the German Statistical Verbund, a separate metadata system had to be developed for the Census 2011. The system has been of modular design. Therefore some of the applications could remain in use after the census had been completed and - if applicable - be employed in other statistical activities as well. Issues of census metadata management included:

#### 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 standardized tool to store such documentation in a structured way in the German Statistical Verbund. In order not to change existing work practices, the first measure to be taken for the Census 2011 was to use the Eurostat session management tool "Circa" as a relatively simple document management system. The documents provided were largely be documentation already existing but structured according to the Census process model (which was near to GSBPM).

#### 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 has been realized. It was based on the Neuchâtel Terminology Model (Part II, Variables and related objects). To realize the potential of metadata and reduce duplicated entries, the variable database had been connected to an output database for analysis.

#### Classifications

Several standard classifications have been used in the census. Since these classifications should be stored in the KlassService database, avoiding duplicated entry, these systems had to be linked in some way.

### **Database of Quality Reports (in use)**

The Federal Statistical Office has built up a database to manage the production of quality reports for internal and external use. The database is structured according to the statistical activities in Germany. Responsible employees can fill in structured forms reference metadata especially about the quality of the outputs and the production of the statistics. Regarding the designated recipient of the quality report the information can be restructured and transferred into several formats. The international standards of ESMS and ESQRS are supported. A direct connection to the ESS Metadata Handler of Eurostat is planned to be accomplished in the near future.

Very close to the Database of Quality Reports a similar project deals with the management of paradata which is gathered during the production process directly from the production tools ("QuiV"). In the future the captured paradata can be merged to certain quality reports which describe the quality of single data packages which are arranged for exchange in the German Statistical Verbund. The collection of these quality reports supports the provision of quality indicators required for the measurement of quality within the ESS for example. The QuiV-Project is being drafted at the moment and is planned to be productive in 2017.

### **Metadata Management System (being drafted)**

In the future a centralized metadata management system shall share and support the exchange of information objects along the production chain to interconnect the production tools and systems. Using the standardized interface of the metadata management system the tools will be able to put and get information objects taking account of access rights and with assistance of mapping functions which convert different information models.

Metadata, which are produced along the production chain by one tool, can be reused by another tool in a different sub-process. This way the metadata management system can even control the whole production process.

The metadata management system is already being drafted (including small prototypes) and plans are scheduled to build up the system in the next two years to support the management of quality indicators ("QuiV", see above) and the dissemination of external public metadata ("Statistical Portal", see above).

### **Costs and Benefits**

Metadata management contributes directly to the realization of major objectives in Destatis' corporate strategy. It enables the further standardization of processes, the harmonization of statistics and the monitoring of data quality. Metadata systems help with the documentation of surveys. To ensure that the public trusts in the data which Destatis and the German Statistical 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 standardization of business processes and IT-systems because it represents a common structure for all statistical activities.

The German Statistical Verbund cannot give a detailed overview of the costs accompanied by the metadata activities. In comparison with all the benefits mentioned above the costs may be rated as rather low.

