WP. 4
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 (i): Metadata in a Corporate Context

THE STATISTICAL AND GEOSCIENTIFIC IBGE's METADATA SYSTEM

Supporting Paper

Submitted by IBGE, Brazil [1]

Abstract

1. This paper focuses on the Metadata (MetaBD) project of the Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics) - IBGE. This project started in the 1990´s and is in continuous evolution, due to users´ new demands and to changes in the Information Technology infrastructure. The MetaBD project allows an effective data administration and enables the planning and design of the surveys and censuses, supporting most of the production processes (capture, editing, imputation, tabulation and dissemination) on a centralized basis.

2. The MetaBD is actively integrated to other institutional "*ad-hoc*" data processing and dissemination systems. It describes both micro-data and data aggregates, and enables access to the official statistical publications of IBGE. Quality indicators for statistical products and processes are stored according to the recommendations contained in the Program of Statistical Cooperation for the European Union and Mercosul and Chile.

3. In addition to statistical metadata, the MetaBD database also incorporates geographical and cartographic metadata, supporting different technical areas in their need for data descriptors.

Keywords: statistical metadata; geoscientific metadata; corporate metadata systems.

I. INTRODUCTION

4. Metadata are useful to identify, locate, understand, manage and use statistical and geo-scientific data originated from surveys, censuses and projects in general elaborated by the Brazilian Institute of Geography and Statistics (IBGE), the national statistical and geoscientific agency of Brazil.

[1] Prepared by Luigino Palermo, luigino@ibge.gov.br

5. As these data sets grow in number and diversity, are produced by different departments and are made available through communication networks (e.g., LAN, internet), it is essential to describe them in a standard and centralized basis, allowing knowledge about data to be shared among technical areas.

6. The great amount of statistical data at IBGE must be described, in order to allow researchers and the Brazilian society to access it both in an organized and safe way. Besides that, geographical data are of a distinct nature and both kinds of data can be kept in many forms, varying from flat files to relational databases or geographical information systems.

7. In the beginning of the 1990's, IBGE developed an initial version of the Metadata system, basically concerned with helping the task of administration of micro-data files delivered to the Database department, usually at the end of a statistical job. With the advance and progress of IT since then, and at the same time with users demands having been changing considerably, that system has evolved to the present version, so called MetaBD (abbreviation in Portuguese for Meta-Database), with new purposes:

- To offer a centralized metadata system to the data producers areas, which can feed the different kinds of metadata along the production phases of a statistical survey or census;
- To make the MetaBD database active to the execution of homologated systems at IBGE, mostly available for processing some of the production phases;
- To become a reference to other institutional products, besides the micro-data files for example: aggregate data, dissemination medias (statistical publications, cartographic documents, CD-ROMs, etc), special collections (Natural Resources area), among others;
- To allow the retrieval of statistical metadata following varied criteria such as: survey theme (population, health, industry, etc), responsible department or geographic level of dissemination;
- Provide free-indexing mechanisms, making it easy for inexperienced users to search for metadata;
- Include metadata provided by geo-scientific production at IBGE.

8. The purpose of this paper is to present, in a concise way, some aspects of the IBGE institutional MetaBD system, and how metadata were incorporated to the daily job of the agency. These aspects will be next organized in 3 on topics: statistical context of metadata, geoscientific context of metadata and implementation issues.

II. STATISTICAL CONTEXT OF IBGE´S METADATA

A. Micro-data

9. The statistical metadata set covered by MetaBD system was the result of a study of the production phases of statistical surveys and census [Silva1997] and also from information obtained in meetings and interviews involving IBGE´s technicians.

10. Nowadays IBGE's statisticians count on facilities to store surveys' metadata into MetaBD database along with the survey production, from the planning phase to the dissemination phase (Table 1).

11. Data and respective metadata can be accessed during the whole survey's cycle of life, including Data Collection and Capture, allowing quality to be assured after each stage of production, by means of data analysis. However users are grouped in clusters, depending on the nature of the task they are allowed to do, and their actions are restricted according to the metadata status.

12. The MetaBD system also offers the storing of editing rules plans of surveys, to be used in the Editing and Imputation phase, through writing them using natural text or CriptaX language. CryptaX [Hanono1996] is a system developed in IBGE for the editing and imputation phase. In this case, when running CryptaX apuration programs, the MetaBD database is active, because it validates the editing rules, using data dictionary informations created in previous phases.

