Metainformation System of the Hungarian Central Statistical Office

Gizella Baracza
Chief Professional Advisor
HCSO
E-mail: Baracza.Lajosne@ksh.hu

Zsófia Ercsey
Head of Section
HCSO
E-mail: Ercsey.Zsofia@ksh.hu

Csaba Ábry
Councillor
HCSO
E-mail: Abry.Csaba@ksh.hu

The article gives detailed information on the metainformation system of the HCSO. It presents the history from the beginning up to the latest improvements. A separate chapter of the paper provides an international overview.

KEYWORDS:
Statistical methodology.
Development program.
Database.

The first chapter gives information on the definition of metadata. The second chapter covers the history of the metainformation system of the Hungarian Central Statistical Office (HCSO) and the various stages of the system improvement from the beginning until today. The third chapter presents the main subsystems of the metadatabase, while the fourth chapter provides an international overview. The fifth chapter includes the results of the latest development available on the website of HCSO.

1. The role of metadata in statistics

Metadata are data on data. They adequately describe, determine data. Therefore, wherever we are engaged in data handling, we have to produce metadata as well. This is especially the case for compiling statistics.

If we say that the value of industrial production decreased by 3.2 percent, we can find two metadata beside the statistical one: the denomination for describing the content (change of industrial production) and the unit of measure (percent).

However, these are not yet satisfactory for the right explanation of data. In order to make data valuable for users, we have to add more metadata. For example the following:

– the time period of data (for example May of xxxx. year);
– the reference time period to which we compared the examined period (for example the same time period of the previous year);
– the definition of "Production value" as a concept;
– the definition of "Industry" or those classes of NACE Rev.2., which belong to the "Industry" branch;
– the scope of observation (statistical units) (for example economic units employing more than 5 persons);
– the type of the survey (full scope or based on a representative sample);
– in case of applying a representative sample, we have to add the method of estimation;
– the method of imputation for non-response;
– the calculation method of indices;
– the application of seasonal adjustment;
– the punctuality of data.

There are several more metadata, which we could add to the previous list, these were just some examples. The more metadata we add to a statistical data, the more certain we become that the user does not misunderstand the statistical data and he/she analyses it well.

Statistical data without adequate metadata are barely understood. The scope of metadata added to data depends on the type of users or on the aim of data usage. A person of the general public is not interested in the method of calculation if he/she wants to know for example the rate of inflation. But economists, researchers or analysers highly need metadata on the methodology used and the quality aspects of statistical data.

Metadata can fulfil several requirements. Among these, the most important one is to give users information on the content and quality of data or on the method of data production. Hence, there are metadata, which support, document the work of persons engaged in data processing.

The automation and integration of statistical data production requires more and more parameters for operation. These are also metadata, and in such cases we can talk about metadata-driven processes.

If we describe data and processes in the same way in every statistical domain, the whole statistical system becomes more uniform and integrated. Standardization is brought to the fore also at the international level, especially in case of data transmission. The need for including standardised metadata in data transmissions arose long time ago in order to understand data better. As a solution, the statistical data and metadata exchange (SDMX) standard was established with the sponsorship of seven international organisations.

The knowledge of planning and handling metadata should be a part of statistical knowledge. Information that is necessary for the practice of statistics as a science and as a profession can be systematized as follows (*Pukli–Végvári* [2004]):

– information on general statistics;
– special information of statistical domains;
– information on planning, maintaining, improving the statistical information system.

The metadata system is a sub-system of the statistical information system. It covers the databases of metadata, as well as the activities and IT tools necessary for handling them.

The international statistical organisations have recognised the importance of metadata, thus their work programmes always contain relevant issues. A major milestone was the establishment of the METIS Group as an international research forum of the subject. It was founded by the UN Statistics Division and the Economic

Commission for Europe. The forum made it possible to exchange experiences associated with statistical work. The main aim of the research was to work out guidelines on models concerning statistical data and metadata (*UN* [1995]).

At the European level, the most important forum for metadata is the META Working Group, which coordinates the work of national statistical institutes in the subject of the metainformation system.

2. History of the meta-information system improvement at the HCSO

The Hungarian Central Statistical Office recognised the importance of metadata in the 1970s. Although the available resources and the intensity of improvement changed, this issue is constantly on its agenda.

The direction and degree of development depended on several issues. Among them, the most important ones are:

– support from the top management;
– condition of the IT background;
– importance of different sub-processes within statistical data production.

The history of the metainformation system improvement of the HCSO can be divided into the stages to be stated next.

2.1. "Heroic times", establishing the bases

In the 1970s, special focus was placed on researching the integration of information systems in the working groups of international organisations and in the HCSO as well. The centre of research was the Computing Research Centre in Bratislava, which was supported by the UN. The research objective was to plan and to introduce the integrated statistical information system (ISIS), and the annual seminar on ISIS served as a forum for analysing the results. Although metadata-related researches were led by Swedish experts, Polish, Czechoslovakian and Hungarian specialists also played a significant role in this field.

The top management of the HCSO established a separate organisational unit for the improvement of the statistical information system. At that time the current database management systems did not exist. In 1974, the HCSO procured the MARK IV

file management system, which could describe data irrespective of the programme and thus, it served as a base for the development of the database. This system was called STAR (Statistical Database System).

Thereupon the production database structure was finalised, to which metadata describing the statistical data content for users had to be connected. Metadata were also stored in MARK IV files. For the realisation of the system, an IBM mainframe and a batch process were available.

The types of metadata connected to the database were the following: *1*. Hierarchy of statistical domains and files; *2*. Measures (including observed and aggregated variables) with their unit of measurement, periodicity, reference period and information on comparability; *3*. Nomenclatures (classifications); *4*. Nomenclature items; *5*. Variety of nomenclatures, which is the subset of elements; *6*. Nomenclatures determining the level of aggregation; *7*. Cross-references between nomenclature items; *8*. Statistical concepts.

Concerning the content, the foundation of the present metadatabase was laid down at that time. While the most important metadata types were already available, upload caused a problem. In order to integrate other domains or fields into the system, their structure had to be planned. This caused a restructuring of the system, requiring great effort from statisticians and IT experts. Therefore, the expansion of the system was very slow.

The connection between the database and metadata was ensured by the naming convention for identifiers.

As the technical conditions did not allow the on-line availability of data, the description of metadata was disseminated through white papers (MARK IV had an excellent text processing solution). The following catalogues were compiled:

– Measure catalogue: variables stored in the database and aggregated measures;

– Catalogue of nomenclatures and classifications: list of the items of nomenclatures, classifications and nomenclature varieties;

– Dissemination catalogue: denomination of measures and classifications associated with keywords;

– List of the definitions of statistical concepts: definitions of the most important concepts.

As this catalogue system was in close connection with STAR databases, it was known as “STAR catalogues” among users. Later, it was supplemented with further catalogues, such as the catalogue of data collections and publications and a catalogue of statistical questionnaires with cell identifiers. The whole system composed the so-called “Data documentation system” (*Baracza* [1980]).

2.2. SOLAR, the first metadata-driven system in the HCSO

In the beginning of the 1980s, a need arose for developing interactive accessibility of the data of the database for users so that they could make aggregation on data after choosing the right metadata. This demand was met by the statistical online data query system (SOLAR) (*Györki–Papp* [1985]). By this time the IBM mainframe became available from terminals.

To realise the system, new metadata were required. The most important one among them was the group of those measures that could be processed together. A cardinal item of the system was the data vocabulary, which contained the description of the system for users and for the software as well (*Györki* [1980]).

The software was metadata driven with respect to dissemination and operation with data, furthermore, there was an opportunity for automation in case of data upload.

