1. Introduction (Statistics Austria)

back to case study home page

1.1 Metadata strategy

History

At Statistics Austria the development of cross-domain metadata systems already began in the early 1970s, with the statistical output database ISIS (Integrated Statistical Information System) which is still in use (see section 4.1). When developing the application DOK (no longer in use) in the 80s, the main focus of interest was on technical metadata (files, programs, variables, code lists), and with WIKNACE 10 years later the first version of a classification database was implemented (a mainframe application limited to the Austrian version of NACE which was replaced by an MS-Access database at the end of the 90s; finally in 2006 a Web application containing approximately 20 major economic classifications was completed).

For many of these projects the IT department can be seen as the main driver. One reason for that may be the pronounced "stovepipe" organisation of the statistical production processes at Statistics Austria: as a crossing point for many projects, the IT department has a more general view not limited to single surveys.

The concept of BASIS 2000+

Theoretical considerations in the field of "Metadata" began in the middle of the 90s. In 1996/1997 two members of a sub-unit of the IT department (mainly engaged in consulting and cross-sectional projects) developed a concept for an integrated metadata system named BASIS 2000+ (Metadata **Ba**sed **S**tatistical **I**nformation **S**ystem).

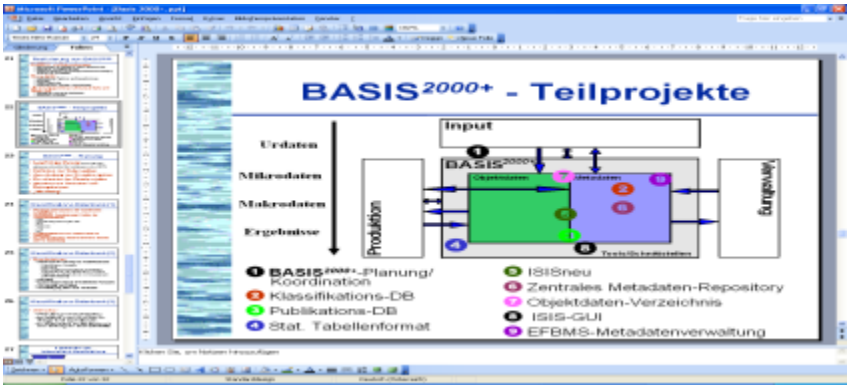

Figure 2: Overview diagram of BASIS 2000+

Some of the general aims of the concept were:
to create extensive, flexible, open, simple and user-friendly access to the object and meta data for both NSI internal and external "statistics users";

•to achieve planned collection, storage and (multiple) use of meta data (which means that they have to be standardised and harmonised);
•to establish norms for data management in general and for interfaces between the software products used for producing statistics;
•to provide support to users via general-use tools and to enforce global solutions covering the entire statistical office, instead of insular solutions or double and multiple developments;
•and finally, always to take into account the diverging and in some cases unknown or unpredictable needs of different user groups.

The main principle of the system architecture was: no single monolithic application, but a modular design consisting of subsystems which can be implemented and enhanced consecutively, based on common standardized data and metadata objects.

The concept was regarded with high interest. But with a view to the high amount of resources needed and the expected rearrangement of production processes, the realization was postponed.
However, by designing and realizing a metadata-driven electronic questionnaire system (e Quest) by the "fathers" of the concept described

above, several basic ideas of BASIS 2000+ were brought to life.

Back on the agenda: STAT+

The legal status of official statistics changed at the beginning of the new millennium. The former Austrian Central Statistical Office was separated from the federal civil service and Statistics Austria was created as an independent, non-profit federal institution under public law. Along with the modernization and the reorganization of Statistics Austria the concept of BASIS 2000+ was picked up again in an "umbrella project" called STAT+. A working group in cooperation with Professor Karl Froeschl, a well-known expert in the field of metadata, refined the general concept and elaborated a business process model called the "4-layer-model". A list of projects which should be implemented with high priority was a second result of the working group. However, no organizational unit existed (nor does one exist up to now) which is concerned with metadata or with any other harmonization project (such a unit would also be responsible for designing and coordinating the development of integrated metadata systems).

