WP. 10
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD workshop on statistical metadata (METIS)
(Lisbon, 11-13 March 2009)

Session 4 Metadata case studies

METADATA CASE STUDY

Submitted by Statistics Austria[1]

I. INTRODUCTION

A. 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 2.2). 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).

2For 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).

[1] Prepared by Thomas Burg (thomas.burg@statistik.gv.at), Wolfgang Koller (wolfgang.koller@statistik.gv.at) and Guenther Zettl (guenther.zettl@statistik.gv.at)

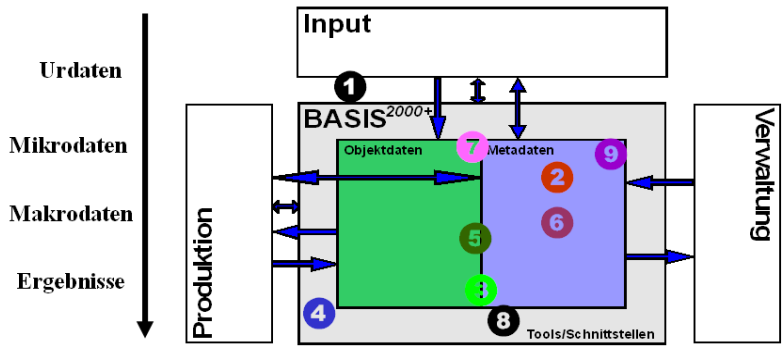

Figure 1: Overview diagram of BASIS 2000+

4Some 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 2.2).

- 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.

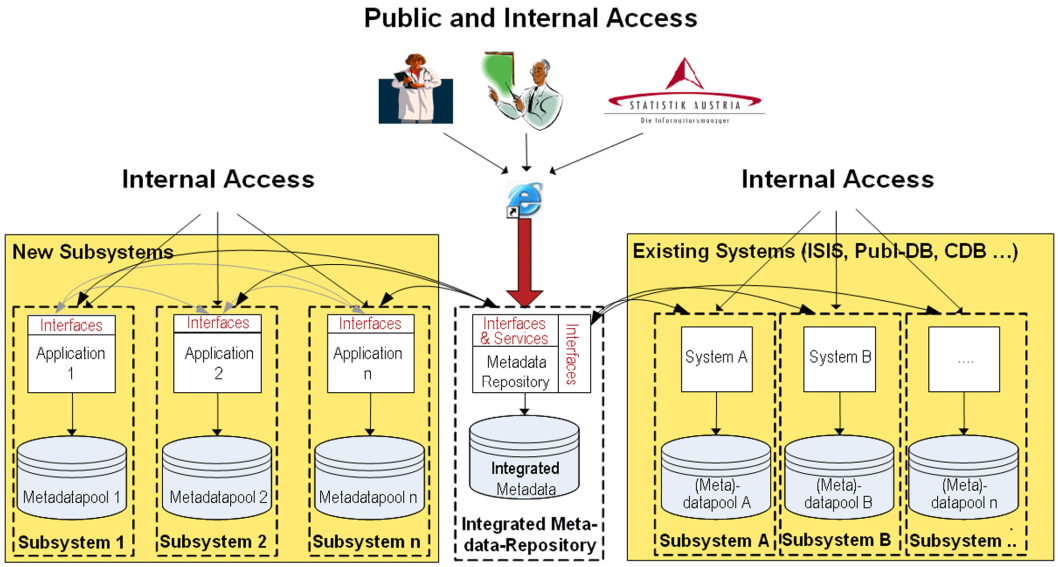

Figure 2: IMS architecture

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.

B. 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.

II. STATISTICAL METADATA SYSTEMS AND THE STATISTICAL BUSINESS PROCESS

A. STATISTICAL BUSINESS PROCESS MODEL

Within the framework of the STAT+ project, a model of the statistical life cycle (called the "4-layer-model" because of the four data systems it defines) was elaborated at Statistics Austria in 2002. The model distinguishes between the following types of statistical projects: surveys, registers and analytical systems.

Surveys are the most "typical" and most commonly occurring form of statistical projects at Statistics Austria. One can differentiate between primary surveys (in which the collection of raw data is one of the steps of the process) and secondary surveys (which process data which already exist and often were collected for non-statistical purposes). There are also mixed types, e.g. surveys in which data from secondary sources are used to augment the data collected by questionnaire. Some surveys are only undertaken once, others are repeated at regular or irregular intervals – although the sets of variables collected in each repetition do not have to be identical. It is therefore useful to further subdivide the survey structure: each survey consists of one or more survey versions, and each survey version consists of one or more survey instances, i.e. concrete executions. E.g., in the case of a survey with monthly periodicity the data collection of each year might be seen as a new survey version with twelve survey instances.

In contrast to the data of a survey instance, which pertain to a certain reference date or period, **registers** are usually updated continuously. Maintaining a register is thus a core process that is typical for a statistical project

of type “register” but unknown for projects of type “survey”. Another fundamental difference is that register data are used as resources for workflows in other statistical projects, e.g. when drawing samples, for addressing, registration of incoming questionnaires and administration of reminders. Commonly, specific (database) applications are developed to carry out these functions.