The system was finalised after a long in-house development phase; the users started to use it and gave positive feedback on its functionality. Although it didn't become widespread, a great deal of experience was gained on metadata needed to develop a metadata-driven system and on the functionality required from such a complex system.

2.3. The introduction of the database management system and the spread of PC clients

In the beginning of the 1990s, there was an opportunity for the HCSO to renew its IT system with HP Unix operation system, ORACLE database management system and PC clients. The main task of this time period was migration of data and programme systems. As the MARK IV files and the files containing metadata were relational, they could be easily migrated to the ORACLE database but for queries and maintenance new applications had to be developed.

Instead of migrating SOLAR, the HCSO wanted to choose another system, which covered more existing software items. As the development of this system was carried out with third generation tools, it was difficult to follow the software/hardware conditions and there was not enough capacity for the development.

2.4. Methodological documentations

The terminology of several statistical domains was compiled in the 1970s and 1980s and was disseminated on white papers. This serial also included the folders of

the most important statistical classifications. In a few domains, summary leaflets on statistical methods were published too. The denominations of the serials were:

– Statistical Concepts;
– Statistical Nomenclatures;
– Statistical Methodological Papers.

In 1995, the top management of the HCSO established a Methodological Working Group to elaborate the so-called basic documentation of statistical domains. As a result, the methodological "assets" of the HCSO was surveyed, and a HCSO presidential order entered into force on the planning of statistical data collections in 1997. This work has laid down the foundation for the compilation of the methodological documentation of the subsequent statistical domains.

2.5. Methodological description in publications

Methodological descriptions have been provided to the tables in the HCSO publications since the 1980s, and on the HCSO website, data have been published in STADAT system since the middle of 1990s. Data are available for users in ready-made tables supplemented with methodological descriptions, which contain the most important concepts of the given statistical domain.

2.6. The support of the statistical data production with metadata

This subchapter provides information on the subsystems of the metainformation system concerning statistical data production.

Data collections, data providers

The recording system of *data collections* was developed within the framework of the metainformation system. The data collections of the statistical service are included in the National Statistical Data Collecting Programme for the compilation of which the HCSO is responsible. (Besides, it also contains data collections enacted by other institutions.)

The records of economic units as *data providers* are covered by the Business Register. Their attributes and the corresponding attribute values are described by the sub-system "Nomenclatures" of the metasystem.

Cross-references between data collections and data providers are recorded by the system of survey control (GÉSA). This controls the dispatch of questionnaires, re-

cords data capture, encodes the reasons for late responses or non-responses. In order to operate GÉSA, new metadata were needed for the accurate description of questionnaires, instructions and supplements and for the determination of the scope of data providers, etc. (*Györki* [1996]).

All three mentioned types of objects partly existed even in MARK IV. The usage of more and more metadata made the systems more complex during transition to ORACLE.

Data preparation

At the end of the 1990s, there was a need for uniform handling of data preparation (data entry and editing). In order to manage this problem, a system called ADÉL was established. The related metadata provides logical and physical description of questionnaires, their correspondence and the definitions of check rules.

Electronic data collection

Since 2005, more and more data collections have been using the electronic data collection system of the HCSO, in which registered users get their own list of questionnaires to be completed. During the fill-in process, the system makes an automated validation, and thus, the supplied data are automatically loaded into the ADÉL system. The database loading and impersonating part of the system is metadata driven. From 2008 onwards, the completed questionnaires received by e-mail are also loaded into the database on the basis of the questionnaire cells–measures, nomenclatures–database cells relation.

Data processing

The standardisation of data processing and the establishment of the metadata-driven data processing system (Uniform Data Processing System) are still in planning phase.

2.7. Data warehouse, dissemination database

A multi-dimensional database management system was procured by the HCSO (ORACLE Express and later Hyperion) at the end of the 1990s. Although external experts were needed for its realisation, the principles and plans to make the system metadata driven were compiled by HCSO experts.

The system covers two main parts: *1*. A data warehouse for the internal users of the HCSO; *2*. A dissemination database for external users.

The data warehouse stores the prepared and documented statistical data and makes them available for interactive use. According to their needs, users may compile tables, graphs, maps in a dynamic way.

The dissemination database is such a part of the data warehouse that does not cover confidential data and is available via the internet.

For the following purposes, the data warehouse needed more metadata besides the ones it had used: *1*. the determination of data request is based on metadata (measures, dimensions, etc.); *2*. there are metadata for the right interpretation of data (definitions of measures and concepts; from 2009 the complete methodological documentation of the statistical domains); *3*. the maintenance of data is metadata driven.

The dissemination database is available on the HCSO website both in English and Hungarian from the middle of 2004.

2.8. Completion of subsystem "Concepts"

The need for completing the statistical concepts has a long history in the HCSO. The IT background had got ready for the maintenance of concepts but the enormous job of uploading the approved, bilingual notions of all statistical domains was only carried out in the time period between 2005 and 2008.

The subsystem "Concepts" is complete by now. It includes approximately two thousand concepts, which have definitions, cross-references, sources, etc. For further improvement of this subsystem, "supervision" is performed, that is to say, consistency and cross-references between the concepts of different statistical domains are checked by methodological supervisors. The supervision is carried out within a certain statistical domain and also between different statistical domains.

2.9. Completion of subsystem "Nomenclatures, classifications"

In the metadatabase of the HCSO, there are hundreds of nomenclatures both for the data collection phase and for data processing, information. Within the framework of the 2007–2009 development, different classifications were selected for inclusion on the website. They were mainly international classifications (for example NACE, ISCO, COICOP, COFOG, etc.). All the prioritised classifications were loaded into the metadatabase, if it has not already been done. For giving adequate information to users, a short description on these classifications (on their content, legal base, structure, history, applications, etc.) was prepared. Currently, there are twenty-four classifications on the HCSO website.

2.10. The methodological documentation of statistical domains on the website

The 2005 HCSO strategy contained the task of supplementing the metainformation system with a new subsystem, which sets down the methodological background of statistical domains. Thus, supplementing the existing system with new metadata, it was established as an integrated part of the metainformation system. It is available from September 2008 both in Hungarian and English on the website of the HCSO (www.ksh.hu/Data/Metainformation). The new sub-system informs users of the content, legal base, concepts, classifications, data sources, methods, dissemination forms, etc. of a certain statistical domain.

After studying the websites of other statistical institutions and international standards, the methodological documentation schema for every statistical domain was finalised as follows:

– Short description;
– Concepts;
– Classifications;
– Methodology of data production;
– Data quality;
– Data sources.

The detailed schema can be found in Annex 1 (www.ksh.hu/statszemle). During the development phase, methodological documents were compiled for approximately hundred statistical domains. This was followed by a revision procedure covering the following stages: *1*. Methodological reading, which was carried out by erudite experts of the HCSO according to general methodological and quality rules. *2*. External user reading, that is, revision of metadata on the website, from a common, external user's point of view. This phase was performed by the experts of the concerned ministries, professional organisations and universities. *3*. Linguistic reading done by HCSO experts specialised for this work. They ensured the clarity, grammatical correctness and uniform use of words.