Implementation of projects

As a consequence of the developments described above, the following projects were carried out (for further description see section 2.1).

•Statistical Table Format STF
•Publication Database
•e-Quest/Web
•Classification Database

Over the years several other projects were launched. Some examples:

•With the re-launch of Statistics Austria's Web site in 2007, external users have been provided with easy access to comprehensive metadata (the standard documentation files and quality reports; see next sub-chapter).
•An online publication directory presents all print publications available at Statistics Austria.
•For internal use only, an MS-Access database of administrative data has been developed.
To a large extent, however, the implemented metadata systems are isolated solutions and not integrated with each other.

Quality reporting

Quality reporting has been a topic at Statistics Austria since the end of the 90s. In order to collect and manage information about the quality dimensions, an MS-Access based application called SYSQUAST was developed. However, due to problems in maintenance (primarily synchronization problems) the system never left the prototype phase.

When at the beginning of 2000 the new legal form of official statistics came into force, one decisive consequence was that TQM (Total Quality Management) was established. Since product quality can be seen as one of the cornerstones of TQM, the creation of a detailed quality reporting system became indispensable. For a producer of statistics (for instance a survey manager) the compilation of a detailed quality report means a heavy work load. For that reason the intention was to create a standardized Statistics-Austria-wide documentation system for statistical projects, into which the process of quality reporting is incorporated.

The problem of isolated and unstructured metadata

In international discussions it is generally acknowledged that metadata play a decisive role both in satisfying statistics users' growing quality requirements and in increasing the efficiency of the internal production processes within an NSI (national statistical institute).
In recent years relevant technical publications have repeatedly stressed that the implementation of metadata systems must be founded on a comprehensive and general model of statistics production and on a long-term master plan (the term "metadata strategy" is often used in this context). Paying too little attention to these preconditions leads to metadata systems which are neither linked with each other nor with the data they document and which lack the ability to cooperate with each other. Often, the same information is stored repeatedly, rendering it difficult to keep the metadata consistent and causing unnecessary effort and costs. In the worst case, the resulting applications rely on mutually incompatible concepts and models, making integrating them ex post an extremely demanding if not impossible task. As with many other NSIs, Statistics Austria finds itself in exactly that situation.
Another problem in addition to that of the single solution approach is that metadata generated during the planning, implementation or execution of a statistical project (the term "statistical project" is here taken to denote all types of statistical production systems - primary and secondary surveys, registers and analytical projects) within the separate stovepipe production systems are often written into working documents or are included in non-structured form in various print publications. It is therefore difficult for internal and external users to locate them; in the worst case, they cannot be accessed at all.

The IMS project

The Federal Statistics Act 2000 installed the so-called "Statistikrat" ("Statistical Council"). This functions as the highest-level body for quality assurance of federal statistics. One of the tasks of the Council's 15 members is to elaborate comments and recommendations pertaining to the statistical work programme.

In several of its comments the Statistical Council has explicitly drawn attention to the importance of delivering comprehensive metadata and of increasing the statistical system's coherence, and has demanded the development of a metadata repository. In this context it has also underlined the central role which the IT department should fulfil in "implementing the requirements repeatedly voiced by the Statistical Council for uniform information delivery, increased quality, enhanced timeliness, easier data access and provision of more comprehensive metadata" (quote from the position paper of the Statistical Council pertaining to the work programme 2007, p. 12)

In 2006 an IT project was commenced (working title: IMS - Integrated Metadata System), the goal of which was to conceive an "integrated metadata repository" based on best practises and international recommendations and to prepare an overall plan for implementing such an information system.
In order to make quick progress in the project and with regard to the limited budgetary and personnel resources,