Analytical projects and systems (as, for instance, national accounts) characteristically do not collect raw data on specific observation units, but use data from other statistical projects and evaluate them or combine them into a coherent, integrated model.

Data which form the input to a statistical project (or which are collected in an early phase of processing, in the case of primary surveys) often pertain to individual observation units, e.g. individual persons, households, enterprises, events, etc., and are termed “microdata”. However, the input may also consist of macrodata, i.e. data pertaining to collectives. In addition to these, metadata also enter into a statistical project and form an important resource for the steps carried out in its processing.

The output of a statistical project consists predominantly of macrodata – cross-classified tables, multidimensional data cubes and time series being the most important categories – and metadata. More rarely, (anonymized) sets of microdata may be produced. Macrodata and certain related metadata are often combined into an “information object” (e.g., a press release consisting of a table and descriptive text). Such information objects may also be composites of smaller information objects (as with a printed publication containing several tables and descriptive metadata – e.g., analytical texts).

The Statistics Austria model of the statistical life cycle distinguishes between the following phases in the production of statistical information (this description applies to statistical projects of the type “survey”. Registers exhibit different core processes – creation, maintenance, and use of the register –, although the contents of a register may also form the basis for production of statistics and information, which can be identified with the relevant phases of a survey. The line between surveys and analytical projects is not always clearly defined – in fact one could certainly argue that analytical projects are a special type of statistical survey creating statistical information from input data, the difference being mostly that methods are applied which differ from those used in “typical” surveys.):

Phase 1: Planning, design and system development

The survey is set up in this phase. Given specific requirements (e.g. EU regulations) and the information needs of internal and external parties (e.g., statistics users in government and the economy), the survey must be prepared to satisfy these as best possible while simultaneously minimizing the effort of the statistics producers and the burden on the data providers.

Output of this phase are metadata of various kinds – e.g., description of the survey’s goals; description of the characteristics to be surveyed; definitions; value domains; classifications; questionnaires and comments explaining them; list of validation rules, etc. The metadata created in a survey may form the input to other phases or other surveys and be reused there.

The development of tools for actually conducting of the survey (e.g., electronic questionnaires, editing software, programs for checking consistency and plausibility) is also a component of the first phase.

Phase 2: Data production

Whereas the decisions taken in the design phase and the metadata and tools which are created there apply to the whole survey or at least a survey version, the focus of the activities undertaken in the following phases lies mostly on the current survey instance (excepting activities in which data from more than one survey instance are processed, as in the creation of time series).

Data production can be subdivided into three sub-phases:

- In pre-production activities such as drawing the sample, printing and posting paper questionnaires, loading Web questionnaires with respondent-specific initial data, etc. are undertaken.

- The actual survey/measurement/observation of the statistical raw data is termed core production. This includes conducting interviews, filling in paper or electronic questionnaires, registering and roughly checking questionnaires which arrive, mailing reminders, data entry from paper questionnaires, etc. In secondary surveys, this sub-phase includes acquisition of the secondary data and, if necessary, reformatting or recoding it. The collected data are stored in the so-called Raw Data System (RDS).

- Post-production includes all activities necessary to improve the quality of the raw data. Among these activities are validation and consistency checks, examination and correction of dubious information, and imputation of missing values. The results of this phase are the "authentic" data (ADS: Authentic Data System); of these several versions may exist, especially in complex and voluminous surveys (e.g., preliminary version at a certain cut-off date and final version at a later date).

A large part of the metadata created during the design phase enters the second phase as input. As also in later phases, however, new metadata also are produced (e.g., the answer rate, which is an attribute of the survey instance).

Phase 3: Statistics production

In this phase, the contents of the Authentic Data System (consisting mainly of microdata) are processed further. To do this, data from other surveys may occasionally be accessed. Some of the processing steps undertaken are aggregation into macrodata, calculation of statistical measures and indices, diverse methods for increasing quality and comparability of statistical information (e.g., seasonal adjustments), and creation of time series. The results are data sets which are stored in the Statistical Data System (SDS) and are at the disposal of internal, often also external users. The SDS primarily contains multi-dimensional data cubes, although anonymised sets of microdata may occasionally also be created in this phase.

In part, the transformations which are carried out here have already been planned in the design phase and are applied to the data of each survey instance. Partly, however, ad hoc analyses may also take place, which use the existing data material in ways not foreseen when the survey was planned. This underlines the importance of comprehensive and easily accessible documentation of all a survey's design decisions, of the data sets and of the transformation processes (in whatever phase of the statistical life cycle they may be created or executed).

Phase 4: Information production

In this last phase, "information objects" such as tables, charts, articles, press releases etc. are created from the data stored in the ADS and the SDS and their metadata and disseminated via various media (internet, print publications, etc.).