HCSO publications always included chapters for methodology, but they varied from one statistical domain to another. Therefore, the new methodological documentation has the following advantages:

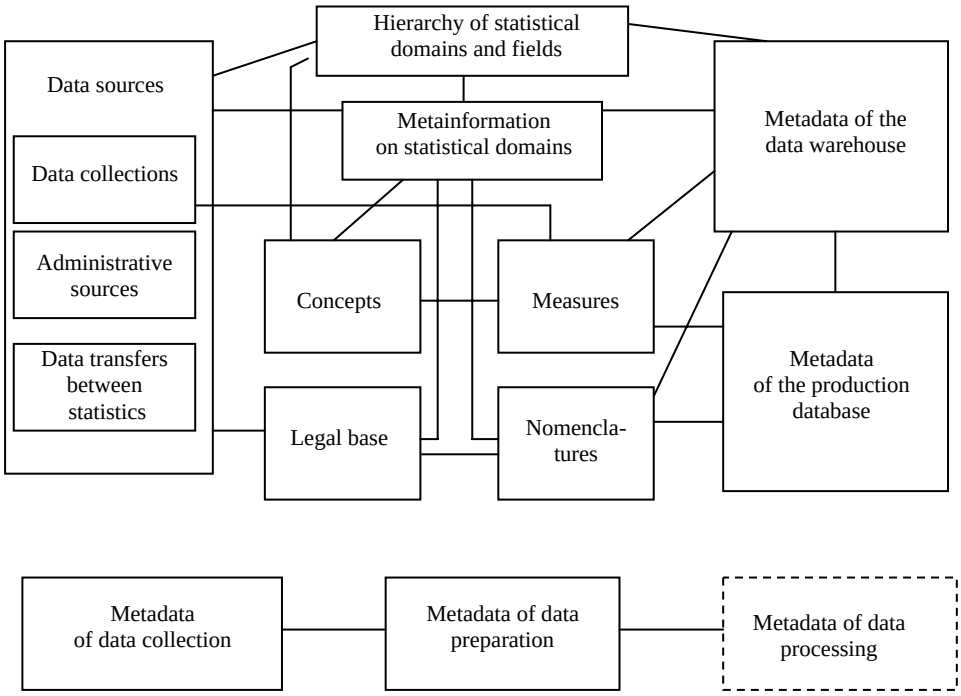
– methodological descriptions became uniform;
– there are more available metadata (new chapters embrace short descriptions, methodology, quality);
– these documentations are more detailed and cover the changes through time;

– the documentations are available for everyone on the website of the HCSO.

3. Main subsystems of the metainformation system

The Hungarian metainformation system is integrated which is well illustrated by Figure 1. In this chapter its main subsystems are presented.

Figure1. Main subsystems of the metainformation system



Note. The box with broken lines refers to further improvement plans.

3.1. The hierarchy of statistical domains and fields

The "*Hierarchy of statistical domains and fields*" is the systemic grouping of statistics by content. Its structure is composed of three levels. The first level includes 6, the second 32 and the third 95 domains and fields.

The hierarchy puts statistical domains into a system, for example:

– Social statistics
  – Employment, labour force, earning
  – Social stratification, living conditions
  – Households, families
  – Housing and public utilities
  – Health care, etc.

An example for the general fields:

– Methodology, metainformation system
  – Statistical production process and methodology
    – Survey design
    – Data collection arrangement, etc.

The "Hierarchy of statistical domains and fields" is the principal rule for the system. Both statistical and metadata (concepts, statistical domains, data sources, etc.) can be searched with the help of this structure. (See Annex 2. www.ksh.hu/statszemle)

3.2. Subsystem "Concepts"

As it was mentioned previously, the subsystem "*Concepts*" contains definitions, their validity periods and sources as well as cross-references between concepts. For example:

Mother tongue

> Definition: Mother tongue is the living language which one learns in childhood (as first language), in which he/she generally speaks with the family members and which one declares to be his/her mother tongue, free of any influence and true to reality. The mother tongue of dumb and infants unable to speak is the language in which their closest relatives regularly speak.
> Source of definition: HCSO
> Related term: Language spoken besides one's mother tongue
> Related term: Nationality

3.3. Subsystem "Nomenclatures, classifications"

The subsystem "*Nomenclatures, classifications*" contains the identifier, the denomination and, if it is necessary, the definition of classification items.

For example:

COICOP – Classification of Individual Consumption by Purpose

01 Food and non-alcoholic beverages
011 Food
0111 Bread and cereals (ND)
0112 Meat (ND)
0113 Fish and seafood (ND)
0114 Milk, cheese and eggs (ND)
0115 Oils and fats (ND), etc.

Subsets can be selected from nomenclature items which are called “Varieties of nomenclature”. There is also an opportunity to make correspondence tables between nomenclatures.

For example: a part of the correspondence table between NACE Rev.1. and NACE Rev 2.

NACE Rev.2:	NACE Rev 1.1:
0111 Growing of cereals (except rice), leguminous crops and oil seeds	0111 Growing of cereals and other crops n.e.c.*
0112 Growing of rice	0111 Growing of cereals and other crops n.e.c.
0113 Growing of vegetables and melons, roots and tubers	0111 Growing of cereals and other crops n.e.c.
	0112 Growing of vegetables, ornamental plants
0114 Growing of sugar cane	0111 Growing of cereals and other crops n.e.c.
0115 Growing of tobacco	0111 Growing of cereals and other crops n.e.c.

* n.e.c. stands for not elsewhere classified.

3.4. Subsystem “Measures”

The subsystem “*Measures*” cover variables, attributes used in data collections and aggregated measures (for example: amount of total harvested production (ton), size of harvested arable land (hectare), average yield of arable crops (kg/ha), etc.)

An important piece of information on a measure is the level of aggregation, which is described by the nomenclature (for example territorial units, group of arable crops, legal forms of enterprises, etc.).

The category "Variety of measure" is used for expressing the different aggregation levels of a certain measure. But further definitions and explanations can be also added.

Measures have a central role in the metainformation system. As the system stores the cross-references of measures, the following information can be found out:

– What statistical concepts do we need to know in order to understand a measure?
– Which data collection contains that certain measure?
– Which data cube in the data warehouse contains the measure? etc.

3.5. Subsystem "Legal base"

The subsystem "*Legal base*" is a record for legal rules, handbooks, recommendations, etc. on which statistical work is based. The enactor can be either a Hungarian or an international organisation. The register contains the denomination and type (act, regulation, etc.) of the legal base and (if available) the accessibility.

For example: Commission Regulation (EC) No 912/2004 of 29 April 2004 implementing Council Regulation (EEC) No 3924/91 on the establishment of a Community survey of industrial production (Text with EEA relevance)

In the metainformation system, the legal base can be attached to a statistical domain, a data source or a classification.

3.6. Subsystem "Data sources"

The subsystem "*Data sources*" covers the main sources of statistics: data collections and administrative data sources as well as data transfers between statistical domains.

An example for metadata on data collection is as follows:

Consumer price survey, 2008

Enactor: Hungarian Central Statistical Office
Legal status: Compulsory data collection
Change of data collection: Unmodified data collection

Frequency of data collection: Monthly
Scope of data providers: Designated shops, repair shops, marketplaces, etc.
Deadline of arrival: 27th–29th of the reference month
Mode of data collection: Representative data collection
Mode of implementing data collection: By enumerators

The next example is for metadata on administrative data sources:

Address register of the Central Office for Administrative and Electronic Public Services (data for the international migration statistics), 2008

Organisation supplying the data: Central Office for Administrative and Electronic Public Services
Frequency of data receipt: Twice a year
Content of data receipt: Data of foreigners having address in Hungary, data of emigrating and returning Hungarians: name, mother's name, sex, date of birth, place of birth, citizenship, date of changing citizenship, marital status, legal title of registration, reason and date of getting into the register

Besides this, the records of administrative data sources cover data on the availability of experts for the given administrative data.

Since, instead of external sources, other statistical domains are the major source of statistics, it is necessary to record data transfers between them. The description contains the measures, aggregation level and periodicity of data transfer. For example:

The final expenditure of GDP
Data transfer from Dwelling construction and cessation
Serial No.: 1
Denomination of transferred data: Size and number of flats
Content of transferred data: Number of flats, dwelling construction by constructor groups
Frequency of data transfer: Quarterly

3.7. Subsystem "Metainformation on statistical domains"

The subsystem "*Metainformation on statistical domains*" is the newest development in the metainformation system. It covers textual information on statistical do-

mains on the one hand and cross-references between the metainformation subsystems on the other. The text description comprises of three main parts:

– Short description;
– Data production methods;
– Data quality.

