DYNA
ISSN: 0012-7353
Universidad Nacional de Colombia

Jiménez-Ramírez, Claudia; Burke, Maria Edith; Rodríguez-Flores, Ivonne
Statistical metadata in knowledge discovery
DYNA, vol. 84, no. 202, 2017, July-September, pp. 270-277
Universidad Nacional de Colombia

DOI: https://doi.org/10.15446/dyna.v84n202.61417

UNIVERSIDAD
NACIONAL
DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Statistical metadata in knowledge discovery

Claudia Jiménez-Ramírez [a], Maria Edith Burke [b] & Ivonne Rodríguez-Flores [c]

[a] *Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia. csjimene@unal.edu.co*
[b] *Winchester Business School, University of Winchester, United Kingdom. Maria.Burke@winchester.ac.uk*
[c] *Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia y Facultad de Informática y Electrónica, Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador. ierodriguezf@unal.edu.co, irodriguez@espoch.edu.ec*

Received: December 5th, 2016. Received in revised form: July 11th, 2017. Accepted: August 25th, 2017.

Abstract
Metadata represents the semantic schema of the data collected over the years by an organization in order to apply the business intelligence approach. However, the metadata normally collected are not enough to facilitate knowledge discovery processes because they are conceived, primarily, for the interoperability between information systems. Research undertaken in this study confirmed the need to enrich data warehousing systems with structured meaningful metadata in order to increase the productivity and efficacy of any investigation, including data management and future business analytics. This need led us to adopt and extend the concept of “statistical metadata”. Thus, our proposed conceptual model of statistical metadata not only considers recognized standards, but also represents other additional properties. This means that our conceptual model allows increased levels of detail about the data and quality of the semantic contents.

Keywords: statistical metadata; knowledge discovery; knowledge management; data analytics.

Metadatos estadísticos en el descubrimiento de conocimiento

Resumen
Los metadatos representan el esquema semántico de los datos recolectados a lo largo de los años por una organización para aplicar el enfoque de inteligencia de negocios. Sin embargo, los metadatos normalmente recopilados no son suficientes para facilitar los procesos de descubrimiento de conocimiento porque están concebidos, principalmente, para la interoperabilidad entre sistemas de información. La investigación realizada en este estudio confirmó la necesidad de enriquecer los sistemas de almacenamiento de datos con metadatos significativos y estructurados con el fin de aumentar la productividad y la eficacia de cualquier investigación, incluida la gestión de datos y la analítica futura del negocio. Esta necesidad nos llevó a adoptar y ampliar el concepto de "metadatos estadísticos". Por lo tanto, nuestro modelo conceptual propuesto de metadatos estadísticos no sólo considera estándares reconocidos, sino que también representa otras propiedades adicionales. Esto significa que nuestro modelo conceptual permite mayores niveles de detalle sobre los datos y la calidad de los contenidos semánticos.

Palabras clave: metadatos estadísticos; descubrimiento de conocimiento; gestión de conocimiento; analítica de datos.

1. Introduction

Recent developments in information technologies, are enabling organizations of all types and sizes to understand and analyse massive amounts of data about their performance and environment. Thus, new management approaches that promote decision-making processes based on data rather than intuition and subjective judgements are becoming more important. The Business Intelligence (BI) approach is a good example of this approach: “BI is a process for analysing data to help managers and other end users to make decisions that are more informed, based on the use of the information for making decisions with more accurate diagnoses and smarter solutions” [1].

BI technology enables organizations to collect and integrate data, prepare for analysis, run queries against the data, generate reports and create dashboards or other types of data visualizations for supporting decision-making processes. The BI approach can also incorporate forms of advanced analytics (called Business Analytics) such as data

How to cite: Jiménez-Ramírez, C., Burke, M.E. and Rodríguez-Flores, I., Statistical metadata in knowledge discovery. DYNA, 84(202), pp. 270-277, September, 2017.

Revista DYNA, 84(202), pp. 270-277, September, 2017, ISSN 0012-7353
DOI: http://dx.doi.org/10.15446/dyna.v84n202.61417

mining, predictive analytics, text mining, statistical analysis and big data analytics for knowledge discovery. Additionally, Information Management involves not only knowledge generation, but also the use and dissemination of information.

A major component for BI is the data warehouse, which serves as a central repository for all data required for analytics and knowledge discovery. However, the use of data from these repositories is not possible without complete and meaningful metadata.