The following figure presents the phases described above, the data systems, the registers and the meta-information systems. On the one hand, the latter provide input and various services for the activities carried out during the statistical life cycle, on the other they also accept the metadata created as output from each phase. Thus metadata systems and registers form an infrastructure layer accompanying the whole production process. The data systems are drawn as broader than the "process arrows" in order to point out that they contain data from various surveys and that an individual phase of a survey may accept input data from more than one statistical project.

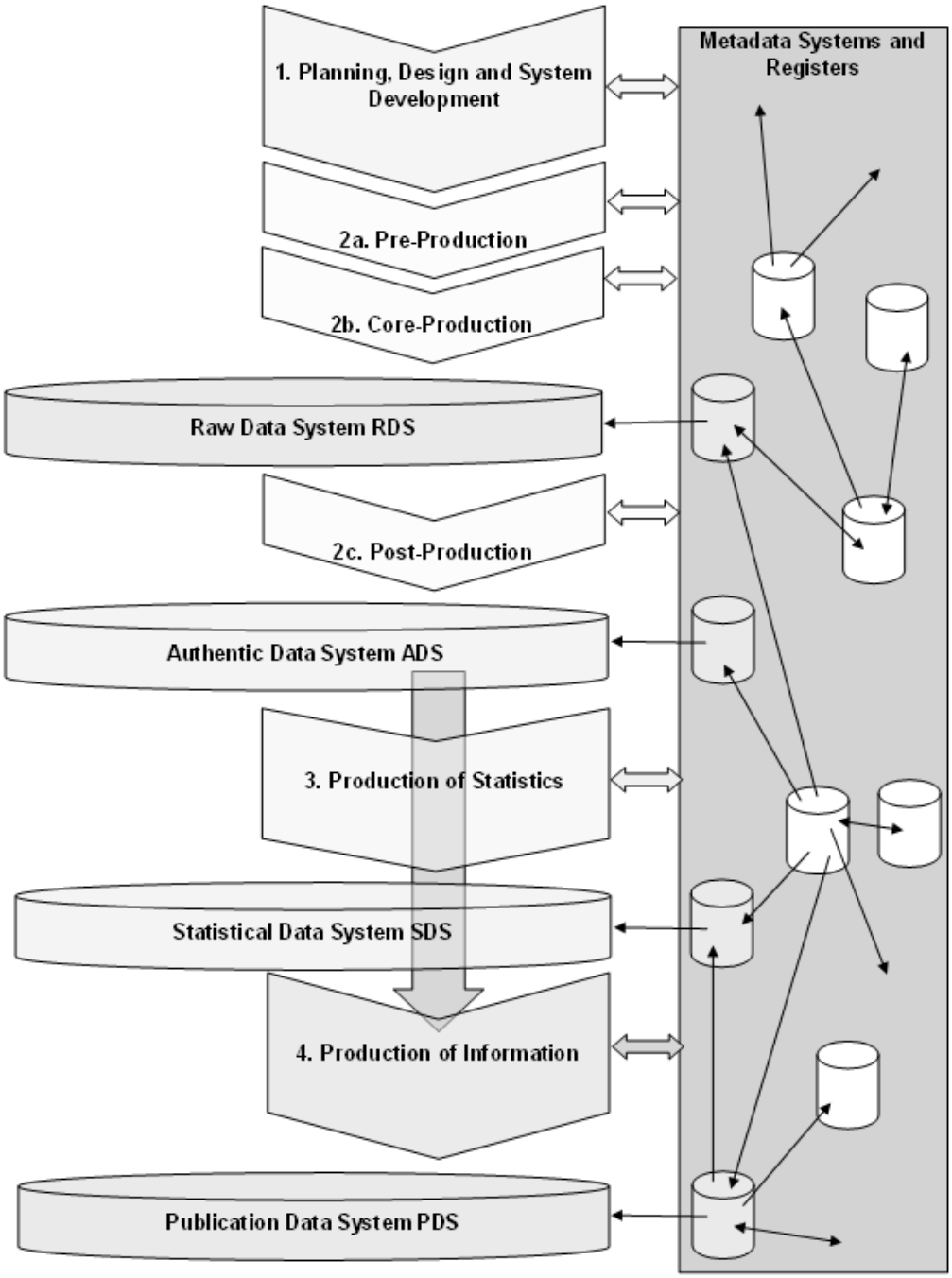

Figure 3: "4-layer-model"

In actual fact the workflows are of course not quite as linear as the figure might suggest; on the contrary, complex control flows (branches, loops) often occur. Moreover, events in later phases may have retroactive effects on the survey's design and lead to adjustment of the current or future survey instances (e.g., changes to the validation rules).

Phases 1 to 7 of the Generic Statistical Business Process Model ("Specify Needs", "Design", "Build", "Collect", "Process", "Analyse" and "Disseminate") can be mapped to the "4-layer-model". The phase "Evaluation" has been considered as an ongoing process, but is not explicitly mentioned in the model. The over-arching process of "Metadata Management" is represented by the "Metadata systems and registers infrastructure layer". "Archiving" and "Quality Management" are not part of the model.