Statistics Austria's top management decided to reduce the scope of the conceptual tasks in the IMS project. The goals and consequently the basic focus of the project were thus stipulated as follows:

"The goal of the system to be developed is to deliver to Statistics Austria's customers (various external users, national and international organisations) that functionality which they require in order to satisfy their needs with regard to statistical information (e.g., to understand statistical results and to have the means to judge their quality). One can start from the assumption that the functional range implemented internationally as "best practise" in various statistical offices will cover the customers' requirements.
Not only external users should profit from the metainformation system. Internal users of statistics also require the metainformation relevant to statistical products and processes (e.g., in order to be able to efficiently reuse statistics produced by the Office or to process them further for specific projects). It can be assumed that a metadata repository will also generate internal benefits with respect to efficiency and quality of the statistical production process."

With the above as the fundamental goals, the main focus of the IMS project was placed on passive metadata (see section 3.1), which are required both by external and internal statistics users, in particular for the functions "finding" and "interpreting" statistical data. With an eye to this, the metadata repository was conceived as a comprehensive documentation system for statistical data and production processes.

The concept provides for collecting the metadata which are hitherto scattered over various production systems and (working) documents, storing them in structured form according to a general model of statistics production, and integrating them by allowing links to be created between the individual elements of documentation and the data they describe.

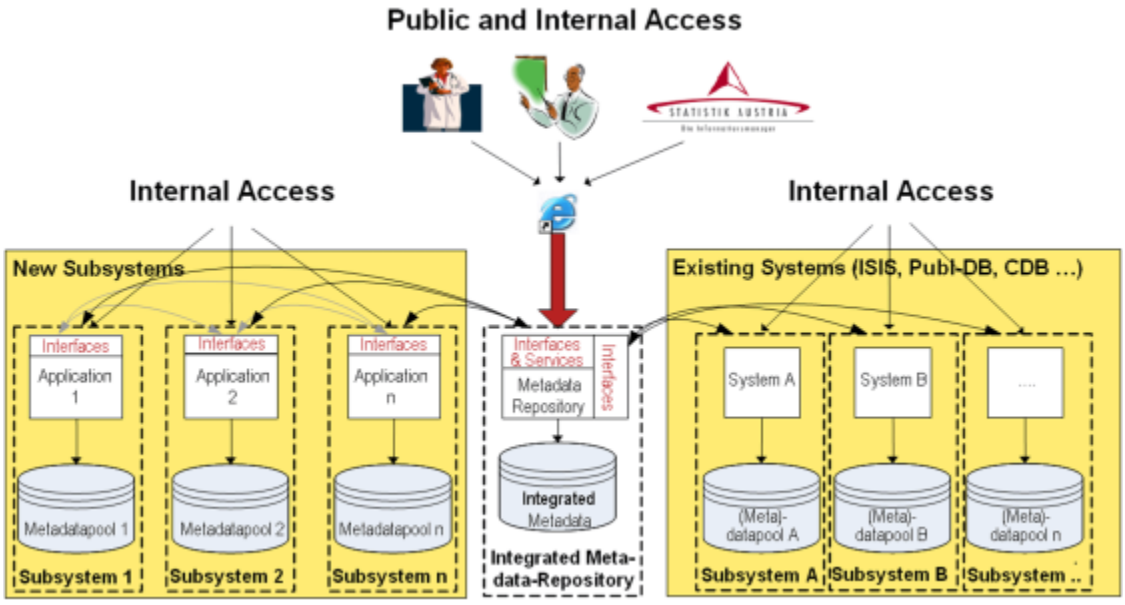

Figure 3: IMS architecture

The right side of figure 3 symbolises the cross-domain applications and stand-alone metadata systems which already exist (ISIS, Classification Database, Publication Database, etc.) and which must be connected to the overall system. The left part shows new subsystems which must be developed. These do not replace any existing IT systems, but are responsible for central and structured entry and consolidation of metadata which at present are scattered over various sources. The proposed sub-systems are:

1.Definitions and concepts
2.Statistical projects
3.Types of statistical units and their characteristics (variables)
4.Value domains

The latter two subsystems are based on ISO 11179, although during the modelling process some areas were simplified and others enhanced as compared to this standard.