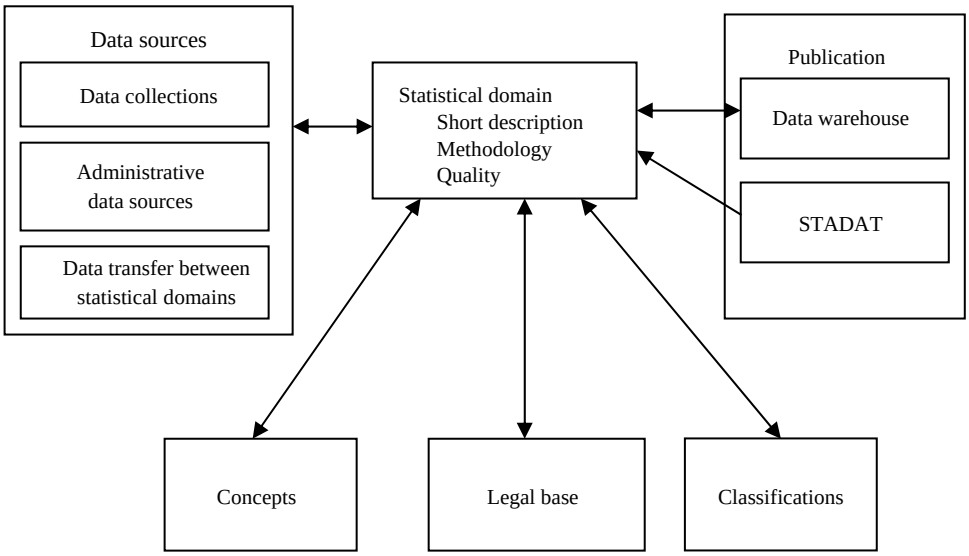
The *short description* covers the following: responsible department; responsible person; purpose and content of the given statistical domain; classifications used; data sources; forms of dissemination; timeliness, revision policy and practice; history of the statistical domain.

The "*data production methods*" part includes the following headings: population, sampling frame; sampling; data capture and data editing methods; imputation; data processing, estimation, calculation and compilation; methodological publication(s).

Main issues of *data quality* are: relevance; accuracy; comparability, coherence (comparability between geographical areas (with other countries, within the country); comparability over time; coherence with other statistics); quality report.

The subsystem "Metainformation on statistical domains" connects the other subsystems of the metasystem to the statistical domains as it is shown in Figure 2.

Figure 2. Connection between statistical domains and other subsystems



These connections make it possible to search easily the concepts, classifications, legal bases and data sources of a given statistical domain.

There is a link from publications to the statistical domains. The button "Methodology" navigates the user to the subsystem "Metainformation on statistical domains".

3.8. Subsystem "Data warehouse"

As the most important part of subsystem "*Data warehouse*", the data cube covers data which can be processed together. Besides denomination, there is information on the reference period and reference observation units in its description.

The dimensions of the data cube are described by the nomenclatures of the metasystem. One dimension may contain several nomenclatures. A hierarchy can be built from the items of dimensions and its levels can be denominated. There are measures in a certain dimension of the data cube. Their descriptions are included in the subsystem "Measures" of the metainformation system. For example:

– Data cube: Vocational education;
– Reference units: Institutes of initial education;
– Reference period: 2005–2008;
– Dimensions: Time period, type of education, field of training in initial education;
– Measures of the cube:
  – Number of students in vocational education (heads;
  – Number of female students in vocational education (heads);
  – Number of students passed vocational exam (heads).

This data cube belongs to the statistical domain "Formal education". Clicking the button "Methodology", the whole documentation can be read.

3.9. Metadata of the production database

Metadata of the production database connects the description of concepts, classifications and measures to the tables of the production database. According to the naming convention, the columns of the tables have measure identifiers and their rows have nomenclature identifiers. That is how the columns of the tables navigate the users to the concerning documentation.

3.10. Grouping the metadata of the production database

Metadata of the production database can be divided into three groups. Metadata of the "*Data collection arrangement*" phase embraces information on data collections, questionnaires, instructions, supplements; on the scope of data providers so that an algorithm can be designed to send questionnaires out and to control data capture; on the organisation and employees of the HCSO; on work flow parameters.

Metadata of "*Data preparation*" include the cells of questionnaires by their physical location; the cross-references between cells of questionnaires and measures, nomenclatures as well as database cells; and the check rules.

The planning of the "*Data processing*" subsystem along with that of its metadata is currently underway.

4. Relevant international standards and the practice of other statistical institutes

This chapter provides information on the international standards as well as on the metainformation system of other statistical institutions and compares the latter with that of the HCSO.

4.1. International standards

Adequacy to international standards has become more and more important for Hungary since its EU accession. To this end, the Eurostat and the HCSO compared the metainformation systems of the European national statistical institutes (NSIs).

Since every NSI wants to meet its own users' requirements in the establishment of its metainformation system, the purpose of these comparisons was not the ranking of NSIs. The aim was always a mapping of best practices.

Among the formerly mentioned standards, the most important ones are SDMX and the Euro-SDMX Metadata Structure (ESMS) because these are the standards which ensure the comparability of different metainformation systems and thereby, the comparability of statistical data.

Structured data and metadata exchange

The abbreviation SDMX means structured data and metadata exchange, but it is actually more than that. Besides the standardisation of the IT background for the statistical data and metadata exchange, it standardises the recommended content and concepts

of the metainformation systems as well as statistical data processing. The most important white paper of the SDMX standard is the Content Oriented Guidelines, which contains the formerly mentioned elements. This standard was established by seven international organisations (Bank of International Settlements, European Central Bank, Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, World Bank). Adequacy to the SDMX standard can hardly be stated since its application is not compulsory. The reason for this is that although the standard was established for the whole world, it is unlikely to prescribe for all the NSIs to restructure their whole statistical system accordingly. However, it is advisable to introduce SDMX into the whole statistical data processing if there is no "traditional" metainformation system at a NSI. Besides, the NSIs are required to map their system to the SDMX standard for ensuring international comparability.

For further information on the SDMX standard please visit www.sdmx.org homepage.

Euro-SDMX Metadata Structure

The establishment of the standard ESMS was induced by the 2008 version of SDMX. It is actually the European version of SDMX (but it does not contain all concepts used in that). This standard focuses on the special requirements of the EU and as a result, ensures the comparability of different metainformation systems and statistical data. Its concepts are published in an EU Recommendation (*EU* [2009]).