The Generic Statistical Business Process Model incorporates many more details than the 4-layer-model. Therefore the GSBPM (with the addition of the four data systems) would appear to be appropriate for use in future metadata projects.

B. CURRENT SYSTEM(S)

- **ISIS:**

ISIS (short for Integrated Statistical Information System) is a statistical output database which was already developed in the early 1970s and has been consistently maintained and developed further since then. It contains thousands of multi-dimensional data cubes as well as metadata of various kinds (e.g., short descriptions of the data cubes and the underlying surveys; keywords and a hierarchically structured topic tree are furnished for data searching) and implements a large part of the Statistical Data System SDS in the life cycle model. Although ISIS is still very modern from the point of view of the conceptual design of its contents, the software itself has reached the end of its life span, as only one programmer now still possesses sufficient technical know-how to maintain the mainframe Assembler and PL/I programs. Because of this, a successor system (ISIS New) is currently being developed on the basis of the Australian company Space-Time Research's SuperSTAR product range.

- **e-Quest:**

e-Quest is a system consisting of several tools for metadata-driven generation of electronic questionnaires, administering them and preliminary processing of the incoming questionnaires. Subject matter experts can design the questionnaires with a user-friendly graphical editor. The active metadata thus specified are stored in XML format and then used to represent the questionnaires dynamically in a Visual Basic 6.0 rich client application (which must be installed by the respondents) on the one hand; on the other hand, however, they are utilized for generating Java and Javascript source code, JSP pages and SQL table definitions for electronic questionnaires accessible via a uniform Web questionnaire portal. e-Quest thus covers important areas of phase "data production".

Currently the project "e-Quest New" is running with the goal of replacing the Visual Basic components by a Java-based solution. Simultaneously, better integration of the stand-alone and Web questionnaire subsystems is being aimed for.

- **Publication Database:**

Using document management software from the company Stellent (which since has been acquired by Oracle) the publication data system PDS was created during the last few years. This stores all publications (i.e., documents of various types, from tables over print publications and press releases to the so-called standard documentations) together with metadata relating to the documents. Since the Web re-launch on June 1st 2007, Stellent is also utilized as a Web content management system. The subject matter experts now create Web pages in the form of standardized Word documents which are automatically converted to HTML and copied to the correct position in Statistics Austria's website on the basis of associated metadata (in particular a hierarchical topic and navigation structure). The navigation structure is also used for generating links to related documents with data and metadata. The online directory of print publications (many of which can be downloaded free of charge as PDF files) was also implemented in the Stellent system.

- **Classification Database:**

In 2006 the Classification Database KDB was released. This allows Web access to almost 20 voluminous classifications such as PRODCOM, NACE, COICOP, SITC and CPA, including comments and correspondences. More than one version is available for several classifications.

Up to now an application for interactive editing and processing of classifications has not been developed.

- **Statistical Table Format STF:**

STF is an XML specification which permits cross-classified tables to be stored together with extensive metadata in a hardware- and application-independent format – for long-term storage, among other uses. Converters from STF to Excel and HTML and from Excel tables to STF are supplied. When Excel tables are checked into the Stellent publication database, they are automatically converted to STF format. ISIS query results can also be stored in STF format.

- **Standard documentations:**

The standard documentations – which can be downloaded as PDF documents over the Web – serve as the most important source of metadata about statistical projects and the quality of the statistical results they produce. The documents exhibit a standardised chapter structure and hitherto describe more than 100 statistical projects or survey versions, in part in great detail (they number between 8 and 100 pages; in many cases further documents are provided as attachments which can be accessed via hyperlinks in the text). Among other things they do carry the disadvantage of usually being written and made available to the statistics' users in a separate and additional work step after the fact, although they contain many documentation elements which come into existence in the early phases of planning and preparing the statistical project. Another weak point is that there are no quantitative quality-indicators included.

This system was implemented through a Word template. Every manager of a statistical project is obliged to use this template when compiling a standard documentation.

The main headlines are the following:

1. Important Hints
2. General Information
3. Statistical Concepts, Methods
4. Production of Statistics, Processing, Quality Assurance
5. Publication (Accessibility)
6. Quality

Every chapter is divided into subsections which are more or less standardized.

- **Release calendar:**

The calendar of planned releases is available at http://www.statistik.at/web_de/ueber_uns/veroeffentlichungstermine/index.html. It consists of two PDF-files which are updated on a regular basis (in the first one releases are sorted by date, in the second by statistical domain).

From the same Web address, a file with information on the dates of data transmissions to Eurostat can be downloaded. There is also a link to the advance release calendar at the SDDS site of the IMF.

The planned press releases of the upcoming week are published at http://www.statistik.at/web_de/presse/presseservice/index.html

- **Database of administrative data:**

This is an MS-Access application available only to internal users which contains information about administrative data sources.