The subsystems which are to be developed will communicate with each other (and also with the Integrated Metadata Repository IMR, particularly with its component "Registry" - see below) via web services. In this way the mutual interdependencies between the individual subsystems can be minimized, with an eye to the concept of encapsulation.

The Integrated Metadata Repository (IMR) occupies a central position between all these systems. It consists of two parts, the "Registry" and the "Catalogue". The latter contains all those metadata which should be accessible for external users over the Web. These will normally consist of a subset of the metadata administered in the IMS subsystems, as the latter will also contain information which is only of interest to the subject matter persons responsible for statistics production. The Catalogue would also allow a comprehensive full text search over all subsystems, i.e. with a single search request the user should be able to locate not only the documents and Web pages stored in the Publication Database, but also data and metadata in ISIS, the Classification Database and the future IMS subsystems.

The second component of the IMR, the "Registry", is responsible for a Statistics Austria-wide unique registration of all the information objects contained in the individual subsystems and in the connected legacy applications. In addition, it administers links between information objects, which will be of various types (e.g., "contains data from" between a table or an ISIS data cube and one or more surveys). These two core functions are prerequisites for allowing the users to navigate from one data or metadata object to another along predefined paths. E.g., starting from a list of types of statistical units such as enterprise, household, birth, etc., one might quickly locate the characteristics which were collected or created during statistics production, "surf" from there to the corresponding value domains or to the surveys / survey versions, to definitions etc.

Further tasks of the Registry are to provide central services required for more than one subsystem (such as administration of users and access rights, status of the registered information objects, ownership of administered items) and the publication of metadata to the Catalogue.
Work on the IMS proceeded in two sub-projects and in cooperation with external partners, among them once again

Professor Karl Froeschl. In the first sub-project subject matter departments were invited to participate in several workshops. The aim of these workshops was to analyze the types of metadata which are used and produced during the statistical production processes of some selected surveys. Although the project leaders were aware of the fact that conducting such a project without strong involvement of subject matter experts is problematic, because of time constraints and the demand for quick results a further integration of the subject matter departments was not possible.

During the second sub-project the specifications of the proposed subsystems were refined in the form of use-case and class diagrams. Top management received regular reports on the progress of the project.

Critical aspects

After discussing the results of the IMS project with top management, subject-matter departments and the Statistical Council it was found necessary to react to the following critical points

•The cost factor:
Considering the uncertainties with respect to the required budgetary and personal resources and the limitations of these resources, a decision on the IMS in its complete stage of expansion was not possible at this stage. Therefore it was decided to start with a single subsystem.

•Stronger involvement of subject matter departments:
Integrating experts from subject matter departments into the development process was seen as indispensable. For this reason, a working group was formed in the autumn of 2008, consisting of the following members:

•Project leader: Head of population statistics (the former secretary general and former head of Quality Management)

•3 persons from the IT department
-1 metadata expert
-1 database expert
-1 external software engineer

•Head of Quality Management

•Expert for classification systems

The mandate of the working group is to discuss and specify the contents and functionality of the "Concepts and Definitions" subsystem of the IMS. This system will allow the centralized collection and administration of various definitions relevant to the production of statistics. By integrating the database with the Web content management system, these definitions should be presented to external users on a new metadata portal on Statistics Austria's homepage.

To prevent yet another isolated solution being developed, special emphasis has to be laid on the possibilities of integration and enhancement as they were defined in the architectural design of the IMS. This means that along with the implementation of the "Concepts and Definitions" database the most important components of the Registry and a part of the Catalogue must be realized as well. This should guarantee that - when the system is enlarged by other important modules later on - this can be done without major effort, which means that there will be no obstacles to further development.