4.2. The practice of other statistical institutes

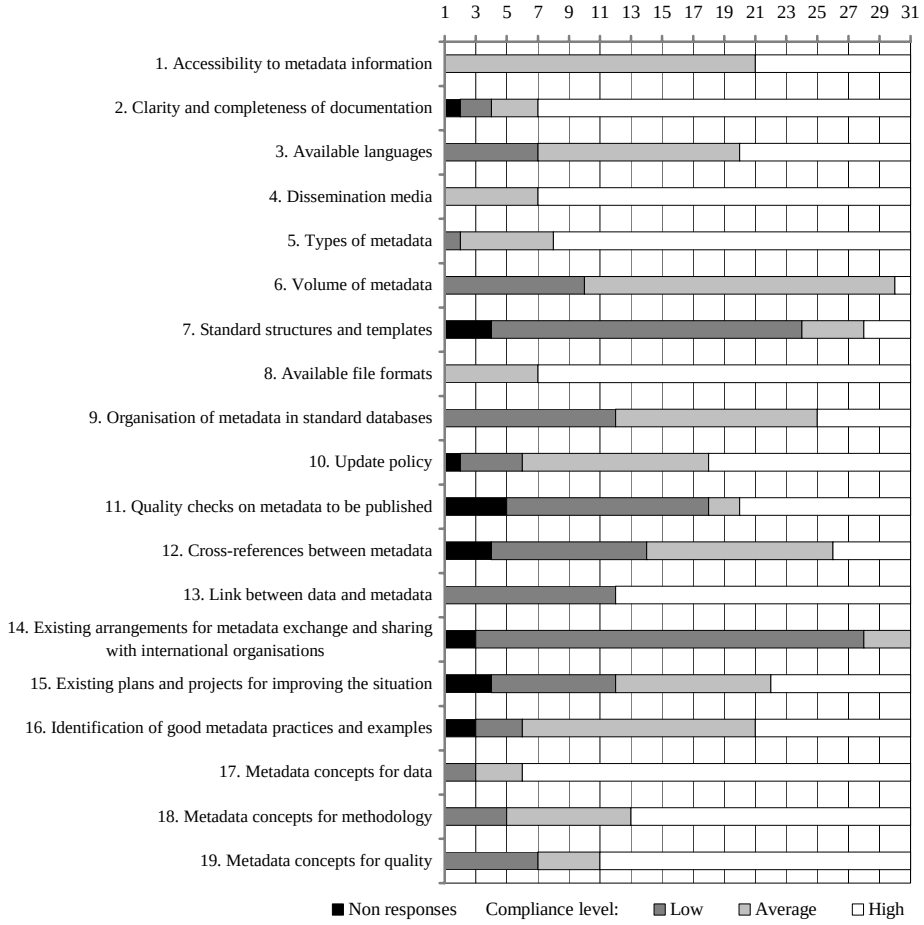
This subchapter is based on the case studies presented at the 2009 METIS Working Group Meeting. As further sources, personal experiences and the analysis of a Eurostat survey were used. This latter was based on a questionnaire on the actual condition of the metainformation systems of different European NSIs and on the analysis of NSI websites containing large amount of metadata. The survey was carried out in 2008, and its results were published in June 2009. The analysis of results was interpreted at the 2009 META Working Group Meeting in Luxembourg. The questionnaire was sent to 31 countries and it contained several questions on the metainformation system. (See Annex 3 www.ksh.hu/statszemle) Based on the analysis, the next statements can be made for all the countries.

The following criteria are highly met:

– Clarity and completeness of documentation;
– Dissemination media;

– Types of metadata;
– Available file formats;
– Metadata concepts for data;
– Metadata concepts for methodology;
– Metadata concepts for quality.

Figure 3. The number of NSIs by compliance level



Source: Document of Meta Working Group, 2009. Eurostat. Luxembourg.

The following criteria are met on average:

– Accessibility to metadata information;
– Available languages;

– Volume of metadata;
– Organization of metadata in standard databases;
– Update policy;
– Quality checks on metadata to be published;
– Cross references between metadata;
– Link between data and metadata;
– Existing plans and projects for improving the situation;
– Identification of good metadata practices and examples.

The following criteria are poorly met:

– Standard structures and templates;
– Information on existing arrangements for metadata exchange and sharing with international organisations.

A more detailed introduction of Norway, Portugal, Canada and New-Zealand is given next.

The structure of the metainformation system

In *Norway*, the metadata need of the statistical data processing system is filled by several sub-systems. They are the following:

– Datadok for archiving;
– Vardok for variables and their definitions;
– Stabas for standard classifications;
– Meta website for making the Norwegian system visible for external users;
– A database for data preparation;
– Metadata for researchers;
– A sub-system of statistical methodological documentation;
– StatBank, which is similar to the dissemination database of the HCSO;
– Metadb containing historical data on social security and national education.

In *Portugal*, the metadata need of the integrated statistical data processing system is also satisfied by several sub-systems. (See Figure 4.)

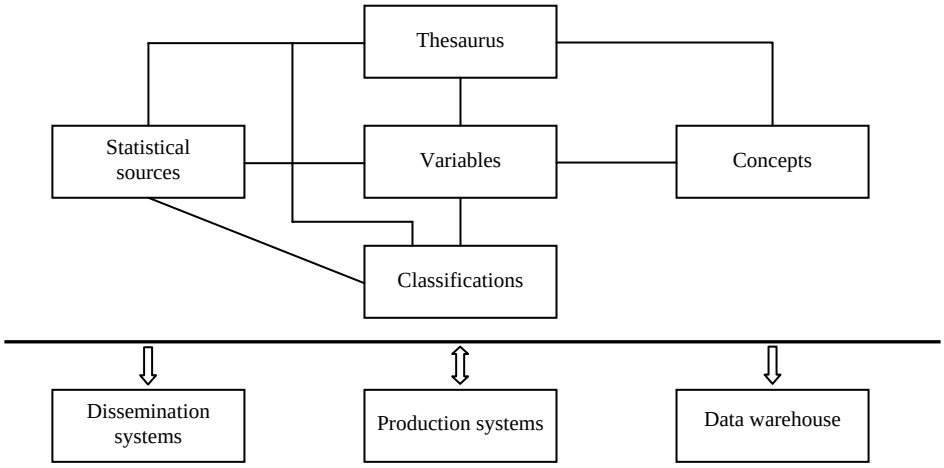
Its sub-systems are as follows:

– Concepts;

– Classifications;

– Statistical sources, which cover methodological documentations, variables, and – in the future – the metadata of administrative data.

Figure 4. The structure of the metainformation system of Statistics Portugal



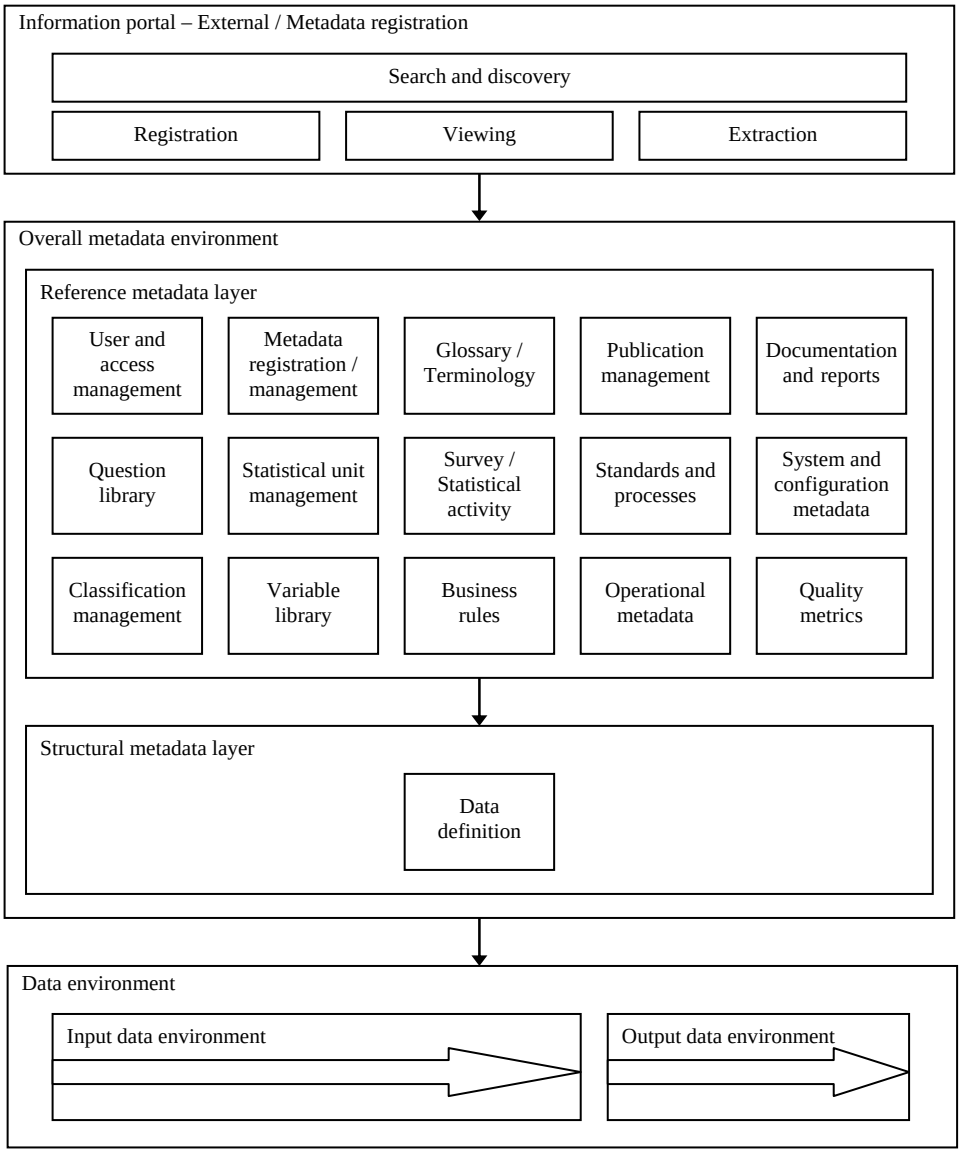
Source: METIS session document, 2009.