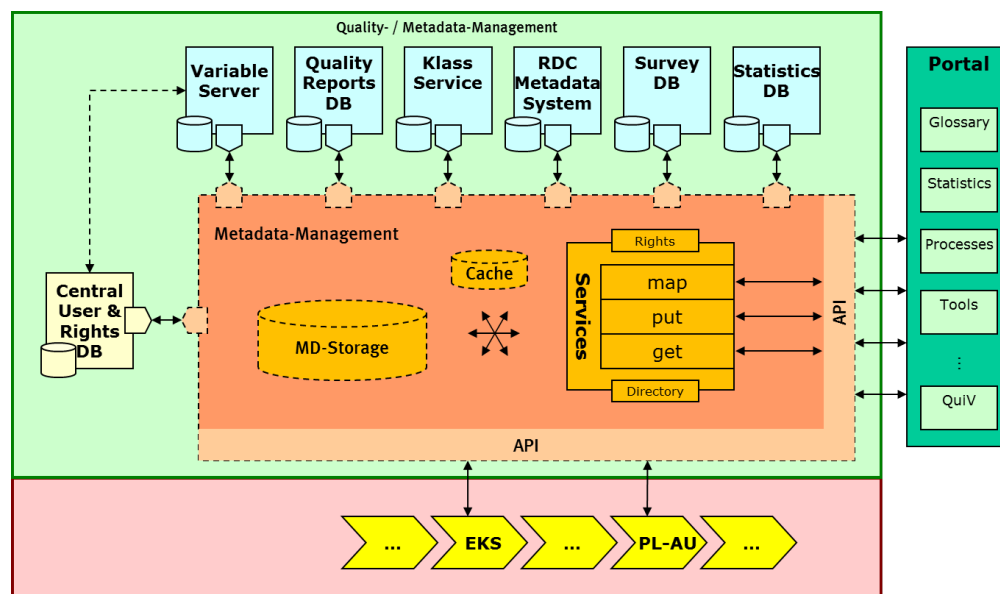
### **Implementation strategy**

The German Statistical Verbund has chosen a step-by-step approach to change smoothly the way how statistics are produced. Many tools and systems already support the production successfully. Therefore any changes and reorganizations should be carefully designed and well-thought. However, to overcome some of the well-known drawbacks of the current production system a centralized metadata management system is designed, which is planned to be implemented in the next two years (2015/2016). Meanwhile a comprehensive information model should be adopted which enhances the current core model significantly. The metadata management system will implement the information model and provide services for a statistical portal and the management of quality information as well by no later than 2017.

### **IT Architecture**

Given that the German Statistical 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 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. In order to connect to the portal, each of the participating applications will need to have web service functions. The metadata management system will interconnect the existing systems and provide centralized functions to receive the requested information objects.



Technically the API-interface is responsible for receiving and distributing the information objects in the right way. External systems which are already contain information objects are connected to the metadata management system via certain adapter which handle the exchange of the objects. Internal services provide support for reception and release of objects especially to take care of access rights and to transfer the objects in different formats regarding the underlying information models. Directory services keep the necessary information about the physical storage of objects and the metadata management even has an internal storage, too. For faster performance caching functions can store objects temporarily.

## 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 1,900 active cube variables in the system at Destatis (for a total of 215 statistical activities) - a figure that has proven to be manageable.

The KlassService contains the classifications necessary for the creation of code lists and for coding the input data. A new variable server was build up to support the production of the census 2011. Although not every functionality was used the variable server could be to be standardized and used in other statistics in the future. Besides variables the server is able to manage statistical units, code lists, measurements etc.

## 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 XML-format "GenML", 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. The import and export of data files in the CLASET format is supported. There is no standard format used in the Statistical Database. Nevertheless, the understanding of the term statistical activity is much the same as in GSIM. Therefore, the Statistical Database can interact with other systems within a distributed metadata systems that allocates different functions to different systems according to GSIM. The database for quality reports "QDB" supports the generation of quality reports in the ESMS and ESQRS format.

## Version control and revisions

In theory, there seem to be two different ways to control the versions of 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.

## Outsourcing versus in-house development

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

- GENESIS has been developed as a German Statistical 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 German Statistical Verbund project. It is being developed by the Bavarian State Office for Statistics and Data Processing as was the original KlassService.
- The Statistical Database is a Destatis project carried out as an in-house development.
- All metadata systems developed for the census have been outsourced under the general rules laid out for the IT-development of the census. Therefore the variable server "MMS" (as part of the census project) was also outsourced.
- The database for quality reports was developed by Destatis as an in-house-project.
- The metadata management system will be a German Statistical Verbund project. The contractor is unknown at this moment though since the invitation for bidding isn't published yet. Furthermore the development shall be coupled with the development of the statistical portal and / or the development of the prospective system for the management of quality indicators.