C. COSTS AND BENEFITS

Metadata systems form a fundamental information infrastructure for the production of statistics. More than 15 years ago, Bo Sundgren wrote the following about this topic:

"Statistical metainformation systems (...) exhibit some characteristics, which are typical for infrastructures:

- *They require collective commitment and relatively large investments, which (at least initially) have to be financed by the organization as a whole.*
- *They have to be designed on the basis of partially unknown needs, some of which require "intelligent guesses" about the future.*
- *They have to be planned for a wide range of usages and users, some of whom may have conflicting needs.*
- *Once they exist, the marginal cost of using them is relatively low, at least in comparison with the initial investment."*

(Bo Sundgren, *Organizing the Metainformation Systems of a Statistical Office*, Statistics Sweden 1992)

When metadata can be utilized to standardize and automate production processes ("active metadata"; see section 3.1), the costs for the development of metadata systems (which in many cases are quite substantial) are balanced by prospective long term monetary benefits, which in the long run may result in major cost savings. One example of this is Statistics Austria's metadata driven electronic questionnaire system e-Quest. Compared to the development of a tailor-made electronic questionnaire for a single survey, its initial development costs were inevitably higher. But now e-Quest facilitates the cost-effective creation of electronic questionnaires. By using the system repeatedly within many different statistical projects, the break-even-point was reached quickly.

The situation in the case of developing systems for the collection and administration of passive metadata is, however, quite different. Passive metadata are an integral component of statistical information. Their availability and easy accessibility contribute to the quality of statistical products, but in many cases do not result in cost reductions (they may even increase the work load of subject matter statisticians). Opportunity costs caused by the non-existence of centralized end-to-end metadata systems are rarely found in accounting systems. Thus high investments are accompanied "only" by a gradual gain in quality (which may not even be recognized by all user groups). Under these circumstances it is understandable that in times of economic crisis the willingness to invest in metadata projects is not high.

The concept of "high-quality statistics" is a dynamic one. The needs and requirements of users are changing and will probably increase in the future, e.g. with regard to harmonization of statistics or the linkage of data with *relevant* metadata items (respectively linkage of metadata items with related metadata items), so that they can be accessed at the push of a button. If metadata are stored in the continuous text of bulky documents, these new requirements cannot be met. The management of metadata in an "atomic" and structured form, however, is a challenge with respect to both financial resources and personnel.

The fundamental principles of metadata management, which have been defined by experts during recent years (and which can be found, for example, in part A of the Common Metadata Framework) will become more and more commonly accepted standards and state of the art for the production and dissemination of statistical information.

The task of implementing these standards can certainly not be carried out at short notice. In this respect, it is not easy to answer the question whether to continue building isolated metadata systems whenever the need for one specific system arises, or whether to strive for an integrated system based on a global architecture. The first approach is certainly less expensive in the short run and produces quicker results, but in the long term it will cause quite substantial "repair" costs.

D. IMPLEMENTATION STRATEGY

Similar to the BASIS 2000+ concept, a modular implementation approach was a major design principle of the IMS. In order to minimize the complexity of the complete system, the individual components (subsystems) should be able to work independently, communicating with each other and the central "Registry" by means of a web service and program interface layer. Thus – considering the limited resources – stepwise realization and gradual commissioning and expansion of the IMS (in the sense of "evolution instead of revolution") should be facilitated.

Regarding the integration of previously existing legacy systems into the IMS, several options are possible. A very simple form of coupling can be realized by manually registering information objects (for example a classification from the Classification Database) in the IMS Registry. A tighter and more sophisticated integration will require some programming effort, so that a legacy system can communicate with other components of the IMS via web services.

III. STATISTICAL METADATA IN EACH PHASE OF THE STATISTICAL BUSINESS PROCESS

A. METADATA CLASSIFICATION

Statistics Austria has no "official" classification of metadata. But during the conceptual work for BASIS 2000+, STAT+ and the integrated metadata system IMS a multidimensional approach – similar to Bo Sundgren's proposal in working paper 7 of the METIS 2008 meeting – was worked out.

In this model metadata content is considered to be the focal point. Statistical metadata appear in many different forms: e.g., as title of a table, text in a document (describing, for example, conceptual objects like a survey, a statistical concept or a validation rule), source code statements in a software program, technical attributes of a file, and so on. In principle these metadata items can be seen as instances of a set of object types, which are connected by different kinds of relations (which themselves are part of the metadata).

Metadata dimensions

These different types of metadata can be investigated from varying points of view. The IMS project team differentiated between the following six dimensions:

1. Dimension "Function":

This dimension describes metadata's purpose. Basically, metadata are required for the following reasons:

- for searching and finding statistical information
- in order to interpret statistical data
- in order to access data
- for processing data and producing statistics
- for managing statistical projects

2. Dimension "Statistical Life Cycle":

Statistics production can be described as a process which transforms input data into output data via several steps and using statistical methods. In Statistics Austria, this statistical life cycle is structured in the form of "statistical projects" of different types (surveys, registers, analytical projects or systems). This is described in more detail in section 2.1

3. Dimension "Users":

Statistical metadata are no end in themselves, but are required by different groups of users for varying purposes. "User" must here be understood in a broad sense, comprising not only persons but also IT systems.