Phases	Metadata Group	Metadata Item
Planning	Survey Objective	Objective
		Main variables
		Concepts
	Resources and Deadlines	- - -
	Target Population	Geographic Coverage
		Investigation Unit
		Survey Scope
		Territorial hierarchy
	Process Type	Process Type (census, survey, administrative registers etc)
	Periodicity	Periodicity
	Survey Plan	Methodology
	Data Collection Method	Data Collection Techniques (CAPI, CASI, CATI, PAPI)
	Questionnaire Definition	Questionnaire Image
		Variable Concepts
Data Processing	Data Collection and Capture	Notes about Survey Occurrence
	Editing and Imputation	Editing and Imputation Plans
		Physical characteristics of data elements - dictionaries (conventional files) and databases
	Data Analysis	Classification
		Cross Tabulation
		Derivation Algorithm
		Quality Indicators
		Dictionaries
Data Dissemination	Data Dissemination	Publication
		Tabulation Plan
		Constraints (e.g., census secret etc.)
		Special Collections

Table 1 - Stored metadata per survey production phase at IBGE

B. Aggregate Data and Tabular Plans

13. In the same way as for micro-data, the MetaBD system can handle descriptions of the institutional aggregate datasets. The main IBGE system for on-line queries of aggregate data, available in the official IBGE web site (www.ibge.gov.br) in the internet [Figueredo2005] is the SIDRA[2] system, and it uses those aggregate data descriptions, keeping MetaBD active also during the dissemination phase.

14. More recently, this interaction between MetaBD and Sidra systems leaded to the development of the Sidra-Tabula system [Hanono2005] for generation and publication of tabular plan's tables of a survey, in a standard way, considerably reducing the time spent on tabulation process.

15. Figure 1 illustrates the communication between both systems, that allows the integration of the aggregate data production process with its dissemination process.

[2] Abbreviation for "Sistema IBGE de Recuperação Automática" (IBGE System for Automatic Retrieval)

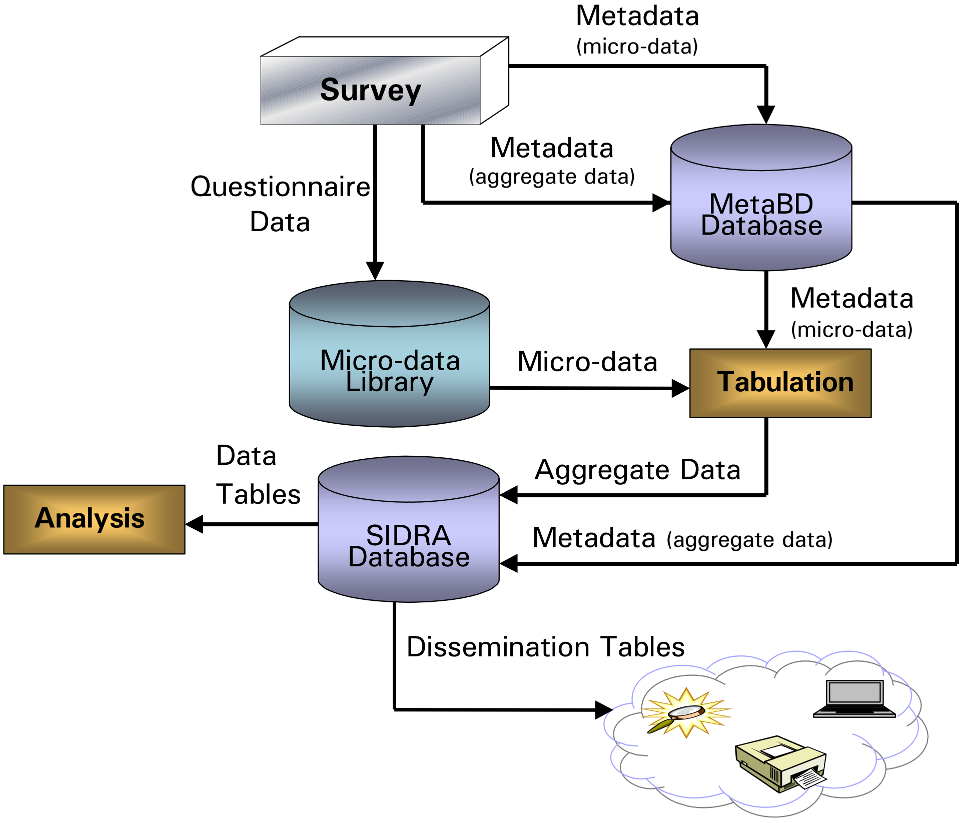

Figure 1 - Integration between SIDRA and MetaBD database

C. Quality Indicators

16. According to the directions of the Program of Statistical Cooperation of the European Union and the Mercosur, IBGE has been assessing and improving the quality of its statistical processes and products [Bianchini2004].