## 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, only a 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 German Statistical Verbund and any external partner. It is not impossible, however. Most of the IT-systems are either owned by Destatis or the German Statistical Verbund. On the other hand there are restrictions by national laws regarding budgetary rights stipulating that it is not allowed to transfer ownership of software for which the development is publicly financed. Any prospective effort to share IT-technology will have to be reviewed by the responsible committees.

## Overview of roles and responsibilities

The coordination of the work related to the management of metadata is done by a small metadata management team at Destatis. The team is part of the IT department and consists mainly of IT specialists.

Metadata used by Destatis' output database (GENESIS) is coordinated by the dissemination department, where an editorial team reviews the cube descriptions. Similar solutions are used at the Länder offices.

Metadata that describes public and scientific use files is stored in the RDC-metadata system. The RDC-system is jointly maintained by Destatis' RDC and the RDCs of the Länder offices.

Production metadata is being cared for by the subject matter departments. Different roles exist in the layout of the .BASE-system.

## Metadata management team

In the German Statistical Verbund there is mainly one expert group, which is responsible for coordinating metadata related activities in the German Statistical Verbund. Above this expert group a steering committee (SteP) facing standardization issues guides the work. At Destatis in 2012 a small unit in the IT department was established dealing with metadata and quality. This unit currently consists of 5 people who are developing specifications as well as prototypes and systems for the management of metadata and quality information. This unit is responsible for the technical issues of SDMX as well.

## Training and knowledge management

There are no courses entirely dedicated to metadata management. Metadata issues are covered by courses on the GENESIS cube database. The various RDCs run courses to introduce internal users to the RDC-metadata system. General training on the topic of metadata is planned in the medium term.

## Partnerships and cooperation

All current metadata management projects are mainly carried out in the German Statistical Verbund participating Destatis and the Länder Offices. In 2008 Norway and Switzerland supported the German Statistical Verbund mainly to build up a classification and a variable server though.

## Other issues

## Lessons learned

- Metadata management is a communication challenge. We found two issues were particularly difficult to communicate:
  1. Metadata management is tricky. Statistical data is inherently volatile. For any given data, an endless number of transformations are possible producing an endless amount of metadata. With the distribution of modern IT-systems there is hardly any limit to producing endless variations of the same dataset.
  2. Metadata can be more than just documentation. The same information that is used to transform (produce) data can be used to document it and vice versa.
- As we are faced with multiple stakeholders, several isolated decisions taken by governing committees and a variety of IT-systems in place, it would probably be useful to develop a metadata strategy. Such a strategy might help the organization to focus on important projects and provide a coordinated approach ensuring that systems are able to interact. Distributing the energy of an organization across too many

unrelated tasks easily drains away resources without delivering satisfactory results. Drafting such a strategy, however, also consumes resources and requires a deeper understanding of the problems.

- The advantages and disadvantages of a metadata model can often only be properly evaluated once an IT-system is in place. It is therefore important to learn from evaluations of existing systems.
- Considerable effort went into formulating metadata models. Having evaluated some of them, we feel that the existing models do bear some similarities. A perfect model may not exist, especially since the resulting implementation usually involves some compromise. No database can be endlessly complex. But on a more conceptual level there seems to be some convergence. Indeed there might even be a structure inherent to the metadata of (official) statistics. Thus, the quest for the "real" metadata model might be less a matter of design than of discovery.
- In a federal system, national coordination usually requires a lot of resources from all partners in the system. Understandably, international cooperation is then often seen as being of lesser importance. Despite this, international cooperation has substantially helped the metadata team at Destatis in the past to understand the subject of metadata management. The development of IT-systems consumes a lot of resources. We feel it helps to build on existing international knowledge and that it minimizes risks and maximizes return on investment.
- 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.

|   |
|---|
| <b>Links:</b>                               |
| <a href="#">DESTATIS Organisation Chart</a> |
| <a href="#">GENESIS-Online</a>              |
| <a href="#">RDC Metadata System</a>         |

## Attachments