We can distinguish roughly between two main groups of users: external ones (which do not belong to the NSI) and internal users. External users are mostly "consumers" of statistics, they may however also be providers of raw data (respondents). Internal users are often both "producers" as well as consumers. Among others, external users may be politicians, scientists, economic enterprises, journalists, private persons or international organisations. In general, the users in these groups differ in the amount of their previous knowledge, the level of detail they wish for in the statistical information they are seeking, and the resources at their command. From the point of view of the amount of metadata they require, one must keep in mind that this may also vary within the relatively heterogeneous groups. Furthermore, the requirements may evolve with time.

4. Dimension "active / passive":

This dimension treats the degree to which metadata play an active role in statistics production, i.e. controlling the process or automating processing steps (e.g. when an electronic questionnaire is generated automatically based on the specification of a survey's questions). With regard to efficient production of statistics one should aim at letting as many active metadata elements as possible be defined directly by the statistical subject matter experts.

5. Dimension "formatted / unformatted":

A distinction can be drawn between formatted and unformatted data. The structure of the former is agreed beforehand (e.g., every record in a file consists of the same sequence of data fields, which in their turn exhibit prearranged characteristics such as data type, length, etc.; or a data file conforms to a predefined XML schema) and thus easily lends itself to automated processing with computers. Unformatted data on the other hand – texts, graphics, voice etc. – are much more difficult and cost more effort to process, especially with regard to IT programs "understanding" their contents. Statistical metadata often occur in unformatted form, e.g. as text in documents.

6. Dimension "manual / automatic":

The criterion by which this dimension classifies metadata is whether they are recorded manually by the persons entrusted with planning and implementing statistical projects, or whether they are created automatically by tools.

Additional Aspects

Apart from these dimensions, which serve as a means to describe and understand the multilayered topic "statistical metadata", other important aspects must be taken into consideration within the context of metadata management and the development of metadata systems.

1. Quality:

When talking about quality in statistics, in most cases the quality of data and statistical results is regarded. In this context, a definition of quality as well as quality criteria have been elaborated, and many NSIs have introduced routines for quality reporting within their institutions.

Compared to data quality, the topic of “metadata quality” has received much less attention. In our opinion, the definition of quality criteria for metadata should become a central task of international working groups in the future.

2. Organization:

This topic comprises organizational questions within an NSI (for instance: is there a central metadata unit? If yes: what are its tasks?), but also issues regarding the registration and administration of metadata items (for example access rights, stewardship, life-cycle status, locking of items while they are updated).

3. Tools:

In the process of software development metadata play a decisive role. In order to produce software of high quality and in an economic way, the availability of tools – to support the management of “software metadata” (including the source code of the programs) and to provide services to alleviate the software engineers’ work – has long been recognized as necessary. Especially when several programmers are cooperating in a software project, the storage and administration of all information items in a central repository seems indispensable.

The production of statistics exhibits a high degree of similarity to the production of software. However, in statistics the advantages offered by specialized tools and a centralized metadata repository are not yet generally accepted.

4. Architecture/Strategy:

Numerous papers point out that the development of a long term strategy forms a necessary and fundamental basis for the step-by-step realization of an integrated metadata system. The elaboration of a “construction plan” as a flexible and extendable architecture is cost-intensive and time consuming, but it is also an investment into a stable fundament which will pay off in the future.

Some important general goals of a metadata strategy are:

- centralization of metadata;
- identification of “atomic“ metadata items, their structure and their mutual relations, and decomposition of so far unstructured metadata into such elements;
- storage of these atomic and structured metadata items in databases;
- integration of isolated subsystems into a complete system;
- end-to-end processing.

B. METADATA USED/CREATED AT EACH PHASE

In the “4-layer model” metadata are represented as an “infrastructure“ layer accompanying the phases of statistical production; in every phase newly produced metadata are stored in the metadata systems and existing metadata are accessed and perhaps re-used. A higher degree of model detail concerning different types of metadata was not attempted, however.

C. METADATA RELEVANT TO OTHER BUSINESS PROCESSES

For the purpose of cost planning and controlling, SAP software is used.

IV. SYSTEMS AND DESIGN ISSUES

A. IT ARCHITECTURE

One of the cornerstones of Statistics Austria's IT strategy is the use of the database DB2 on the mainframe computer. Web applications run on Linux servers (which are logical partitions of the mainframe). Server programs are developed in Java and deployed with IBM Websphere as application server. To a large extent, PC software is also written in Java.

However, Statistics Austria still has a relatively heterogeneous, historically grown IT landscape which is composed of a large number of legacy systems and components. As a consequence one can find a multitude of programming languages, development environments and IT architectures. A few examples:

ISIS: The statistical output database is a mainframe application which was developed completely by Statistics Austria. PL/I and assembler were used as programming languages.

ISISGui: At the beginning of the new millennium a graphical user interface for ISIS was implemented as Java applet.

e-Quest: The electronic questionnaire software which is installed on the PCs of the respondents is a Visual-Basic 6.0 application and stores data and metadata in the Microsoft relational database MSDE (at Statistics Austria DB2 is used instead). A number of supplementary programs were written in VB6 or Delphi.