17. A quality methodological study was conducted by a group of experts from European Union and Latin-American agencies in 2002, which was considered strategic subject in the scope of this cooperation program. The methodological study listed nine quality dimensions that should be observed: relevance, accuracy, timeliness, punctuality, accessibility, transparency of product, comparability, coherence and exhaustiveness.

18. Matrices of quality indicators by type of statistical process (censuses, surveys and administrative registers) were created for each phase of the process, in a total of about fifty different quality indicators.

19. A standard report was specified to be elaborated at the end of each statistical process, comprising three parts: characteristics of the process, product quality and process quality. The quality indicators displayed in this written report would permit a comparison of quality between data collections, and would highlight the progress or the strengths and weaknesses in the statistical process and their products.

III. GEOSCIENTIFIC CONTEXT OF IBGE´S METADATA

20. The projects in the areas of Geography and Cartography are of a different nature: the items of the geographical and cartographic metadata depend on the theme/subject being processed, whereas in the statistical context the majority of the items are the same, with their contents varying according to the survey/census being focused.

21. This observation lead us to approach the treatment of these metadata in the new version of the MetaBD with a different strategy, where each project of the geoscientific area of IBGE is individually analyzed, and afterwards has its metadata incorporated, what in general implies in creating new metadata entities.

22. Presently, there are two projects which have been successfully incorporated to the MetaBD, both in Cartography: the 2000 Population Census Digital Municipal Maps and the Digital Index Map (which was already available in CD-ROM). In both cases the goal was to extend and to make more democratic the access to the information contents of IBGE maps.

23. At the moment, the first project in the field of Geography is been implemented, more specifically in the theme of Natural Resources, describing IBGE's accumulated intellectual patrimony about vegetal and animal species - some of them threatened of extinction - kept in an ecological reserve belonging to the institution and located in the Brazilian savanna in the Center-West Region of Brazil.

IV. IMPLEMENTATION

24. The system implementation was performed by analysts and programmers of the MetaBD project. In a few occasions, they counted with the orientation of external consultants, especially when a new technology was being used.

A. MetaBD Database

25. The conceptual data model of the new MetaBD system was conceived after several meetings gathering researchers and database experts of IBGE; from the results of these meetings, emerged the first design of the MetaBD's database, whose macro aspects can be observed in figure 2.