The reason for choosing *Canada* and *New-Zealand* for the analysis is the fact that the systems of these countries contain elements, which became part of the SDMX standard. The statistical data processing model of New-Zealand was the role model for the generic statistical business process model (GSBPM) in SMDX. It is a uniform model used throughout the world. Statistics Canada has made the New-Zealand-model-based GSBPM complete by adding two further sub-processes: "Archiving" and "Evaluation". Although these systems could be analysed in several ways, we emphasize only their organisational issues and structure, as the most cardinal ones. (See Figures 5 and 6.)

As it is shown in Figure 4 and just like in the formerly mentioned countries, the metadata need of the statistical data processing is met by several sub-systems. In the figure, for better understanding, the connections between the sub-systems are not illustrated. However, in reality, these sub-systems are connected to each other forming an integrated metainformation system.

Statistics Canada currently uses a metainformation system called IMDB (Integrated Metadatabase), which is handling metadata for data preparation and dissemination. (See Figure 6.) The metadata need of other statistical data sub-processes are satisfied by several additional metadatabases. Statistics Canada is working on a system for managing metadata through the whole statistical data processing.

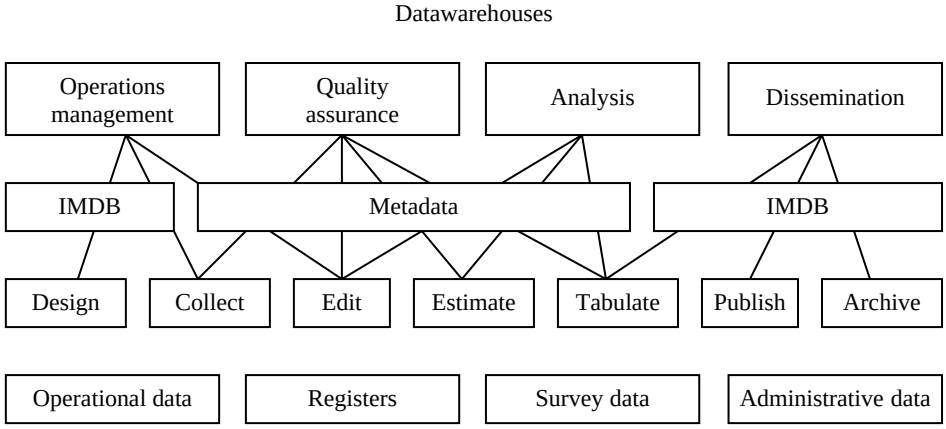
Figure 5. The structure of the metainformation system of Statistics New-Zealand



Source: METIS session document, 2009.

In the HCSO, the structure of the metainformation system (see Figure 1) contains the results of the latest improvement, which aimed at improving the metainformation system of the institute.

Figure 6. The role of IMDB at Statistics Canada



Source: METIS session document, 2009.

It is rather difficult to compare the metainformation systems of different countries. There is a great deal in their content that is similar. It is clear that the aim of every NSI is to make metadata suitable for describing more phases of the statistical data processing than it is done today in order to establish an integrated system.

Organisational issues

In the HCSO, the responsibility, concerning the metainformation system, is shared. The subject-matter departments are responsible for the metadata of those statistical domains, which they are in charge of. The IT Department is having control over the IT background. The Statistical Research and Methodology Department is responsible for working out the methodology, drawing up and enforcing the common rules and for checking quality. The Dissemination Department is accountable for disseminating metadata in different publications. The responsibilities are clear and distinct as recorded in the Establishment and Operation Rules. Since responsibilities and tasks are shared in the HCSO, there are four persons engaged in the methodological and IT background of the metainformation system. Besides, approximately fifty persons working at the subject-matter departments are responsible for maintaining metadata of subject-matter statistics. The advantage of the Hungarian structure is that by applying common rules, the special needs of subject-matter statistics can be taken into consideration.

There are several solutions for placing a metainformation system within an organisation. Some organisations apply a centralised (for example Statistics Portugal), others a decentralised method (for example Statistics Finland), and of course, there

are also cases for mixed applications (Statistics Norway, Statistics Canada, Statistics New-Zealand, HCSO). Every NSI chooses the most favourable solution for itself according to its traditions.

5. Metainformation system developments on the HCSO website

The meta-website displaying metadata fills an old gap. In its planning phase, meta-websites of other national statistical institutes (their content, functionality, etc.) were taken into account. We considered only those models exemplary, which showed an integrated system. The strength of the solution chosen by the HCSO is that it is database-based; therefore cross-references between subsystems and objects can be stored. As a result, users have an option for the navigation between objects on the website.

The number of metadata available on the meta-website, 2009

Denomination	Number
Definition of concepts	1 996
Classifications	24
Item definitions of classifications	22 339
Statistical domains	104
Of which:	
Statistical domains including concepts, classifications, legal base, respondents	104
Statistical domains including also short descriptions	77
Statistical domains including also methodology and quality	61
Further metadata:	
Data collection as a data source	247
Administrative data as a data source	98
Legal base with a link to the statistical domain	144

The core of the system can be summarized as follows: metadata can be achieved by users either using the link between data and metadata or starting from the root menu.

The first case means that there is a “Methodology” button next to the statistical data, which navigates the users to the corresponding metadata (methodological document).

In the second case, users can have information on the statistical domains (their concepts, classifications, data sources, etc.) irrespective of the statistical data. (For further information, please visit the website of the HCSO (http://portal.ksh.hu/pls/ksh/ksh_web.meta.main?p_lang=HU)).

Metadata are updated daily on the webserver owing to which the users can have up-to-date information on the statistical domains.

In the previous year, the "Metainformation" website had an average of 1 650 Hungarian and 250 English speaking visitors per month. Visitors may send their remarks via e-mail address (meta@ksh.hu) to the subject-matter statisticians and to the developers of the metainformation system to support the further improvements.

References

BARACZA, G. [1980]: *A statisztikai meta-adatbázis információleíró alrendszere*. KSH Rendszerfejlesztési közlemények. Vol. 1. pp. 21–34.

DÖRNYEI, J. [1983]: *The Role of Metainformation in Statistical Integration*. 44th Session of the International Statistical Institute. 12–22 September. Madrid. Working Paper.

