

REFERENTIAL DIRECTORIES AS METADATA DISSEMINATION

Mauro Sergio S. Cabral, IBGE Foundation

ABSTRACT

We get statistical information through surveys or administrative records. It is usual to get a lot of information without integration and comparison among them. This has been happening because the information producers are not well known. The information providers do not know exactly what they are producing each other. Meanwhile, the new information technologies are looking for standards to facilitate the communication and the interaction between producers and consumers of information. It is necessary to go towards the tendency of this technology in order to have a more democratic dissemination of information.

On the other hand, plenty of information normally is critical to the consumers and generally it is not produced on time and its cost is enormous. The more information producers know each other more quality is incorporated into the produced information and the consumers get the desired information more easily. Besides, the more information producers interact each other faster the information is produced and lower is the cost. Therefore, it is necessary to look for alternatives to organize the information according to the trend of the information technology.

KEYWORDS

Data administration, Information consumers, Information producers, Information network, Internet, Metadata, Referential systems, System coordination

1. INTRODUCTION

We have an enormous difficulty for comparing similar surveys from different statistical information producers. Generally, each survey is produced using a proper methodology and the users of the generated information do not know the correspondent methodologies used. So, we need to know this kind of information about the data. In other words, we need to know a sort of metadata.

Sometimes we need to get the data from one survey to use in another. So, we need to know some characteristics about the data. In other words, we need to know another sort of metadata. More and more we have to know the data about the data. We are speaking about metadata. Nowadays we need to include the metadata into the available information. People always need a lot of information about the data before to get them. Moreover, there is similar information spread over several places. It is necessary to choose the best information. Sometimes we can use one set of information into a survey but we cannot use it into another. So, we need to find out the more adjustable data when we have a collection of similar data. Metadata are vital if we want to offer a quality service in a society whose information needs are growing all the time. Therefore, it is very important to organize all the available metadata.

It seems to be a good idea to have referential directories associating the correspondent metadata to the data in order to solve the main problem. The problem, of course, is to look for the desired information. These directories can be classified by thematic areas (e.g. education, transportation,...) and by levels like federal, municipal, and so on. The contents of these directories are exclusively data descriptions although it is also possible to get the wanted data through referential systems. Referential systems could have links to several directories and their correspondent data files. This idea is based on the fact that many statistical and geoscience agencies are now engaged in the development of metadata systems. The referential directory is an important tool for that kind of system.

The first important step to search for an information is to find the possible places where it is stored. The second step is to verify how the correspondent data are organized. Finally, the last step is to know how to get the desired data. The referential directories must satisfy the users with these three steps even though it is not necessary to get the desired data directly through the system that controls the correspondent directory.

The implementation of these metadata directories must be done through homogeneous areas where the concepts and procedures are close to a standard. Several directories from different areas could be linked to provide a new level of information. The approach to create these directories is presented in a conceptual manner and the first steps to develop a kind of referential directory in IBGE Foundation is presented as a practical experience.

2. A CONCEPTUAL APPROACH

There are different manners to face the problem we have to reorganize the information in referential directories.

However, we can choose some common aspects to several approaches. On the basis of these aspects we can describe the main steps as follows:

2.1 The Environment Definition

Firstly, it is necessary to know where we want to reach. This does not mean the information only belongs to the environment that we are studying. In fact, the information as a producer of knowledge must transpose the frontiers. What we are searching for delimitation to the environment of the information is to find out the data producers and data consumers' areas. The first ones are responsible for data updating while the last ones are very important in the dissemination process. Studying this environment, the perception of these areas and the relationships among them must be recognized in order to understand the most important data flows. The more systematic these dataflows are, the more they could be analyzed in order to understand the information into that environment.

Secondly, we must verify the level of homogeneity that belongs to the chosen environment in relation to the flows of information. It is the opportunity to revise concepts and adopted standards. This is very useful because the information needs to have the same shape to be well understood.

Finally, before defining directories and referential systems, we must study the origin and destiny in each cycle of the information perceived in the studied environment. It is very important to verify where we have lot of activities adding information to the information and where few activities are modifying the information.