Among other results, the working group has elaborated a proposal for a metadata portal and a prototype of the internal metadata management tool (which of course should also be easily extensible). The next step will be to write a detailed requirements analysis document, on the basis of which a precise cost estimate will be made.
If the costs stay within the scope of the available budget, the system will probably be implemented by an external software company.

The working group has also started to debate a second subsystem for the management of metadata which documents statistical projects (as a successor to the standard documentation files). This subsystem should also be able to fulfil EU requirements regarding the documentation of statistics by integrating the Euro SDMX Metadata Structure ESMS.

1.2 Current situation

At Statistics Austria, a written and formally adopted corporate metadata strategy does not exist, but the IMS concept could serve as a fundamental cornerstone of such a strategy.
In comparison to international best practices, a number of weak points were identified in the IMS project.

- No standardised processing of raw data

The pronounced stovepipe architecture of statistics production in Statistics Austria leads to major differences between the workflows and tools utilized in different statistical projects. Thus, for example, there is no standardised data storage for raw data and validated microdata: the formats range from sequential files with varying record structures on the mainframe over relational databases to data sets in varying PC formats. Apart from the difficulty that information about the various raw and authentic data sets and their structure can often be discovered only by asking the responsible subject matter experts and/or programmers, the lack of standardisation in this area also requires important processing tools such as editors and validation programs to be tailored to each stovepipe separately, making development and
maintenance inefficient.

- Metadata not standardised and mostly unstructured

The principle often underlined in relevant technical literature, that metadata should be captured in a standardised form at the moment they come into existence, and thereafter should be reused, is only inadequately put into practice at present. On the contrary, metadata are often stored redundantly in work documents or appear in the continuous text of bulky documents. In this form, it is difficult or impossible for internal and external users to access them, and IT access to individual documentation elements or attributes of metadata objects is normally also impossible. This not only causes opportunity costs and additional effort, but also damages the coherence of the statistical system.

- Essential metadata systems missing; isolated applications

Most of the existing metadata systems are isolated applications. Additionally, essential metadata systems such as databases for definitions, datasets, variables, or value domains are missing.

- Some limitations to online-user functionality

A further weak point concerns the search for data and metadata. Since the Web re-launch in 2007, a full text search facility restricted to documents in the Publication Database (including Web pages) is available; however, it does not cover the statistical output database ISIS and the Classification Database. Searching may therefore be cumbersome and time-consuming, as it may entail multiple search requests to multiple systems (which may not even be known to the lay person) with differing search mechanisms and user interfaces.

As regards other searching tools, since the Web re-launch a list of keywords (index) and a thematic search based on a hierarchical topic tree exist, but again these refer only to the Publication Database. Other information systems such as ISIS possess a differing thematic structure. Searching for data based on a list of statistical projects or on types of statistical units and their characteristics is not supported.
With respect to linkages between data and metadata (which would allow users of statistics to navigate quickly and easily within a "semantic net" between objects of various types) the Web re-launch and the use of the Publication

Database as Web content management system have brought considerable progress - for example, links from a Web page to related print publications, standard documentations or press releases are displayed. However, these links are created indirectly by associating topics from the topic/navigation tree defined in the Publication Database to the individual documents. Specific information objects cannot be connected explicitly via various types of relationships (e.g., the relationship "is published in" between a table in Excel format and a print publication). Links to information objects which are not checked into the Publication Database as documents are also impossible to create.

Projects in progress

Apart from the metadata working group described above, the following metadata-related projects are running at present:

- **ISIS New:** replacement of the statistical output database ISIS, based on the Australian software SuperSTAR.
- **e-Quest New:** up-to-date version of the electronic questionnaire system e-Quest, implemented as a Java application based on Eclipse RCP (Rich Client Platform).
- **New Business Register**: a completely new version of the business register.