DÖRNYEI, J. – GYÖRKI, I. [1975]: *Report for the Group of Rapporteurs on the Industrial Statistical Database System*. Seminar on Integrated Statistical Information Systems and Related Matters. Bratislava. Working Paper.

EU (EUROPEAN UNION) [2009]: Commission Recommendation of 23 June 2009 on Reference Metadata for the European Statistical System (2009/498/EC). *Official Journal of the European Union*. L. 168. pp. 50–55.

GYÖRKI, I. [1980]: *A statisztikai meta-információrendszer és meta-adatbázis*. KSH Rendszerfejlesztési közlemények. Vol. 1. pp. 7–20.

GYÖRKI, I. [1996]: *Survey Control as a Subsystem of the Statistical Information System*. Seminar on Integrated Statistical Information Systems and Related Matters. Bratislava. Working Paper.

GYÖRKI, I. – PAPP, E. [1985]: *SOLAR: Statistical Data Base with Interactive Query Facility.* Seminar on Integrated Statistical Information Systems and Related Matters. Bratislava. Working Paper.

GYÖRKI, I. – RÓNAI, M. [1999]: *Metadata Management*. Conference of European Statisticians. UNECE (United Nations Economic Commission for Europe) Work Session on Statistical Metadata. 22–24 September. Geneva. Working paper.

PUKLI, P. – VÉGVÁRI, J. [2004]: A statisztika: tudomány és szakma. *Statisztikai Szemle*. Vol. 82. No. 1. pp. 5–30.

SUNDGREN, B. – MALMBORG, E. [1994]: *Integration of Statistical Information Systems: Theory and Practice*. 7th International Conference on Scientific and Statistical Database Management. 28–30 September. Charlottesville, Virginia. Working Paper.

UN (UNITED NATIONS) [1995]: *Guidelines for the Modelling of Statistical Data and Metadata*. New York, Geneva.

UN (UNITED NATIONS) [2000]: *Guidelines for Statistical Metadata on the Internet*. Statistical Standards and Studies. No. 52. Geneva.

Website of METIS-wiki http://www1.unece.org/

Website of OECD www.oecd.org

Website of SDMX http://sdmx.org/

Website of Statistics New-Zealand http://www.stats.govt.nz

Website of Statistics Norway http://www.ssb.no/

Annex 1

Schema of the methodological documentation on statistical domains

1. Descriptions, content
- 1.1. Descriptions
  - 1.1.1. Identifier of the statistical domain
  - 1.1.2. Denomination of the statistical domain
  - 1.1.3. Denomination and availability of the responsible department
  - 1.1.4. Identifier, name and availability of the responsible person
- 1.2. Legal base:
  - 1.2.1. Hungarian
  - 1.2.2. International (EU and/or other international organisation)
- 1.3. Purpose of the statistical domain
- 1.4. Content of the statistical domain
  - 1.4.1. Short description. (Most important collected and aggregated measures)
  - 1.4.2. Classification used in the statistical domain (for example sex, age, qualification, etc.)
- 1.5. Data sources – short description
- 1.6. Publications
  - 1.6.1. Forms of dissemination
  - 1.6.2. Timeliness, revision policy and practice
- 1.7. History of the statistical domain

2. Concepts and definitions used in the statistical domain
- 2.1. Definition of concept
- 2.2. Source of definition
- 2.3. Validity period of definition
- 2.4. Cross-references between concepts

3. Classifications of the statistical domain
- 3.1. Denomination of classification
- 3.2. Varieties of classification
- 3.3. Items of classifications (code, denomination, definitions)
- 3.4. Cross-references between classifications

4. Data production methods
- 4.1. Population, sampling frame
- 4.2. Sampling
- 4.3. Data collection methods
- 4.4. Data capture and data editing method
- 4.5. Imputation
- 4.6. Data processing, estimation, calculation
- 4.7. Methodological publication(s)

5. Data quality
- 5.1. Relevance
- 5.2. Accuracy
- 5.3. Comparability, coherence
  - 5.3.1. Comparability between geographical areas (with other countries, within the country)
  - 5.3.2. Comparability over time
  - 5.3.3. Coherence with other statistics
- 5.4. Quality report

6. Data sources
- 6.1. Data collection
  - 6.1.1. Reference year
  - 6.1.2. Identification code
  - 6.1.3. Denomination of data collection

6.1.4. Enactor
6.1.5. Conductor
6.1.6. Legal status
6.1.7. Change of data collection
6.1.8. Frequency of data collection
6.1.9. Scope of data providers
6.1.10. Deadline of arrival
6.1.11. Type of data collection
6.1.12. Data collection methods
6.1.13. Status of data collection
6.1.14. Year of change
6.1.15. Year of cessation
6.1.16. Questionnaire
6.1.17. Instructions to questionnaire
6.1.18. Supplements to questionnaire

6.2. Administrative data sources
6.2.1. Reference year
6.2.2. Identification code
6.2.3. Denomination of data received
6.2.4. Content of data received
6.2.5. Data supplier organisation
6.2.6. Responsible person
6.2.7. Frequency of data received

Source:
http://portal.ksh.hu/pls/ksh/ksh_web.meta.menu?p_lang=EN&p_menu_id=110&p_session_id=55153688

Annex 2

The hierarchy of statistical domains

- Population statistics
  - Population
    - Number and structure of population
    - Calculated population
  - Vital events
    - Live birth
    - Death
    - Marriage
    - Divorce
    - Foetal losses
    - Migration
      - Information on mobility - census
      - Internal migration
      - International migration
- Social statistics
  - Employment, labour force, earning
    - Institutional labour statistics
    - Community labour statistics
    - Occupation - census
  - Social stratification, living conditions
    - Living conditions, poverty, social exclusion
    - Time use
  - Households, families
    - Demography of households and families - census
    - Budget and equipment of households
  - Housing and public utilities
    - Buildings, dwellings - census
    - Real property of the municipalities
    - Dwelling management of local government
    - Dwelling stock
    - Housing credits
    - Construction permits
    - Dwelling construction and cessation
    - Public utilities
  - Health care
    - Primary health care
    - Health care services
    - Personnel and equipment of health care system
    - Prevention
    - Morbidity, accidents
    - Health care accounts
  - Social provision
    - Retirement allowances, sick-pay, family and social benefits
    - Child protection
    - Social services
    - European system of social protection statistics (ESSPROS)
  - Education
    - Educational attainment - census
    - Formal education
    - Lifelong learning

- Culture, sport
  - Cultural services
  - Entertaining and other cultural activities
  - Sport
- Justice
  - Court cases
  - Discovered crimes and perpetrators
  - Convicts with definite sentence
  - Offences

General economic statistics

- National accounts
  - The production of GDP
  - The final expenditure of GDP
  - Income accounts
  - Foreign direct investment
  - Input-Output tables, supply and use tables
- Demography, productivity, expenditure of business units
  - Business demography
  - Nonprofit organisations
  - Performance and expenses of enterprises
    - Annual performance indicators of enterprises
    - Short-term statistics
  - Raw material consumption statistics
  - Investment
- Research, development, innovation
  - Research and development (R&D) statistics
  - Innovation statistics
- External trade statistics
  - External trade in goods
  - External trade in services
- Energy statistics
- Financial statistics
- Price statistics
  - Consumer prices
  - Industrial producer prices
  - Construction producer prices
  - Agricultural prices
    - Agricultural producer prices
    - Agricultural input prices
  - Prices for external trade
  - Service producer prices
  - Calculation of purchasing power parity

Economic statistics by divisions

- Agriculture, forestry, fishery
  - Land use
  - Economic Accounts for Agriculture
  - Crop production
  - Livestock and animal products
  - Statistics on fruit and vine plantations
  - Forestry
  - Fishery
  - Structure of agricultural activity
- Industry