While we are studying the delimited environment it is possible to notice the technological infrastructure. Having in mind that referential systems need a minimum of functionality, it is necessary to have conditions to set up good flows of information between each origin and destiny points. The more technological standards exist in the studied environment the less "interfaces" and "front-ends" are necessary to set up a communication among the several points of the environment.

2.2 The necessary tools

All the information, which goes from an area to another, has characteristics from its environment. Technical information is different from bibliographical information that is different from statistical information and so on. Therefore it is necessary to recognize very well the characteristics of the information.

When we recognize the information's characteristics we can point out the main aspects for the users of the information. We have to think about the users who only need to know the information and the users who also need to modify the information. We can distinguish two good tools in both cases. Where we need to deepen into information to produce other information normally we need to have more data to produce more information or more data. In this case we need a structure of a metadata bank. Where we only need to search for an information to be catalogued or compared to similar information usually we want to look up the information spread out several files that belong to the correspondent environment. In this case we need a structure of referential bank. In both cases, after we have defined the content (metadata or metainformation) of these databases we must adopt a pattern to register this information. The pattern is very important to facilitate the interaction among the different systems that could access these databases.

In many cases, these two structures of data bases can be joined into only one that we call Referential Directory. In this case, this data structure details the elements of the information to the users who need data operations. Besides, this data structure also presents the information to the users who only need to know the elements of the information. This is a more comprehensive tool to be used.

2.3 The information network

A referential directory spreads the information's knowledge in the environment where the dataflows occur. However, this is not enough. These dataflows are very dynamic and always require a good updating level. The advances of the telecommunications technologies are contributing to this dynamism more and more. Likewise we use the database technologies to create the referential directories we must use the telecommunications technologies to link these directories to the several users.

The definition of these networks must distinguish the areas that transform the information (data operation) and the areas where the information is just requested. Comparing to the "client/server" schema, the first ones would be the metadata servers and the last ones would be the clients. So, in this schema the servers have a very important role because they have to update the metadata into the referential databases. The updating is fundamental in the information dissemination process. When this updating does not occur we can cause delay to the knowledge of the

involved areas and consequently wrong information can be caught by the users. Therefore, these metainformation or metadata flows need to be very fast using all the available resources of the set up networks.

2.4 The referential systems

The structure defined in relation to the storage tools and telecommunications network, involving all the areas in the environment with their correspondent dataflows, would represent the backbone for the referential schema. This can be understood as the body of the referential system. It would be lacking the soul of this system. The soul would be the applications programs, which would control all the system users' necessities. The set of these programs we could call as referential systems. These systems would take care of metadata servers (operating and transforming data) and the metadata clients (presenting the information like browsers). These systems would be adjusted for each kind of user.

Another important characteristic of these systems is to be adjusted to the environment. It is necessary a constant validation of: the metadata incorporated into the referential directories; the database tools used in the process; the dataflows across the network; the areas involved in the environment; and consequently the systems that support these environments. The referential systems will only present useful information if any change in the environment can be detected and adjusted. If this does not occur, the knowledge of the information will be damaged. So, it is very important to have some criterion to evaluate the referential systems adjustment.

2.5 The referential systems coordination

The proposal of referential systems is based on a decentralized system working in a client/server schema. However, it is necessary the role of a coordinate agent. This agent would control the use of the system by the clients and mainly by the servers in relation to the metadata included into the referential directories.

All the system can be affected becoming out of date, inefficient, without this kind of agent. This can provoke the abandonment of the users. However, the coordinator must not be the metadata administrator of the system. Each server area (information provider) is the responsible for the metadata to be stored into its referential directory. The coordinator can define some standards to be followed but must not create difficulties to the dataflows among all the components of the system. The most important function of the coordinator is to ensure the functionality of the system and the quality of its information.

The importance of this coordination is to verify the agents and the sort of information that have a huge demand and those that do not have any kind of demand. Based on these kinds of observations many decisions can be taken to get better the referential systems. Moreover, these observations can allow us to obtain better data transfer between two areas in the referential systems, minimizing time and costs.

In general, the choice of the coordination in statistical systems will always be the agent that has the necessity to disseminate statistics and deal with this sort of organizations. Normally this agent is already some kind of coordinator in the statistical production and dissemination process.