Publication Database: The core of the Publication Database is a commercial document management software (Stellent). Extensions and adaptations developed especially for Statistics Austria are written in Java and a proprietary Stellent script language.

IMS: If the decision is made to realize the integrated metadata system IMS, the implementation will be carried out in accordance with the IT-strategy (storage of metadata in DB2 on the mainframe, business-logic-layer as Web application under IBM Websphere on Linux, client software written in Java).

B. METADATA MANAGEMENT TOOLS

Due to the existence of several isolated metadata systems, there also exist a number of different metadata management tools. Some examples:

Standard Documentations: The standard documentations are written in Word using a predefined template, stored in the Publication Database, converted to PDF and finally replicated on the Web server.

ISIS: Management tools for ISIS are mainframe applications written in PL/I or assembler.

e-Quest: Statisticians define electronic questionnaires using a graphical editor ("e-Quest Metadata Manager") which was developed in Visual Basic 6.0.

Classification Database: Administration software for interactive editing and processing of classifications does not yet exist.

IMS: The development of a management tool which can easily be extended by plug-ins, based on the Eclipse RCP (Rich Client Platform) architecture, is planned.

C. STANDARDS AND FORMATS

Classification Database: The Classification Database was based on the Neuchatel model; however, the system is not a complete implementation of that model.

IMS: Two of the proposed IMS subsystems ("Variables" and "Value Domains") are based on ISO 11179.

D. VERSION CONTROL AND REVISIONS

Version control of metadata is a complex topic. Within the framework of the IMS it is planned to specify version objects whenever version control is found necessary during analysis (for example, a survey would be composed of one or more survey versions). This approach was also used in e-Quest and proved successful.

Regarding the IMS subsystem “Concepts and Definitions“, it was decided that identifying versions of definitions is not necessary because handling them would be too difficult. It is the opinion of subject matter experts that most of the definitions can remain static. If there are changes in some definitions, these will be described in the corresponding text of the definition.

Concerning revisions, it should be a feature of the metadata repository software that prior versions of an item are not deleted, but remain available.

In the Classification Database versions of classifications can be accessed.

E. OUTSOURCING VERSUS IN-HOUSE DEVELOPMENT

Due to limited human resources the outsourcing of software development often is inevitable. However, project management is still carried out by employees of Statistics Austria.

The Classification Database was developed internally (although a prototype was written by an external company).

F. SHARING SOFTWARE COMPONENTS OF TOOLS

- - -

G. ADDITIONAL MATERIALS

- - -

V. ORGANIZATIONAL AND WORKPLACE CULTURE ISSUES

A. OVERVIEW OF ROLES AND RESPONSIBILITIES

Statistical projects are planned and compiled in four subject matter divisions. The survey manager is responsible to compile a standard-documentation. When electronic questionnaires are used for raw data collection, the corresponding e-Quest metadata are defined by the subject matter divisions as well.

Statistics Austria does not have a unit which is responsible for centralised metadata systems. The responsibilities for specific types of metadata are distributed among various organizational units.

An important role is played by the “Registers, Classifications and Methods“ department (approximately 70 persons). One of its sub-units, the GIS unit (Geographical Information Systems), is concerned with metadata because of the EU directive “INSPIRE“.

The administration of the statistical database ISIS is done by a unit attached to the Directorate-General.

Regarding the topic “quality of statistics”, a “Total Quality Management (TQM) board” was established in 2000, consisting of five members. In 2003, a quality management unit (consisting of two persons) was installed, which is directly subordinate to the Directorate-General. The tasks of this unit are to initiate, to plan, to co-ordinate, to implement and to control TQM measures in consultation with the Directors General and in close co-operation with the TQM Board.

It is the particular duty of the quality management unit – in co-operation with the TQM Board – to define actual quality projects within all TQM areas, and it is responsible for the co-ordination, the planning and management of projects (for example, the implementation of an in-house system for quality reporting and of so called "feedback talks", staff opinion surveys, etc.).

With respect to metadata, the IT department plays a very important role. As was mentioned above, the initiative for many projects related to metadata originated in this department. Members of the IT department have a leading function in the conceptual planning and realization of metadata systems (for instance, e-Quest, ISIS and ISIS New, Publication Database, conceptual design of the IMS architecture).

B. METADATA MANAGEMENT TEAM

See section 5.1

C. TRAINING AND KNOWLEDGE MANAGEMENT

The training programme of Statistics Austria includes two courses related to metadata:

- Course on the definition of electronic questionnaires with the "e-Quest Metadata Manager" software.
- Course on quality reporting.

As part of TQM, once a month the standard documentation of a certain statistical project is discussed by a group of experts and the quality subgroup of the Statistical Council. After the discussion the concerned standard documentation is reworked according to the proposals resulting from the discussion.

D. PARTNERSHIPS AND COOPERATION

Some metadata projects were conducted in cooperation with the well-known expert for metadata Prof. Fröschl (University of Vienna).