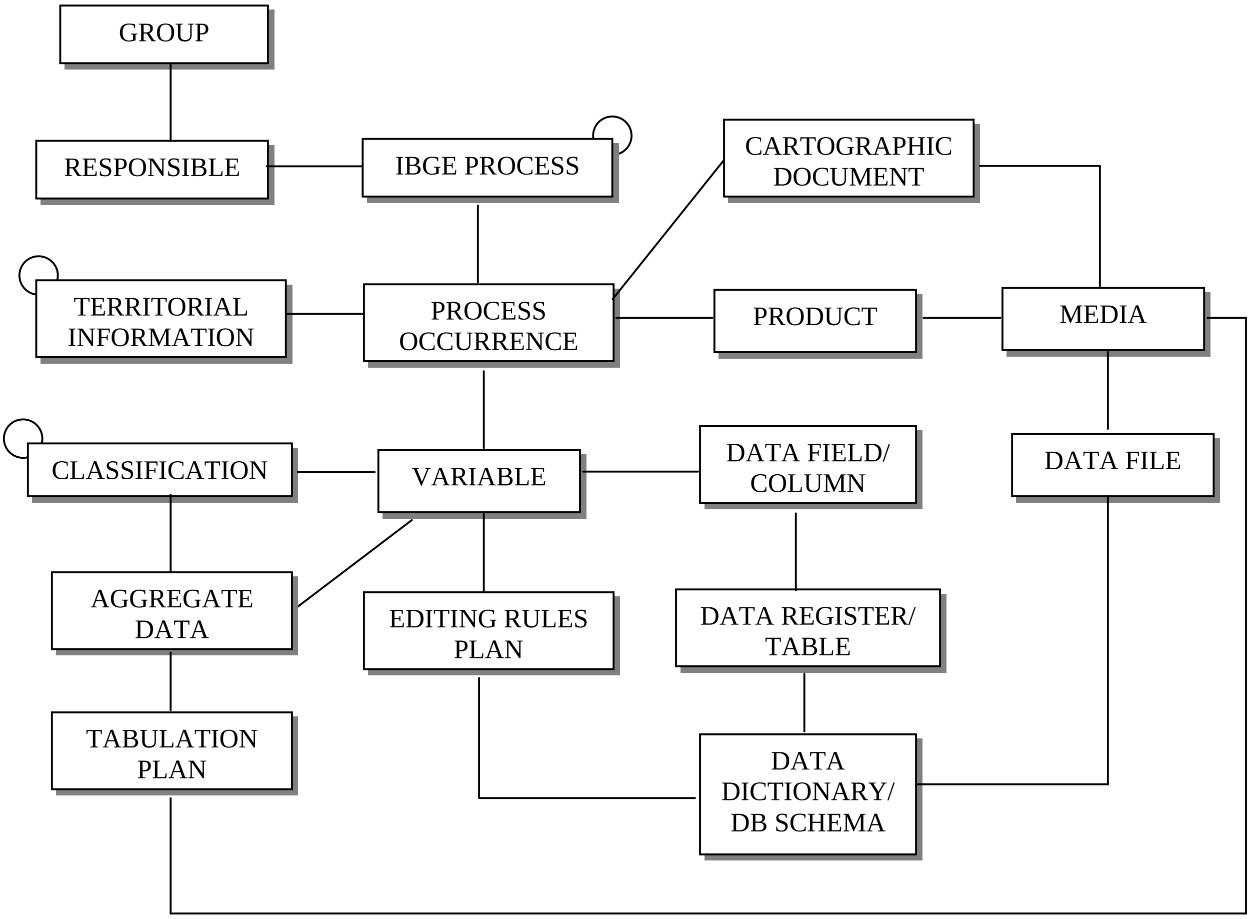

Figure 2 - MetaBD Main Entities

26. The MetaBD database became operational in January of 2000, but its data model still keeps evolving up today, getting new extensions to cover nonexistent contents. Among the most recent extensions, we can highlight the geoscientific metadata for cartographic documents, the tabular plans' metadata, and the quality

indicators' metadata. For this last, a data model extension (figure 3) was conceived, updating some previous entities and creating others, and of course adapting metadata access applications, which will be commented further on this paper.