2.6 The culture among the areas

Besides the role of the coordinator what is more important in the referential systems is the culture set up among the areas. It is worthless to have an excellent referential system in relation to the technological and methodological point of view if users do not exist. Therefore, it is necessary to observe the interaction among the several areas (clients or servers) linked to the referential systems.

In fact, it is fundamental that all the areas linked to the information network use the systems. The only way to improve the referential system is the constant utilization by the several agents of the systems.

This is the most difficult part to reach the success in the implementation of the referential directory involving all the life cycle of a survey.

3. A PRACTICAL APPROACH

IBGE is the Brazilian Statistical System Coordinator. Besides statistics data, IBGE is also responsible for the geoscience area, including geodetic, cartography, geography and natural resources of the country. Therefore it has an enormous collection of data to administer. When the results of the censuses and surveys are published the majority of these data and metadata are stored in what is called IBGE-Database (IBGE's institutional Archive). The information

dissemination process is widely based on the organization of the IBGE's Archive. We could claim that, the more organized the IBGE's Archive is, the more efficiently will society demands be attended.

The proposal of the referential directory is an alternative to improve the information dissemination process in IBGE. In the following items it is presented what has been done in IBGE to create a referential directory.

3.1 The environment definition

The delimitation of the environment where the information has a good level of homogeneity was based on the operational structure of the IBGE. So, it is distinguished three areas: the geoscience area, the statistical area and the dissemination area. After this identification some points of consumption and production of information were identified in relation to their correspondent links. The operational structure also served to find out the points in each area. Having knowledge of these points, it was possible to draw the main information flows (origin and destiny points). Moreover, it was possible to identify some important data files, concepts and standards in each environment. In this phase, it was possible to recognize some producer areas (performing data operations) and some consumer areas (only requesting data) although the information dataflows have not been deeply analyzed. This will be very important to choose the servers and the clients in the "client/server" schema in the network definition phase.

The information mapping, in the first version of the referential directory, was just defined to draw the inside institutional dataflows. So, the links that IBGE has to other institutions will be drawn in the future.

3.2 The tools definition

In relation to the creation of the necessary tools for the intended referential system, the IBGE already went towards this target. Few years ago it was developed a database containing the metadata for its statistical surveys. However, the actual version of this database, called Metadata Bank, it is very much limited and it was developed to be managed by a centralized system in an IBM mainframe platform. This version only deals with statistical data stored in magnetic data files.

The project of IBGE's referential directory expands the thematic areas mapping the geoscience areas and the information dissemination areas too. Moreover, this project will allow any kind of media (magnetic data file, CD-ROM, publishing,...) be catalogued.

The simplified data model of this new Metadata Bank presents several entities, as follows:

1- Alphanumeric file

It describes the alphanumeric files generated by the surveys and incorporated in the IBGE's Archive (institutional database).

Some attributes: Identification, name, format file, type of the file, description...

2- Referential file

It presents the databases and files generated by the surveys and not incorporated in the IBGE's Archive (institutional database).

Some attributes: Identification, name, responsible area, description...

3- Classification

It describes each category of the variables used by the surveys.

Some attributes: Identification, name, quantity of categories, the provider of the classification, last updating, description...

4- Category

It defines each category associated to the correspondent classification.

Some attributes: Identification, name and description.

5- Publishing description

It presents a set of elements to identify the collection of books and maps.

Some attributes: Identification, title, author, ISSN/ISBN, responsible area, description..

6- Data dictionary

It describes the format of the files generated by the surveys and incorporated in the IBGE's

Archive(institutional database).

Some attributes: Identification, name, format file, situation (available or not), responsible user, description,...

7- Element of the information

It describes the elements of the information. These elements have definition/description of the information units considered important for dissemination.

Some attributes: Identification, name and description.

8- Institution

It identifies the survey's producers.

Some attributes: Identification, name and description.

9- Physical item

It describes each field of each register. The physical item corresponds to the information unit.

Some attributes: Identification, format, range, physical position, invalid representation, type of classification (if a categorized item), ...

10- Territorial level

It describes the territorial structure's levels

Some attributes: Identification, code, name, description...

11- Survey occurrence

It describes the period of each survey.