Members of Statistics Austria participate in international conferences and work sessions (METIS, Metadata Working Group).

E. OTHER ISSUES

- - -

VI. LESSONS LEARNED

- It is not a new discovery that the subject of statistical metadata is an extremely complex one. Even now, almost three and a half decades after Bo Sundgren first used the term, different individuals may still mean quite different things or place emphasis on different aspects when speaking of metadata. This phenomenon is even more pronounced when these persons stem from different areas of expertise: senior management, subject matter statisticians, methods specialists, IT experts etc.

- Papers prepared for the METIS work sessions and the Common Metadata Framework (especially the case studies) have proven very useful, as they provide arguments for discussions with statisticians and top management.

- The creation of an integrated system consisting of more than isolated solutions is difficult when there is no organizational unit the main responsibility of which is to deal with the subject of metadata and its usefulness for the NSI – and which is also granted the requisite authority and enjoys the support of top management, so that it can achieve the introduction of integrated and centralized metadata systems even against the possible resistance of subject matter departments.

- That metadata projects are best carried out using an interdisciplinary approach (and not as IT projects) has long been recognized in expert circles. In practice, however, it appears that the qualified subject matter statisticians continually suffer from such a heavy workload that they have no time to spare for complicated conceptual work (e.g., Statistics Austria has reduced the number of personnel by about a third since its separation from the federal civil service in the year 2000).

- Many statisticians associate the concept of "metadata" with the notion of "additional work" (which for instance actually was the case when the standard documentations were introduced). This leads them to resist new metadata systems.

- The idea of developing specialized tools for editing, administrating and (re-)using metadata with an end-to-end approach regarding the statistical life cycle often encounters resistance among statisticians because the introduction of such tools will result in changes to work processes which they have been familiar with for many years.

- Statistics, however, is not the only field of activity in which the creation and usage of metadata can be seen as part of the job description. In order to produce software of high quality and in an economic way, the availability of tools – to support the management of "software metadata" (including the source code of the programs) and to provide services to alleviate the software engineers' work – has long been recognized as necessary. Especially when several programmers are cooperating in a software project, the storage and administration of all information items in a central repository seems indispensable.

- The production of statistics exhibits a high degree of similarity to the production of software. However, in statistics the advantages offered by specialized tools and a centralized metadata repository are not yet generally accepted.

- As was already said in section 2.3, with regard to the development of systems for the collection and administration of passive metadata, the cost factor presents a particular obstacle. Passive metadata are an integral component of statistical information. Their availability and easy accessibility contribute to the quality of statistical products, but in many cases do not result in cost reductions (they may even increase the work load of subject matter statisticians). Opportunity costs caused by the non-existence of centralized end-to-end metadata systems are rarely found in accounting systems. Thus high investments are accompanied "only" by a gradual gain in quality (which may not even be recognized by all user groups). Under these circumstances it is understandable that in times of economic crisis the willingness to invest in metadata projects is not high.

- The concept of "high-quality statistics" is a dynamic one. The needs and requirements of users are changing and will probably increase in the future, e.g. with regard to harmonization of statistics or the linkage of data with *relevant* metadata items (respectively linkage of metadata items with related metadata items), so that they can be accessed at the push of a button. If metadata are stored in the continuous text of bulky documents, these new requirements cannot be met. The management of metadata in an "atomic" and structured form, however, is a challenge with respect to both financial resources and personnel.

- The fundamental principles of metadata management, which have been defined by experts during recent years (and which can be found, for example, in part A of the Common Metadata Framework) will become more and more commonly accepted standards and state of the art for the production and dissemination of statistical information.

- The task of implementing these standards can certainly not be carried out at short notice. In this respect, it is not easy to answer the question whether to continue building isolated metadata systems whenever the need for one specific system arises, or whether to strive for an integrated system based on a global architecture. The first approach is certainly less expensive in the short run and produces quicker results, but in the long term it will cause quite substantial "repair" costs.

What metadata should actually be collected for and provided to external and internal users, and in what form should they be provided? This is a fundamental question on which opinions within Statistics Austria are divided. The search for an answer should not be postponed just because it is clear from the start that up-to-date solutions will require high investments in time and money. The answer should rather be given as soon as possible in order to ensure from the start that the solutions – which must be planned and implemented step-by-step in accordance with budgetary constraints and on a long-term time scale – will be built to last.

VII. ATTACHMENTS & LINKS

"IMS (Integrated Metadata System) – An Architecture for an Expandable Metadata Repository to Support the Statistical Life Cycle" (working paper 5 of the 2007 METIS workshop): http://www.unece.org/stats/documents/ece/ces/ge.40/2007/wp.5.e.pdf

"e-Quest: A Metadata-Based System for Electronic Raw Data Collection" (working paper 15 of the 2002 work session on electronic data reporting): http://www.unece.org/stats/documents/2002/02/edr/15.e.pdf

"Data Warehouse in a Statistical Office" (working paper 9 of the seminar on statistical information systems ISIS 2000): http://www.unece.org/stats/documents/ces/sem.43/9.e.pdf