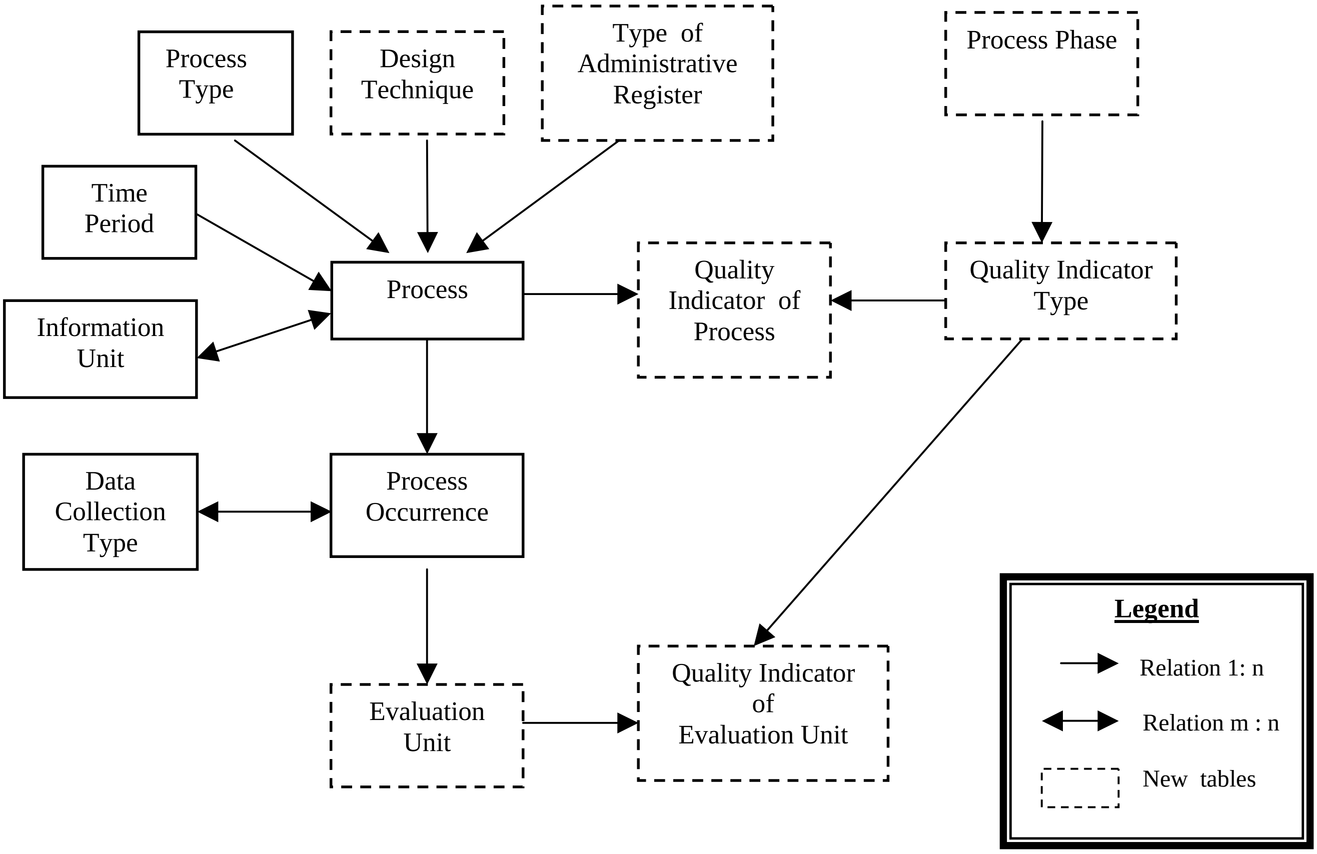

Figure 3 - MetaBD Data Model Changes for Storing Quality Indicators

27. At present, the MetaBD scheme is logically divided into following sub-schema, exclusively for management reasons[3]:

a. IBGE Collection processes (censuses, surveys etc.) and Geoscientific projects - basic metadata that is not time dependent. Examples: name, acronym, periodicity, unit of observation, geographical coverage, background, methodology etc. (12 tables);

b. Statistical or geoscientific events - metadata about surveys, censuses, and others but specific on their time occurrences, including variable concepts and pertaining documents (17 tables);

c. Questionnaire - statistical metadata (9 tables);

d. Editing and Derivation - metadata of editing rules plans and algorithms of derived variables (11 tables);

e. Dictionary or Schema - metadata about dictionaries of micro-data files and about relational databases (23 tables);

f. Territorial - metadata of geographical levels and items (8 tables);

g. Aggregate - metadata of aggregate data (21 tables);

h. Publication - metadata of tabular plans (6 tables);

i. Cartographic - metadata of cartographic documents (33 tables);

[3] The number of tables enclosed in parentheses includes the entities themselves, their relationships and auxiliary control tables.

j. Results - metadata of products (15 tables);

k. Media storage - products' metadata (21 tables);

l. Quality - metadata of the quality of the statistical processes and products (8 tables);

m. Safety - metadata for users, groups, and their authorizations (12 tables).

28. Following the data processing orientations of the institution, the relational ORACLE DBMS was chosen, which version nowadays ratified is Oracle 9i. The Oracle Designer tool has made database administration easier, for tasks such as the chartering of sub-schema and the semi-automatic generation of SQL scripts - to create about 200 tables and their indices.

29. Considering the exposition of IBGE's metadata to foreign institutes and partners, and foreseeing the availability of the MetaBD in the Internet, there was a concern in storing name and description textual fields in another two languages: English and Spanish. Description fields are stored as Oracle CLOB (i.e., Character Long Object) objects.

B. Metadata Access

30. The MetaBD system offers two applications for accessing metadata, available in the Windows environment: one for metadata loading and updating, and another for querying metadata.

31. The metadata loading and updating application is of the type Client/Server, developed in Visual Basic, which demands an initial set up, besides the needed basic software: Oracle Client and some ODBC driver for Oracle. Its use is restricted to the technical staff members - responsible for the data production, loading and updating the metadata - and to IBGE's Data Administration Team.

32. The Data Administration Team has access to all metadata and is in charge of defining authorizations based on each users group's profile and on sets of operations allowed to that group over MetaBD tables. Profiles and sets of operations are implemented using Oracle roles, in such a way that this safety mechanism is naturally supported by the DBMS.

33. The menu structure of the application is organized according to metadata context: statistical, geoscientific and administrative[4]. Menu options that do not apply to a specific user profile are not displayed. Thus, when specifying a user of the loading and updating application, it is necessary to inform the surveys under his/her responsibility, as well as the group to which he/she belongs (or his/her profile, to be provided in a new group, in the case he/she does not fit any group).

34. Concerning statistical metadata, it is also possible to store images (e.g., the survey's logo, the survey's questionnaire) or still to store links pointing to documentation files written in some textual editing application like MS-Word.

35. The MetaBD database can be viewed through a user-friendly web application, so far only for internal users. The decision-making process to make it available in the Internet depends on institutional policies and includes subject matters such as the definition and quality of the items to be released.

36. The querying application was entirely built on the MS.Net platform, using MS Windows 2000 Server as operating system and IIS as web server software, and can be reached starting from a browser (like Internet Explorer or another) and does not require any software component to be set up.

37. When accessing the querying application's home page, the first screen to be displayed brings information on recent data and metadata incorporated to the institutional databases. The MetaBD project team also uses this space to inform about implementation of new application facilities.

[4] In general this includes auxiliary tables, for instance: pre-defined attribute values.

38. Among the most attractive facilities of the querying application is the possibility of producing files containing the dictionary layout in different formats/languages, to be used off-line by other data processing and analysis tools, like SAS [Palermo2003], REDATAM and CriptaX.

39. The querying application also offers some reports about MetaBD database entities: Survey and related Variables, Data Dictionaries, Classifications, Quality Indicators.

C. Hardware Infra-structure

40. The hardware infrastructure of the MetaBD system can be seen in figure 4. Database and web servers of the MetaBD system share same equipment of those of the SIDRA system, and are based on Intel Pentium Xeon Dual-processors machines.

41. A fast cross-over channel connects the database and the web servers, streamlining the metadata access for queries made via web. The database server is connected to the local area network of IBGE for execution of both access applications, but limited only to internal use (corporate network). Presently only one external Brazilian organism, belonging to the planning sector of the national government is able to query the MetaBD database over the internet.

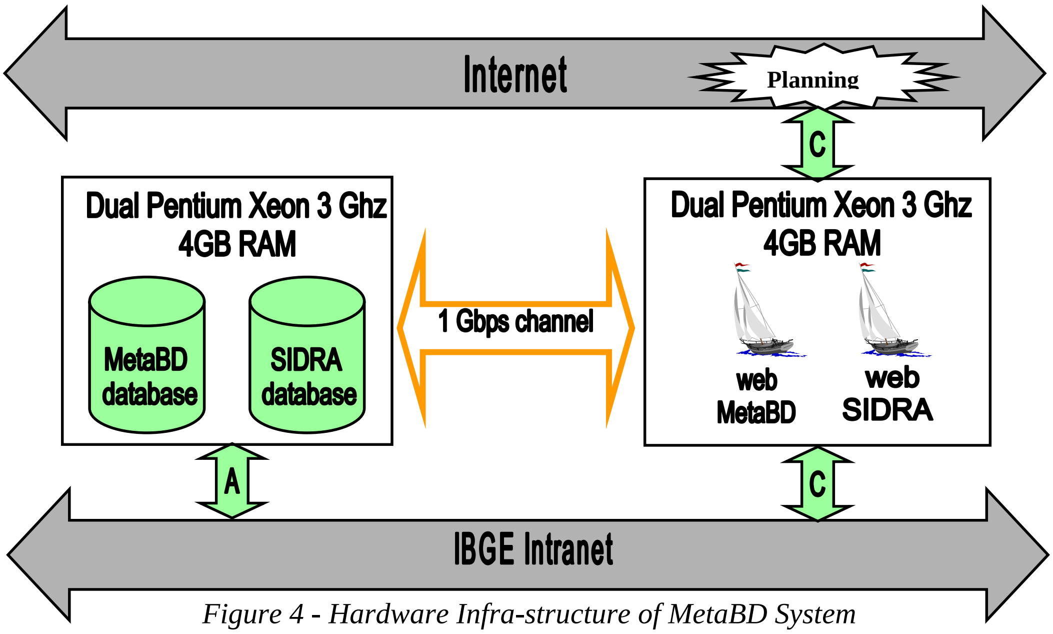

Figure 4 - Hardware Infra-structure of MetaBD System

D. Foreseen Changes and Improvements

a. Improvement of the descriptors' quality (e.g., reducing conflict of concepts);

b. Development of a new 3-thier levels version of the loading and updating application, suitable for web use;

c. Design of a better web interface for the metadata querying application, then allowing internet access to MetaBD database, and perhaps some adjustments for supporting less popular browsers;

d. Release of a sub-set of geoscientific metadata to other national geographic agencies, according to the Federal Geographic Data Committee - FGDC, to meet staff members demands;

e. Take into consideration free-software options, since the Brazilian government is moving towards this direction. The use of PostGreSQL database management system is under study;

f. Integration into new editing/imputation/analysis tools.

V. CONCLUSIONS

42. In spite of IBGE’s limited resources, the most relevant motivation of the institutional metadata system has always been to transform data into information - for the benefit of staff members not involved in data production and of society in general. This emphasizes IBGE’s institutional mission which is “to portray Brazil with the necessary information to the knowledge of its reality and to the practice of citizenship.”.

43. To achieve this goal, the Data Administration Team and the staff members responsible for IBGE’s data production have been documenting the data, adding - with the help of the MetaBD system - its semantic metadata (concepts, background, methodology, sources, geographical coverage etc.), syntactic metadata (micro/macro data fields’, registers’ and files’ descriptions, relational databases’ descriptions etc.) and pragmatic metadata (editing rules plans, tabular plans, derivation algorithms etc.), without which those data would be either incomplete or useless.

44. In the future, the challenge to overcome is the appropriateness of the MetaBD system to metadata patterns, a subject that was not taken into account when the system was designed.

45. Presently the MetaBD database stores, in statistical metadata, approximately 117 surveys and censuses (being almost 6,000 time events of these collection processes), over 50,000 variables described in 824 dictionaries and stored in 2,764 physical files.

46. In the geoscientific area, the numbers are also expressive, considering that only two cartographic projects were carried out up to now, in a total of 12,363 stored maps in 8,194 physical files.

VI. REFERENCES

[Bianchini2004] Bianchini, Z.M. & Albieri, S. - “A prototype of the Quality Indicators System at the Brazilian Institute of Geography and Statistics”. *European Conference on Quality and Methodology in Official Statistics,* Mainz, Germany, 2004.

[Figueredo2005] Figueredo, L.A.G.A. & Masello, J. - “SIDRA - Banco de Dados Agregados - Arquivo de Dados Agregados - Definição e Carga” (SIDRA - Aggregate Database - Definition and Loading). *Technical Note 01/05, Diretoria de Informática, IBGE,* Rio de Janeiro, Brazil, 2005.

[Hanono2005] Hanono, R.M Figueredo, L.A.G.A. and Masello, J. - “Processo de Tabulação SIDRA-Tabula” (SIDRA-Tabula Tabulation Process). *Technical Note 02/05, Diretoria de Informática, IBGE,* Rio de Janeiro, Brazil, 2005.

[Hanono 1996] Hanono, R.M. and Barbosa, D.M.R. - “CRIPTAX - A Generalized Editing Application Generator”, *Work Session on Statistical Data Editing*, Voorburg, Netherlands, 1996.

[Palermo2003] Palermo, L.I. and Masello, J. - “Metadados e Geração Automática de Programas” (Metadata and Code Automatic Generation). *Work presented in the Brazilian Annual SAS Users Group Meetin*g, São Paulo, Brasil, 2003.

[Silva1997] Silva, A.F. - “Aporte da Ciência da Informação à Organização de Metadados para Acervos Estatísticos” (Contribution of the Information Science to the Metadata Organization for Statistical Data Libraries). *MSc Dissertation, Pos-Graduation Course in Information Science, UFRJ/ECO and CNPQ/IBICT,* Rio de Janeiro, Brazil, 1997.