Some attributes: Identification, name, period, description...

12- Publishing occurrence

It presents the essential elements for identification of books and maps.

Some attributes: Identification, title, author, edition, responsible areas....

13- Survey

It describes the surveys and projects produced by IBGE or another institution.

Some attributes: Identification, name, period, types of dissemination, responsible area, situation(available or not), description...

14- Register

It describes each one of the different data collections for each data dictionary

Some attributes: Identification, name, length, description...

15- Thematic and cartographic representation

It describes all the thematic and cartographic documents available in IBGE in the geoscience area.

Some attributes: Identification, international code, title, scale, responsible area, representation type, description...

16- Specific theme

It describes the basic themes used in cartographic and thematic representations in IBGE's geoscience area.

Some attributes: Identification. specific theme and description

17- General Theme

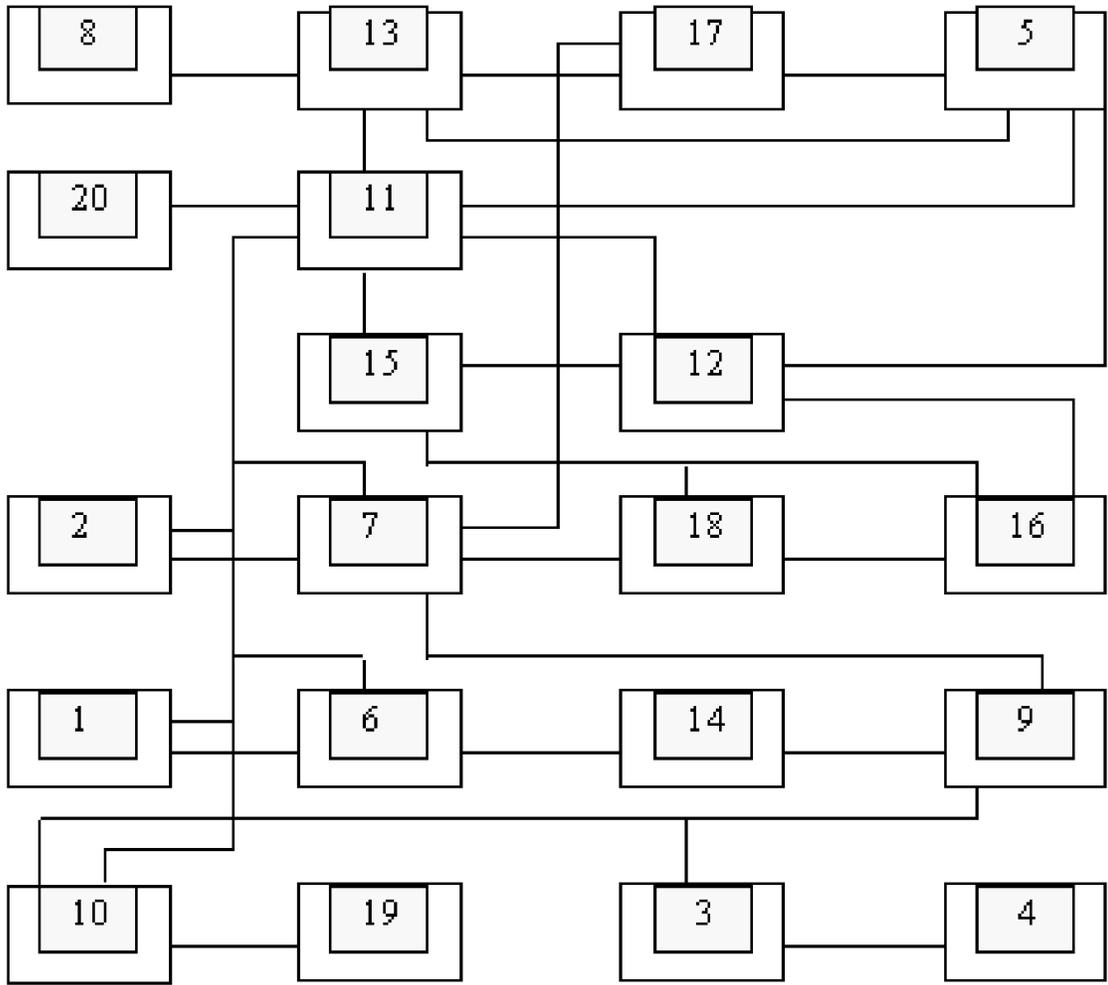
It describes the major themes studied in IBGE

Some attributes: Identification, theme and description.