- Subannual industry statistics
- Annual statistics on industrial production
- Monthly statistics on industrial production
- Construction statistics
- Internal trade
  - Retail network
  - Retail trade turnover
    - Retail trade turnover by groups of shops
    - Retail trade turnover by commodity groups
- Tourism, catering
  - Accommodation services
  - Organized tourism
  - Tourism indicators
  - Tourism demand
- Transport
  - Road network
  - Stock of road vehicles
  - Modes of transport
    - Railway and intermodal transport
    - Inland passenger and road freight transport
    - Transport by pipeline
    - Inland waterway transport
    - Air transport
  - Road traffic accidents involving personal injuries
- Business service statistics
- Information, communication
  - Post and telecommunication
  - Internet services
  - Information technology (IT) services
  - ICT usage in enterprises
  - Content services

Environmental statistics

- Air pollution
- Biodiversity
- Agricultural environment
- Energy and environment
- Forest
- Environmental health
- Environmental industry
- Noise
- Environmental expenditure
- Waste statistics
- Transport and environment
- Water statistics

Regional statistics

Annex 3

Self assessment questionnaire on national metainformation systems

	EUROPEAN COMMISSION **EUROSTAT** **Directorate B: Statistical methods and tools; Dissemination** **Unit B 4: Reference databases**

QUESTIONNAIRE ON THE ASSESSMENT OF NATIONAL METADATA

(11/08/2008)

QUESTIONNAIRE

Contact Details

Please provide the following contact details:

National Statistical Institute	
Postal Address	
Name	
Department	
Position (e.g. database manager, metadata coordinator, compiler,…)	
e-mail	
Telephone	
Fax	

Please answer the following questions taking into consideration the overall content, functionality and structure of your national metadata system.

1. Please provide information on which types of metadata are made available on the web by your organization (e.g. reference metadata, structural metadata, methodological notes, classifications and codes, questionnaire descriptions, release calendars,…). Please provide also additional information in annexes if deemed necessary.

Types of metadata	Explanation	Comments

2. Does your institute have a corporate strategy for metadata or independent systems are built as needs arise? (Choose one or more answers and provide additional comments if needed)

	Yes	No	Comments
We have a corporate metadata strategy			
A metadata strategy is being formulated			
We use independent systems as needs arise			
We are developing a metadata-driven system			
Metadata are mainly based on documentation systems			

Additional Comments:

If published, where can your corporate strategy for metadata be found?

3. Does your organization have an organizational entity (e.g. unit, department, etc.) dealing with metadata?

Yes No

If yes, please specify this entity as follows:

Unit/Department – Name of the Department
A working team or internal committee
Other organizational issues Please describe: …… …….…….

4. Which are the responsibilities of this entity (Choose one or more answers and provide additional comments if needed)

Types of responsibilities	Yes	No	Comments
Standardisation of metadata			
Coordination of metadata			
Storage and maintenance of metadata			
Dissemination of metadata			
Quality checks of metadata			
Other responsibilities Please describe: …… ……			

5. Please indicate the media used for metadata dissemination:

Media used for the dissemination of metadata	Links (if available)	Description and Comments[1]
General web-site (static HTML pages and files ready for download)		
Statistical online database (database from which users can extract data and metadata by using a query)		
Publications		
Other Media (i.e. CD- ROM, special IT applications)		

[1] Provide a short description referring the types of metadata disseminated, the frequency of dissemination, and the degree of domain coverage.

6. Please describe the existence of search facilities for metadata available on your webpage and the respective IT formats used for disseminating metadata:

Search Facilities	Yes	No	Comments
1. Sitemap/table of contents of the website			
2. Subject matter classification			
3. Keywords			
4. Free text search			
5. RSS feeds[2]			
6. Other search facilities Please specify: …… ……			

IT formats used for disseminating metadata	Yes	No	Comments
– html			
– pdf			
– doc			
– xml			
– Other IT formats (statistical package, database format) Please specify: …… ……			

[2] RSS is a family of Web feed formats used to publish frequently updated content such as blog entries, news headlines, and podcasts in a standardized format.

7. Please indicate the language(s) in which your national metadata is available on your general web-site and within your statistical on-line databases:

Language	General web-site	Statistical on-line database
English		
National languages: …….……		

If your metadata is available in more than one language, is the information available identical between the national and English language?

Yes No

Additional Comments:

8. Are you using the following metadata standards:

Standard structures / Templates	Yes	No	Comments
DQAF/SDDS			
SDMX CDC (Cross-Domain Concepts) and ESMS			
Metadata Registry Standard ISO/IEC 11179			
Data Document Initiative (DDI)			
Dublin Core			
Neuchâtel classification model			
Neuchâtel variables model			
Other standards (please specify): …… ……			

9. Please indicate whether your metadata are stored in a database or in a file system and to what extent this database is available for external users:

Types of storage of metadata	Yes	No	Comments
Stored in a file system			
Stored in a database			
Other storage (please describe): …… ……			

Are your national metadata available free of charge for external users?

Yes No

If NO, please specify what metadata is to be purchased by external users.

Additional Comments:

10. Please indicate if you use the following statistical concepts within your national metadata:

Concept Name	Use at national level		
	Yes	No	Comments
Contact details			
Metadata update			
Short description			
Classification system used			
Concepts and definitions (main variables)			
Statistical units			
Reference area (geographical coverage)			
Time coverage			
Unit of measure			
Reference period			
Legal acts and other agreements			
Confidentiality			
Data release policy/calendar			

Concept Name	Use at national level		
	Yes	No	Comments
Frequency of dissemination			
Dissemination format			
Accessibility of documentation			
Quality management			
Relevance			
Accuracy and reliability			
Timeliness and punctuality			
Comparability			
Coherence			
Cost and burden			
Data revision			
Type of source data			
Frequency of data collection			
Data collection methods			
Data validation			
Data compilation			
Adjustments			

Please provide additional information regarding the following issues:

If you use additional statistical concepts in your national metadata, please list these concepts used in your organisation but not listed above:

Provide hyperlinks to additional information pages related to metadata which you use:

Additional comments:

11. Are you using the following means for ensuring the quality of metadata:

	Yes	No	Comments
Standard structural metadata (e.g. codes for data tables)			
Standard metadata terminology			
Metadata release calendar			
Date of update of existing metadata			
Other quality measures (please specify): …… ……			

Additional comments related to the quality assurance of metadata

Does a central quality control of metadata exist in your organisation?

Yes No

If yes, please specify what this central screening and validation refers to	Yes	No	Comments
Respect of the national metadata standard			
Readability of metadata			
Translations of metadata			
Accessibility of metadata			
Other issues (please specify): …… ……			

Other Comments:

12. Does your institution exchange metadata with other international organizations? How is this information exchanged?

Organization receiving your national metadata	Metadata standards used	Electronic format used for metadata exchange	Comments
Word Bank			
OECD			
IMF			
ECB			
United Nations			
Other (please specify): …… ……			

13. How does your organization assess and satisfy users' needs related to metadata?

	Comments
Can user contact the national institute with questions or problems on metadata? How?	
What are the mechanisms and IT tools used for the collection of the feedback of users?	
How is this feedback from users used?	
Are FAQ, helpdesk or help links available ?	

Additional Comments:

Improvements of your national metadata systems

14. Does your organization have any current or future projects and plans related to the improvement of your national metadata system?

Yes No

If yes, please list the projects and plans here; use also the sample table below to describe each of the plans or projects in more details.

Note: Copy the sample template below to describe ongoing and planned plans and projects if more space is needed

Sample table:

Name/title of the plan or project:

Main objectives:

Metadata standard to be adopted (if applicable):

Challenges and solutions:

Expected date of completion and full implementation:

15. Finally, if you have additional comments on topics not directly addressed in the questionnaire you may express them in the space below: