WP. 26
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 work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (iv): Implementation

DEVELOPMENT OF A METADATA SYSTEM AT THE CROATIAN BUREAU OF STATISTICS

Supporting Paper

Submitted by Central Bureau of Statistics of the Republic of Croatia[1]

I. INTRODUCTION

1. The project to develop a central metadata repository and public macro database, in the frame of Integrated Statistical Information System (ISIS), started at the Croatian Bureau of Statistics (CBS) in 2002. The primary target of ISIS is to manage a completely metadata-driven automated processing system, which requires a well-structured metadata system. The outcome of the project, which ended in 2005, was a new metadata model called CROMETA #2. The model is based on the METANET Reference Model ^{TM} and customized for needs of CBS. The new model includes specifics of a previous version of metadata system at CBS, as well as concepts necessary to run PC-Axis as a main dissemination tool and to connect with BRIDGE^{NA} as a tool for maintaining classifications.

2. The model has been physically implemented on relational database management system and accompanied with an application for browsing and maintaining metadata. The development is ongoing, and both the metadata model and the maintenance tool are in the phase of testing. Physical implementation of the metadata model doesn't cover the whole model, just a part of it. The plan is to include all metadata needed for successful creation of so-called 'generators' – programs for building applications for data processing. The final metadata system solution should be presented in 2007.

II. HISTORY OF METADATA IN CENTRAL BUREAU OF STATISTICS

3. Central Bureau of Statistics of Republic of Croatia started to develop the metadata system on the client/server platform a few years ago. The start was actually in 2000, when the first version of metadatabase was designed and implemented. The main reason for the development was to create a completely metadata-driven automated processing system. Production of statistics in CBS was carried out on a mainframe and just a small part was developed for a client/server environment. Data processing in a mainframe environment was solved quite well with so-called 'generators', the applications for generating programs for data processing. This technique covers large part of all surveys that are carried out on mainframe, since 80% of them are very similar. The growing costs of maintenance of two platforms forced CBS to decide to transfer production to a

[1] Prepared by Maja Ledic Blazevic majalb@dzs.hr

client/server environment and use the same old idea of 'generators'. The precondition for realisation of that plan was to have well-structured metadata.

4. The other reason of same importance for the development of a metadata system was to harmonize statistical methods and data according to international standards, and to have compatible, or at least comparable data with international statistical organisations.

5. The first metadata repository, named *Inventory*, was developed based on a custom data model created in CBS. The implementation was carried out on a relational DBMS, MS SQL Server, as it was a pre-existing environment for database development at CBS. This repository was initially filled with collected metadata from a few selected surveys, mainly in the area of conceptual and operational metadata. There were problems in metadata collection because subject matter staff were not accustomed to thinking conceptually instead of considering contents of data. Other sources of metadata were also included in this inventory of metadata in CBS, for example, from the publishing department and Program of Statistical Activities for CBS that is approved by Parliament.

6. The idea of building 'generators' based on that first metadata system was not implemented because in the beginning of 2002 a joint project of CBS and Statistics Sweden (SCB) was launched. This project was sponsored by Swedish Agency for International Development Cooperation (SIDA) and supported by many consultants hired by Statistics Sweden. The primary goal of the project was to develop a central metadata repository and a public macro database. Early stages of the project were more dedicated to public macro database development. At that time, 2002-2003, a decision was made to use PC-Axis software as the dissemination tool at CBS. It was first offered as the dissemination tool with data from the Agriculture Census 2003 and was very well accepted by statisticians and external users.

7. *PC-Axis* has its own metadata with a specific model behind it. Data was collected from some statistical surveys and added to the new statistical warehouse. Cubes of aggregated data were defined based on that micro data and described by metadata. Then PC-Axis could be used to produce statistical tables based on a user's selection of themes, matrices and variables. The metadata model of PC-Axis covered different areas of metadata to *Inventory's* model, and didn't fulfil the needs of CBS. Another solution was needed. Data and metadata prepared for *PC-Axis* was sufficient to develop a customized prototype of metadata-driven web solution for data dissemination in 2004.

8. In 2002 BRIDGE^{NA} was introduced at CBS as the metadata maintenance tool. After testing how much it covers CBS needs, only module ClassE was chosen to be used as a maintenance tool for classifications. It is still in use today. Behind ClassE lays the Terminology Model, a very developed model of metadata dealing with classifications.

9. In 2004, a decision was made to start development of a custom metadata model, based on Reference Model ^{TM} that was developed during the METANET project of Eurostat (2000-2003). Reference Model ^{TM} includes already the above-mentioned Terminology Model, as well other models of statistical metadata, and it is their common denominator.

10. In 2004-2005 the focus of development was set on building a central metadata repository rather than a public macro database. At that time the project was divided in two subprojects: metadata methodology development; and technical implementation. The final outcome of this stage, and the entire project, is the CROMETA #2 metadata model and its implementation.

III. METADATA METHODOLOGY

11. As noted before, a new metadata model was based on Reference Model ^{TM,} but existing metadata consumers, *Inventory*, *PC-Axis* and BRIDGE^{NA}, also had to be included. There was no need to work on BRIDGE^{NA}, but other metadata had to be mapped to the Reference Model ^{TM}. Thorough analysis provided specifications on how to integrate specifics of *Inventory* and *PC-Axis* Macrometa into the new model.

12. A group responsible for metadata methodology, consisting of both statisticians and information technology experts, worked in a technical implementation group on specifications for how to build metadata. These specifications were created on the basis of Reference Model [TM] analysis. A specification consists of a concept, its characteristics and references to other concepts. The extensions of Reference Model [TM] were created in the form of additional characteristics of existing concepts when possible, and additional new concepts and their references when needed. It is important to stress that not many differences were found between metadata models, and not many extensions were added.

13. The Reference Model [TM] was hard to understand due its complexity, especially for non-methodologists in the technical implementation group. Therefore, a document was created about the whole model and metadata process, important in many ways. The main idea of the document was to recognize basic metadata concepts and their relationships. It provided simple vocabulary for non-experts and examples to understand the concepts. It contains procedures how to process metadata and in which order. The rules and behaviour of concepts and characteristics were also described for that purpose.

Examples from the document about metadata process:

*All objects that are observed at CBS are defined as **Statistical object types**:*

- *Person*
- *Household*
- *Enterprise*
- *Commodity*

A statistical object type may further be divided hierarchically into sub-types, as far as needed.

- *Person:*
  - *Employed person:*
    - *Full-time employed person*
    - *Part-time employed person*
  - *Unemployed person*

*All objects that are generally measured at CBS are defined as **Global variables**:*

- *Income*
- *Age*
- *Activity*
- *Occupation*

*When defining what characteristics to observe for a certain object, statistical object types are combined with global variables to form **Object variables**.*

- *Income of person*
- *Age of person*
- *Income of enterprise*
- *Activity of enterprise*

14. When developing a methodology for CROMETA #2, metadata concepts were categorized into nine metadata sections or areas according to the logical relationships between them. These sections can be described as follows:

a. Studies and questionnaires

This section contains metadata regarding the various studies/surveys carried out by the statistical office, for example, all metadata regarding studies, versions of them, general method for performing them, etc. Examples of metadata concepts in this section are Study, Study Version, Questionnaire, Question, Interview method, Population, Coverage type, etc.

b. Variables and measurements

This section contains metadata about variables collected within the frame of the statistical activity, as well as the methods and ways of measuring them. In this section variables are described from a

more general point of view, regardless of use and implementation in different studies. Examples of metadata concepts belonging to this section are Global variable, Object variable, Measure unit, Basic measure unit, etc.

c. Processing and validation rules

This metadata section could be described as a combination of variables and measurements versus studies and questionnaire sections, meaning that variables and measurements are put in the context of studies/surveys. Here are metadata regarding processing of studies, including validation, production processes, registers, cubes and tables created. Examples of metadata belonging to this section are Context data element, Data collection, Derivation rule, Register, Cube, and Table, etc.

d. Classifications

This section contains all metadata concerning classifications. Examples of metadata concepts in this section are Classification family, Classification, Classification version, Classification item, Correspondence table, etc.

e. Publications

This section contains metadata concerning publication and dissemination of statistics. Examples of metadata concepts belonging to this section are Publication series, Publication, Edition, etc.

f. Organisational structure

This section contains all metadata concerning the organisation, related organisations, persons working within the organisation and the responsibility of the latter, etc. Examples of metadata concepts belonging to this section are Organisation, Organisation scheme, Person, Organisation level, etc.

g. History and version handling

Metadata objects as instances of metadata concepts may exist in an indefinite number of versions. Version management is extremely important in order to keep consistency in metadata and data over time. The main part of the history and version handling is implemented through methods applied on metadata objects. Every metadata object has several versions valid in some period of time and is described with a certain status. An example of metadata belonging to this section is Update information, used for logging all changes made to a metadata object over time.

h. General metadata concepts

All metadata contains some general characteristics, and they are sometimes metadata concepts themselves. All these general metadata concepts are kept within this section. Also, there are metadata concepts used in all, or at least several sections, and they too are placed in this section. Examples of metadata belonging to this section are Language, Keyword, Footnote, Theme, Statistical object type, Status, etc.

i. Access and authorization rules

This section contains metadata about how metadata and data could be accessed by internal and external users and the authorisation rules applied to metadata and data. Examples of metadata belonging to this section are User group, Privilege, Access condition, Access form, etc.

15. It is important to stress that metadata concepts in CROMETA #2 can also be grouped by any other type of categorisation, e.g. according to the approach of the OECD, it is possible to categorize metadata by usage (1).

16. The priorities in implementation are also defined from the document about metadata process. It was not intended to implement everything from the CROMETA #2 model, but the concepts that are first on the priority list. It was decided to work first on general metadata, history and version handling, access and authorization rules, organisational structure and variables and measurements. After those sections, focus went on studies and questionnaires and publications. Since classifications are handled by BRIDGE^{NA}, they are just referenced from CROMETA #2 for value domain definitions. Processing and validation rules section is planned for

development in 2006 because of its importance for subsequent project of ‘generators’, which are tightly connected to this section.

17. The document about metadata process provided identification of starting points of metadata collection. It was concluded that metadata can be independently collected from different sources, and collection was carried out in that way. Metadata was collected from *Inventory*, Program of Statistical Activities of CBS, *PC-Axis* Macrometa and few appointed surveys. Finally they were successfully transferred to CROMETA #2.

18. Metadata methodology development was a very important part of the work done, because it is a necessary precondition for successful work on technical implementation. Knowledge about the metadata model must be at a high level to create visible results. All participants working with metadata systems must have clear understanding of metadata and related concepts. The experience from *Inventory* shows that statisticians who tried to enter metadata had problems in that area, for example statistical objects and variables were mixed up.

IV. SOME SPECIFIC METADATA CONCEPTS OF CROMETA #2

19. As mentioned before, there were some extensions of the Reference Model [TM] in CROMETA #2. Some of them will be described below. They are defined as totally new concepts resulting from mapping work, e.g. Edition of publication or Geographic collection level. Alternatively, as new characteristics of existing concepts from Reference Model [TM], e.g. sequence number of a theme needed for specific predefined order of themes presentation, different from alphabetical order.

A. Hierarchical structures

20. The organisational structure of a statistical office can be changed over time. New departments can be established, or moved to another statistical division. Sometimes two departments are merged. A new concept of Organisation scheme (or structure) is therefore added to the metadata model. It represents the hierarchy of the organisations in some period of time. Relationships like parent and child organisations exist in the frame of this object and they are not directly connected to the organisation.

21. In the same way, the themes of statistics that are processed in surveys also create a hierarchy. This is a separate metadata object in CROMETA #2 called Theme structure. Even more, it is possible to define two different theme structures at the same time. In real life, for example, the theme structure used at CBS differs from the theme structure at EU, and some studies carried out in the EU are still not yet established in CBS. Naturally there are themes then without studies belonging in this so-called ‘EU structure’ but all the studies are connected to a theme belonging to ‘CBS structure’.

22. Similar to these ideas, questions in questionnaires are handled in the hierarchy separately from the concepts of question and questionnaire. This differs to Reference Model [TM] where ‘question’ is connected to a set of ‘sub-questions’ in the concept itself. In CROMETA #2 model the hierarchy of ‘main’ and ‘sub-questions’ is moved from the ‘question’ concept. This allows reuse and unification of questions and could improve the layout of questionnaires because very often the questions of same meaning are asked in many different ways.

B. **Status and version handling**

23. It is very important to explain history and version handling. Each metadata concept can have several versions that are valid in exclusive periods of time. At any one moment, one version is currently valid. This is determined with the Status of the metadata object. It is defined that there are six possible states of metadata. They are:

a. Under development
b. Released
c. Authorized
d. Archived
e. Frozen
f. Deleted

24. By default status is set to 'Under development', but it can be immediately set to 'Released' or even to 'Authorized' if the person who adds the object has sufficient privileges. There is one and only one version at a time that has status 'Authorized'. This is the currently valid version. 'Released' means only that work on metadata object is done and waits for approval. Once when metadata object gets status 'Authorized', it enters the queue of versions that have top (or first) version and a strong relation between themselves, meaning that they have a known predecessor and successor version. After a version with 'Authorized' status is replaced with new version with 'Authorized' status, it automatically gets its status changed to 'Archived'. It then becomes the predecessor version of the new one, and the new one becomes its successor version.

25. Status of an archived version can be changed to 'Frozen' and this protects it from any changes in the future. Status 'Deleted' is for keeping information about metadata that are deleted from some reason. Metadata in deleted state can be restored back by an administrator of the system. Metadata can be physically deleted only when they have status 'Under development'. Whenever some significant change on metadata must be applied, it is recommended to create a new version. If changes are small, for example, spelling corrections, they can be done without defining new version. Such changes are kept in the Update information, a general property of all metadata concepts in CROMETA #2.

C. **General properties**

26. General properties of metadata concepts are defined in a specific way in CROMETA #2. They are either generated from relationships with general metadata concepts, or they are attributes of the object. General metadata concepts, such as Language, Keyword, Footnote, Synonym, Milestone, Status, etc., can be applied to all metadata. Some of them are mandatory, like Status, but the majority are not. They can be freely used whenever needed for a better description of some object. Some relationships of metadata objects and general objects can be of type 1: n, like Keyword, Footnote, Document, Registered users, Contact persons, User groups, etc.

27. Each metadata object can have title and description in an unlimited number of languages. For any new language it is enough to add it to metadata system and then add language dependent characteristics in the newly added language. There could also be several types of title for each metadata concept, for example official, short, alternative, etc. This gives the opportunity to present metadata in different forms, for example different length of column and row headers in tables.

D. **Meta-metadata**

28. In the Access and authorization rules section, User group types are defined and they are connected to Privileges defined in the same section. User group of particular type inherits privileges from User group type. In practice, anyone has the right to browse or read data, expert user groups can add and edit their metadata, while only the administrators have the right to delete them as well.

29. Privilege is an example of many meta-metadata defined in CROMETA #2. Meta-metadata provides common vocabulary and whenever some property of metadata concept could be categorized by some definition, it was implemented as meta-metadata. Properties described by meta-metadata can be validated during the capture of metadata and this makes the overall solution more generic. Meta-metadata can also be

maintained through the common interface. Examples of meta-metadata in CROMETA #2 are Document type (legal base, methodology guidelines etc.), File format, Organisation level, Graduation, Publishing type, Variable type, Interview method, Sample method, Matrix type, etc.

E. **Metadata sources**

30. Regarding classifications, CBS is using the ClassE module from BRIDGE[NA] software as a tool for classification maintenance. When classifications development started in CBS, thorough studies of the Neuchâtel terminology model was initiated. This model was selected since it is both a terminology and conceptual model, and that makes it generally applicable and independent from the software platform. Successful tests of including classifications in CROMETA #2 have been carried out. The integration of the classifications section will enable full advantage to be taken of all the general solutions relating to classifications. For the moment, reference is made from CROMETA #2 to BRIDGE[NA] classification database when value domains are defined based on classifications.

31. As mentioned before, metadata has been collected from several sources and this process continues. There were five appointed surveys chosen as the metadata providers, but at the moment not all of them are used. Whatever could be used from the previous metadata system and other sources, has been already transferred to CROMETA #2. After the presentation in November 2005, it was decided to add more concepts necessary to compare CBS statistics with Eurostat, through the Compliance Database. The previously mentioned theme structure allows creation of another structure according to the Statistical Requirement Compendium, and measurement of the compatibility of variables used in CBS's studies to the ones defined in EU statistics. That proves that the model is flexible enough to be adapted when necessary.

V. TECHNICAL IMPLEMENTATION

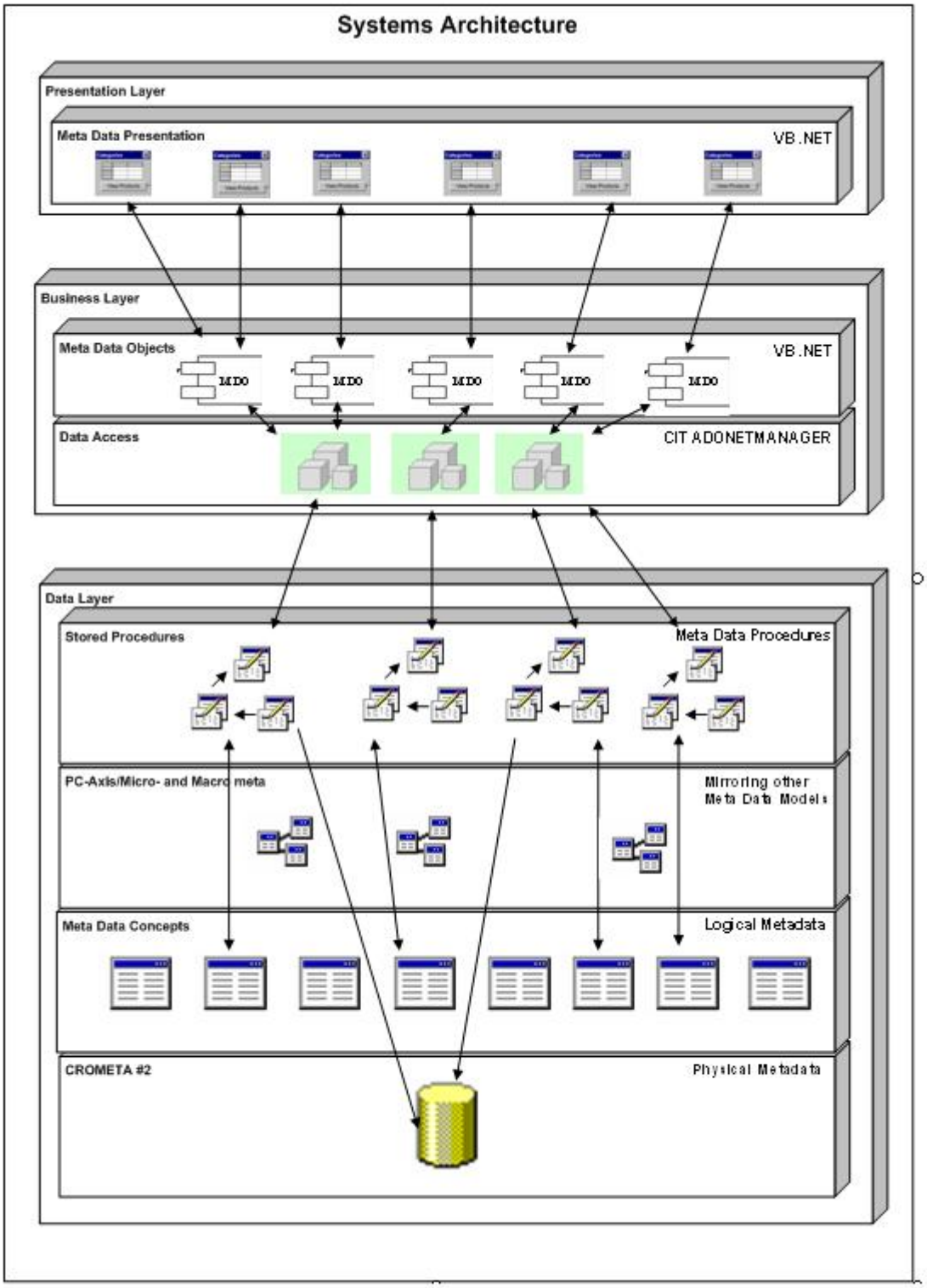
32. This section is about technical implementation because it not much is visible of the metadata model without physical implementation and an appropriate tool for browsing and maintaining metadata. The development tools used for conceptual design of metadatabase was Sybase Power Designer 9.1. The same software was used to create the physical data model that was then generated on the Microsoft SQL Server 2000 RDBMS1. All database development was carried out on the Microsoft SQL Server platform2, while CROMETA #2 maintenance tool has been developed in VB .NET3. Applications were running on Windows 2003 servers on presentation in Zagreb in November 2005. Actually, since development is still ongoing, the solution is already moved to VB .NET5 and tested on Microsoft SQL Server 2005.

33. The technical architecture follows common techniques for modern system development, fronting a multi-tiered, scalable solution, customized for a multi-user environment. It consists of three main layers (see Picture 1)

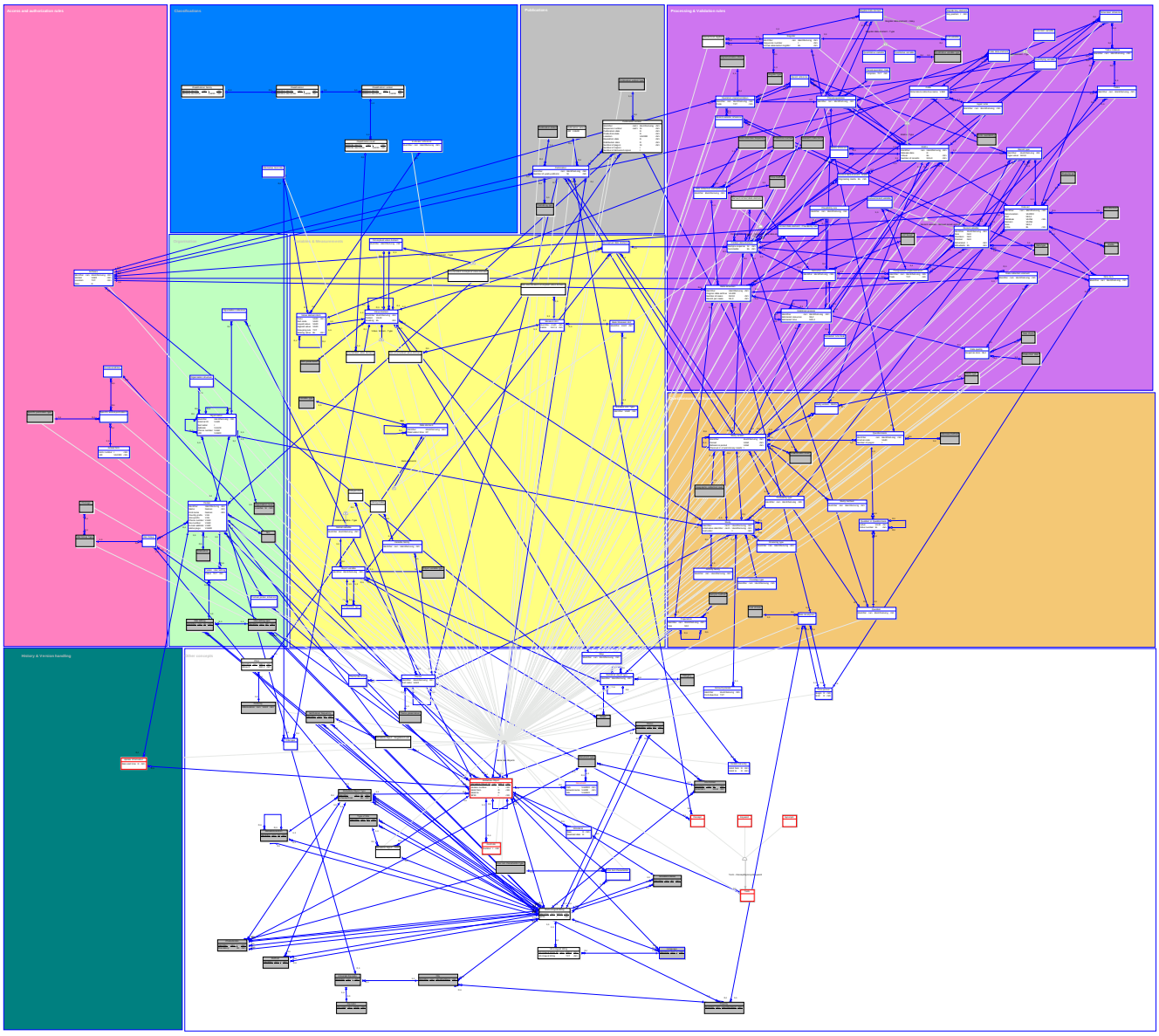
a. Data layer
b. Business layer
c. Presentation layer

34. The Data layer consists of four tiers. The physical storage tier is a normalised data model in 5th normal form (see Picture 2). It has more than 170 concepts when looking at the conceptual model and more than 230 tables in the physical database, connected through numerous relations. For faster retrieval of data, to avoid multiple joins, the logical exposition tier is defined and it is actually a denormalised data model. There is also an alternative view tier, used for mirroring additional models from other metadata consumers (like PC-Axis Macrometa), and at the end the access tier in the form of stored procedures used for managing data.

35. The Business layer is managing the business logic of the metadata solution and consists of three tiers. The lowest part is the data access tier, managing all query execution. On top of that is data communication tier responsible for receiving, formatting and forwarding requests from the top-situated business tier. The latter logically reflects the conceptual metadata model and implements metadata concepts through classes according to an object-oriented approach. This tier provides possibility of accessing metadata from other applications and it is the integration point for metadata consumers of the ISIS.



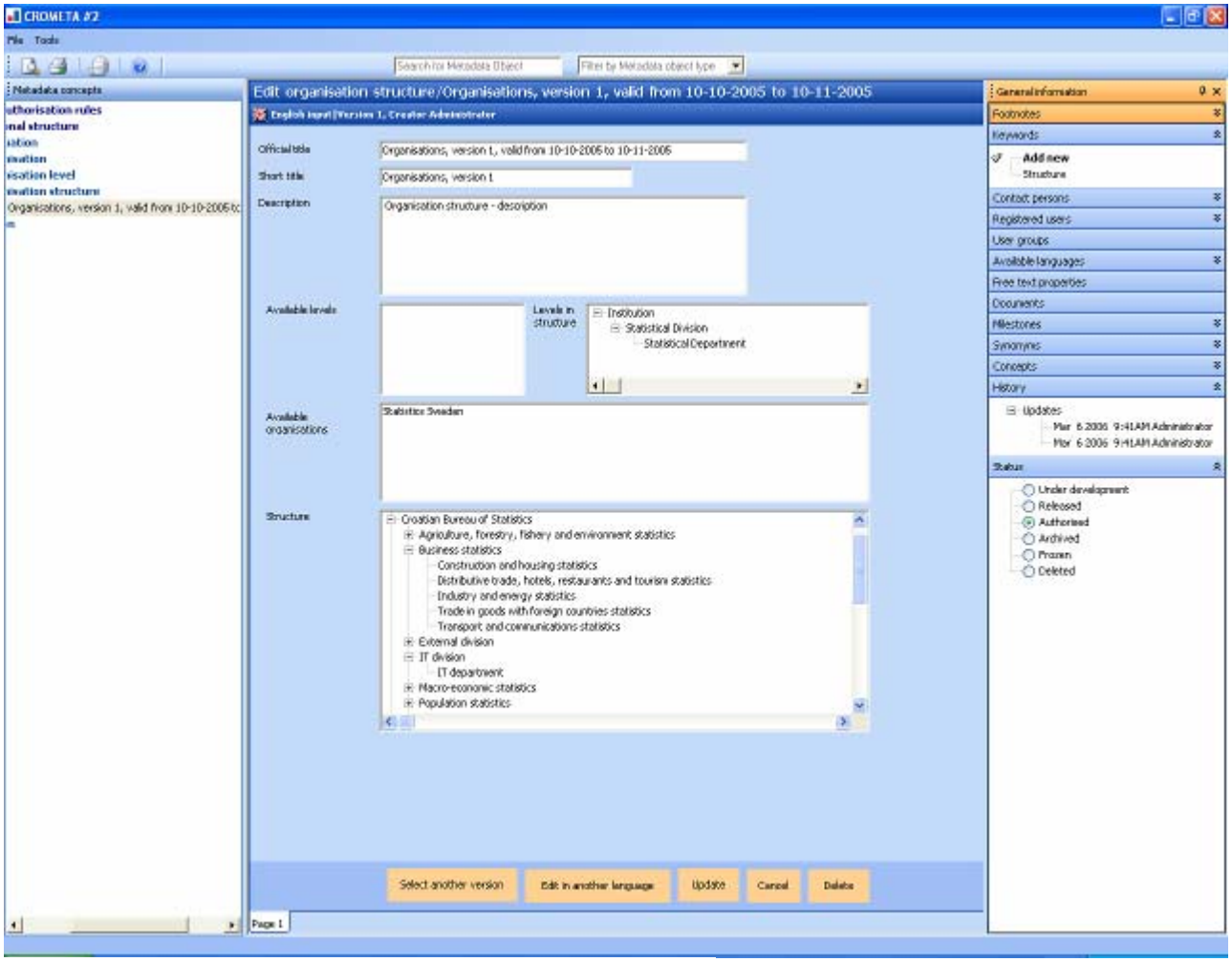
Picture 1. Systems architecture of implementation of CROMETA #2



Picture 2. Physical model of CROMETA #2

36. The presentation layer is actually the CROMETA #2 administrative and maintenance tool. It is a single user interface for all users to enter, administer or maintain metadata, when browsing or searching the metadata repository. This is a thin client solution. All processing and execution is carried out on the application and database servers.

37. The maintenance tool has been developed as a user-friendly interface to the central metadata repository. The development of the tool is ongoing and whatever could help users to navigate better through metadata is either integrated or planned to be. The application presents metadata grouped in sections and listed alphabetically. At the moment alternative views and wizards are developed for non-experts, to guide them through metadata definitions in the right order according to the metadata process. All metadata concepts can be added in many languages and the user interface itself is multilingual; default language is set according to user's choice. General methods for adding, editing and deleting data are established for all metadata concepts. All general properties are easily retrieved and changed. It is possible to search based on titles of metadata objects and by types of metadata objects, but this quick search will be replaced with the advanced one in the next versions. Users have the possibility to subscribe to metadata objects, for example, to select an object and specify the frequency of e-mail notifications about any changes on metadata object they have chosen.



Picture 3. Maintenance tool of CROMETA #2

VI. CONCLUSION

38. The development of a metadata system is a long-term process. The experience of CBS is that a metadata methodology must be strictly defined and well developed before going into technical implementation, but some questions can be raised afterwards that cause adaptations in the methodology. System development is an iterative process and many tests remain to be done before it can be put in regular production. The attempt of implementing the Reference Model [TM] resulted in establishment of the first versions of a central metadata repository in CROMETA #2 and respective maintenance tool. On the basis of the feedback from statisticians who will provide metadata, the maintenance tool will be adjusted and further developed.

39. It is the explicit target of CBS to have the first surveys processed in the automated system in 2007. That means a lot of work on the metadata system and 'generators' during 2006. The prototype of the new website integrating macro and metadata is already tested and now the bulk of data from the mainframe must be transferred to the statistical warehouse. In the future statisticians should eventually have all the tools they need for complete metadata and data maintenance and dissemination. Before that it will be necessary to include and motivate statisticians even more to accept the metadata concept and pass all their knowledge to CROMETA #2. Any misunderstanding of metadata concepts could lead to meaningless statistical data.

VII. SOURCES

(1) CBS/SCB-SIDA Project, Presentation Zagreb, Croatia, November 30th 2005, Development of a Central Metadata Repository and a Public Macro Database at the Central Bureau of Statistics of Croatia