18- Sort of information

It describes the several kinds of information used to thematic and cartographic representation

Some attributes: Identification, categories' names, description



19- Territorial unit

It defines each one of the territorial units for each territorial level

Some attributes: Identification, code, name and description.

20- Folder

It identifies folders containing complementary metadata when necessary and are not incorporated in the IBGE's Archive (institutional database).

Some attributes: Identification, name, description, ...

We can represent this data model as follows:

We think IBGE will only have one referential tool. A system will be developed to allow the data accessing to several servers and clients inside the areas of the IBGE. In the future, when the external dataflows are included probably it will be necessary to build a more referential tool to support a larger demand.

3.3 The information network definition

At this moment IBGE is being transformed from a centralized structure based on an IBM-mainframe to a decentralized structure based on a "client/server" schema. This is happening in its headquarters in Rio de Janeiro but the network will be spread over the several offices set up in each state of Brazil. This large network will be very useful to support the referential directory of IBGE.

In the first version, all the areas of IBGE spread all over Brazil could access the information from the referential directories. So we would have some servers (areas with data operation) and several clients (only requesting data) in many points of the network. However, it is quite sure that one bigger server containing all the institution's metadata will be necessary.

When this version is completed, it will be observed how to use the referential directory improving the information dissemination process. IBGE is linked to other public organizations through a large federal network called SERPRO (federal organism for income tax data

processing). Using this network IBGE is accessed by several ministries and correspondent offices. So, the access to the referential directory could be reached by a large group of users through this federal network. Likewise, IBGE is linked to a public network called RENPAC (packets national network - X25 protocol). Using this network an enormous quantity of people could also access the referential directory in IBGE. Moreover, IBGE is linked to the INTERNET as an information provider. Therefore, it will be possible to have the referential directory's information through the Internet to the society disposal.

Based on these telecommunications networks it will be possible to expand the defined environment for the referential directory. The first version of the environment is only delimited over the IBGE areas. In the future the environment can be mapped including other statistical organisms. It means that other servers and clients areas can join to this model allowing the improvement of these dataflows among several of these areas represented on this larger schema.

This is very important to IBGE because it is always necessary to receive data from surveys or administrative records produced by many other statistical offices. In this situation, the knowledge of the metadata through these larger referential directories would be possible by electronic transfer (using EDI patterns, using FTP protocol, ...) between IBGE and any other server in the network. This could be very important to minimize time and costs to produce surveys in IBGE. We can represent the IBGE network as follows:

IBGE NETWORK

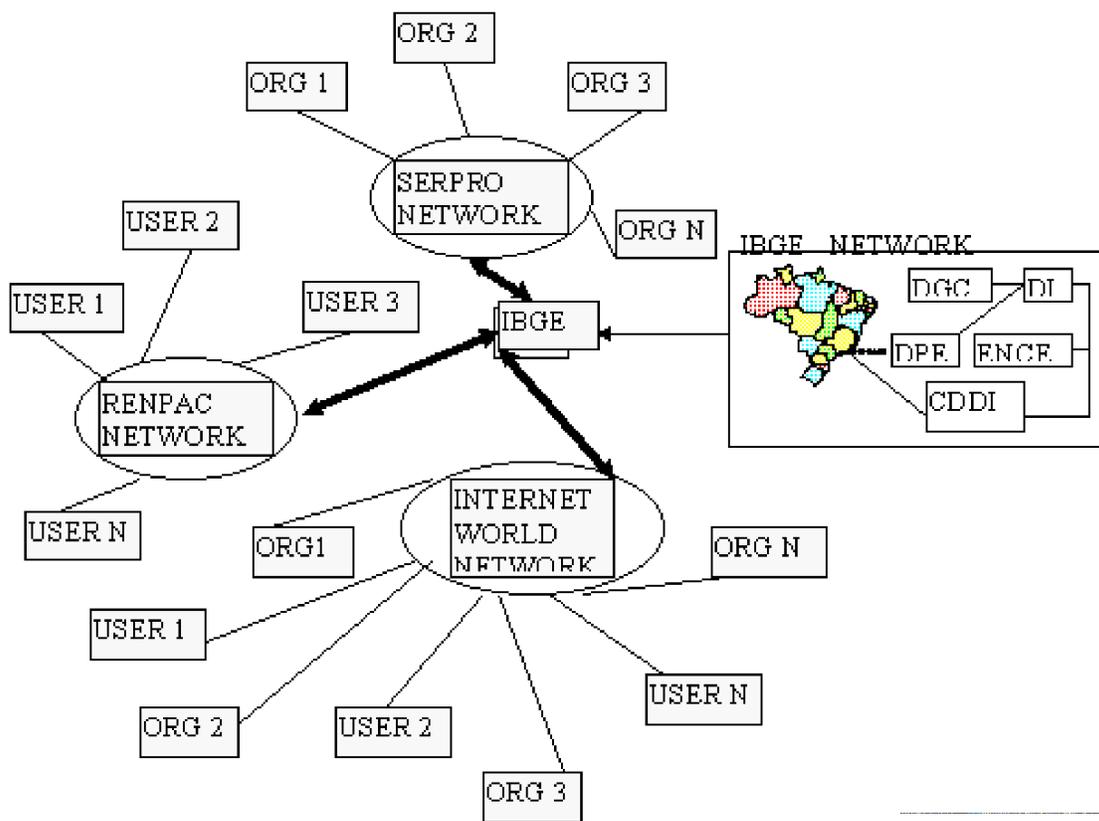


Inside the IBGE network, which is spread all over Brazil, we can point out the main areas located in Rio de Janeiro:

DGC - Diretoria de Geociências (where surveys on geodetic, cartography, geography and natural resources are produced)

DPE - Diretoria de Pesquisas (where all the censuses and several statistical surveys are produced)

DI - Diretoria de Informática (where all the informatic and telecommunications support are provided as well as some systems developing requested by other areas)



ENCE - Escola Nacional de Ciências Estatísticas (where students area graduated in courses of geodetic and cartography, statistics, and some post-graduation courses are provided)

CDDI - Centro de Documentação e Disseminação de Informações (where all the information produced by IBGE or other associate organism are put at the society disposal)

3.4 Referential systems

As said before, in the conceptual part of this paper, the referential systems need to distinguish the producer areas from the consumer areas. A user can be a producer and a consumer at the same time although normally each area can be classified in only one kind. This classification is very important because the applications of the referential system must observe these kinds of services. Likewise, these systems must verify the information updating because if this does not occur the referential directory becomes less important to the users. IBGE does not have the necessary experience to develop such systems yet. The system developed to support the actual Metadata Bank does not comply these premises.

The idea is to use the "client/server" structure, which is being installed in IBGE, to develop the applications oriented to information producer areas (servers) and also develop applications oriented to information consumer areas (clients). So, it is intended to have copies of the parts of the referential directory in many points of the network where the metadata are updated. It is also thought to have one bigger server containing all the available metadata for information dissemination.

This decentralized structure is very important to allow the providers areas for updating directly theirs metadata into the correspondent copy of the referential directory. We have been learning day by day that it is almost impossible to choose an administrative area to be responsible for this updating. It is extremely necessary that each provider area be responsible for its own data administration.

3.5 The referential system coordination

The coordination of the referential system probably will be the dissemination area - CDDI (Information Documentation and Dissemination Center). This is the area that is responsible for statistical and geoscience information dissemination to all the society. This choice is very important because when the internal network is established (completing the version of this first environment) the CDDI will be the best way to expand the delimited environment involving other external public institutions and important sectors of society. The CDDI deals with these flows of information and has some knowledge about what kind of information is needed from IBGE to these external institutions as well as the reverse flows. This experience will be fundamental to create a larger information referential system.

Another important characteristic of CDDI is its responsibility in developing standards for concepts and procedures using adequate documentation. This experience is extremely important for the index process that must be implemented into the referential system. This process is very useful to help the users to search for the desired information in the referential directory.

3.6 The culture among the areas

When the first Metadata Bank was put at users' disposal this experience showed us that it is not sufficient to explain how to use properly the system. It is necessary to show them that the tool (in our case the referential directory) needs to be incorporated to the survey's life cycle. Having this in mind, it will be extremely useful to involve all the interested areas, from the first to the last step to implement the referential system of IBGE. The idea is that the referential directory has to be used by all the involved areas as a tool for institutional data administration. Moreover, the big secret to facilitate the knowledge among the users is that the referential system must be very easy and friendly enough.

4. CONCLUSION. and the Internet ?.

We think that is possible to get a referential directory if we follow the steps described in the conceptual part of this paper. So, we have to define the environment, to choose the main databases to be catalogued, to draw the network backbone, to define the referential systems and to choose the appropriate coordinator. Some investment will be necessary in the telecommunication infrastructure, of course. However, it is extremely necessary to have good ways to facilitate the dataflows to access the desired information inside the referential directories. However, the most difficult barrier to be transposed is the cooperation of all the areas involved in the project and the metadata loading and updating. This likely will require much organization and operational effort from each area classified as information provider (the servers). We must remember that if we would like to access the information it is necessary to organize them before. Likewise, if we want updated information it is necessary to have a systematic process of updating. Even the world network, the INTERNET, requires information providers to organize the data and make them available to the world society.

By the way, people could ask if the Internet would be the correct via to control the referential directories. We have tried to explain that it is necessary a certain control in each delimited environment to have a referential directory. Using the Internet we do not have this kind of control.

Nobody controls the Internet. Internet operates as an anarchic cooperation. To comply the premises of metadata organization and systematic updating it is necessary to have a kind of coordinator as described before. However, we are not saying that the information inside the referential directories cannot be accessed via Internet. It is sufficient that one area, which belongs to the referential system has its referential directory available into the Internet network. So, all the links of the correspondent referential system could be available to access the information. Normally this will happen because nowadays Internet is the main via for information dissemination all over the world.

Finally, we can say that this idea to have a referential directory is not new in Brazil. The Brazilian government tried in 1984/85, through a special agency of information and modernization (SEI/SEMOR) to create a referential directory containing the metadata about all the databases in the Brazilian public organisms. The experience failed because the available information technologies, in that occasion, did not allow to create the necessary structure and the organisms did not join themselves to build up an alternative to maintain the information updated in the referential directory. Moreover, the adopted strategy was an enormous challenge because it tried to create in only one step the referential directory at top level. They wanted to create the Brazilian referential directory for all databases produced by the public organisms in Brazil. In fact, it was a tremendous challenge.

This presented proposal has the intention in creating a referential directory at IBGE level

(IBGE's areas) as the first delimited environment. In practice, the project of the IBGE's Referential Directory is divided in three parts: the Referential Directory definition and the applied methodology; Hardware and software definition; and Referential Directory implementation and Dissemination. This is what IBGE thinks to develop during all 1996. After that, in a second phase, it is thought to link, step by step, the IBGE to other public organisms (likely by thematic areas: education, transportation,...) as other environments. Finally it would be possible to create an enormous information network in Brazil containing all the necessary metadata about statistical and geoscience surveys. However, when the first step is completed all the metadata can be disseminated all over Brazilian territory and all over the world using the Internet. So, this referential directory will be an enormous repository to metadata dissemination.

REFERENCES

- 1 - Cabral M.S, Fluxos e Bases de Dados Estatísticos na Administração Pública Federal, Tese M.Sc. Engenharia de Sistemas, COPPE/UFRJ, 1990
- 2 - Cabral M. S. & Guedes A.P., Metadata Bank, Annual Research Conference, Bureau of the Census, 1993
- 3 - Statistical Journal of the United Nations Economic Commission for Europe, Vol. 10, Number 2, 1993
- 4 - SEI/SEMOR, Diretório Referencial de Bases de dados na Administração Pública Federal, MCT/DF, 1986
- 5 - SEI/85, Relatório da Comissão Especial de Integração dos sistemas de informação no serviço público, CE/20, MCT/DF, 1985