4. Statistical Metadata Systems (Statistics Austria)

4.1 Metadata 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.

4.2 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.

4.3 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.