In the literature, we found different models and standards for metadata specification. According to the European Statistical System [2] the ones which are mainly used by the National Statistical Institutes (NSI) are the Common Warehouse Metamodel (CWM); the Nordic Metamodel; the Generic Statistical Information Model (GSIM); the Data Documentation Initiative (DDI) standard, ISO/IEC 11179 and the Statistical Data and Metadata eXchange (SDMX) standard. Unfortunately, these kinds of models and standards aim to manage the processes between statistical organizations, facilitating the exchange and interoperability between different data sources rather than considering the usefulness of metadata in all phases involved in knowledge discovery and data management. This became evident in a previous study which was concerned with constructing a data warehouse for a research study about innovation in Colombia [3]. We confirmed in that study that conventional metadata are very limited to managing basic information such as data format, type and user privileges. For this reason, conventional metadata cannot ensure the continued viability and usability of data. and unless changes are made, this is likely to be a continuing future issue.

The data for the study of innovation was based on two well-known and recognized Colombian governmental institutions: the Administrative Department of National Statistics (DANE) and from the Administrative Department of Science, Technology and Innovation (COLCIENCIAS). Although these institutions offered some metadata, the amount of available data was limited and not sufficient for our purposes to facilitate the data analysis and the direct access to the data for each member of the research team, (Further collection of information from these bodies would have increased the time and cost of the project)

Further academic research work confirmed the necessity of enriching data warehousing systems with structured meaningful metadata in order to increase the productivity and the efficacy in any investigation, as well as in data management and future business analytics. This necessity led us to adopt and extend the concept of statistical metadata. Thus, our proposed conceptual model of statistical metadata not only considers recognized standards, but also represents other additional properties. Thus, our conceptual model covers increased details and descriptions about semantic contents of data and their quality.

However, the survey and census data were valuable resources not only for governmental departments and academic researchers, but also for the private business sector. These data sets constitute valuable and irreplaceable resources that must be managed in a way that encourages their widest possible usability. In this way, metadata are essential for compensating for the distance in time and space between the source and the usage of the data [4]. Therefore, all kinds of producers or stewards of data should take into account that data without the appropriate metadata are useless and even harmful in some cases, because of the possible misinterpretation of data. Nevertheless, the need for more comprehensive metadata demands both time and effort, and, the aim of this paper is to propose a new conceptual model which enriches conventional metadata but with minimal relevant elements, moving towards a more simple but complete model, which is much more practical for analysts and decision makers.

In the commercial field, suppliers such as IBM, Informatica and Collibra, which are leading companies according to the latest report of Gartner 2017 [5], address the metadata management through their products which are mainly oriented to data governance. These products provide functionalities for business glossary management. The business glossary is a repository for collecting and sharing the business terms and their meaning in a specific organizational context. The glossary aims to provide a set of definitions of common vocabulary for technical and business users [5,6]. A business glossary generally includes term definitions, term names, term definition examples, term acronyms, term abbreviations, security aspects and compliance restrictions. Furthermore, the business glossary usually has metadata about the data, such us the name (name of the person creating the term definition), the data steward names (assuming there is not just one steward) and the data steward´s contact information (phone, email, location) [7]. Nevertheless, the business glossary is usually stored outside the warehouse as the purpose and use is mainly concerned with facilitating communication. In this research, we propose to store administrative and technical metadata together in the data warehouse in order to use them in knowledge discovery phases.

This paper has seven sections. This first section introduces the topic and sets out the aim of our work. The second section, presents a review of literature regarding data warehousing systems and metadata, followed by two case studies which form the third section and confirm the necessity of metadata in knowledge discovery. In the fourth section, we present the materials and methods used in our research work. The results of this work, in the form of the presentation of our new proposal of the conceptual model for statistical metadata, are outlined in section five. In addition, in the sixth section, we present the discussion about our results. Finally, we present the contributions and conclusions.

2. Data warehouse and metadata

A data warehouse is an integrated, time-variant and non-volatile collection of data in support of the decision making process. The distinctive characteristics of a data warehouse is that it contains historical, granular and integrated data, so that various groups of people can examine and analyse the same datasets and achieve a single version of the truth [8].

The data warehouse definition given by Inmon (1991) is valid today, but the data warehousing architecture has evolved to 2.0 version (DW 2.0). According to “DW 2.0 architecture” proposed by Inmon (2008) and "The Data Warehouse Lifecycle Toolkit" by Kimball (2013), metadata are the cornerstones of data warehousing systems. However, this requirement of metadata contrasts with the fact that many information systems do not implement them at all or do so only partially. With complete metadata, data reusing is much easier and this would help avoid duplicating work in data analytics or research studies which have been undertaken in all types of organizations.

2.1. Metadata

The National Information Standards Organization defines metadata as: "*structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource*" [9]. However, the definition which is better known (in metadata) in any area is "*data about data*", making it more general, independent of context, but rather more ambiguous.

Currently the definition of metadata which has also become popular is that which concerned the organizing of Web resources for facilitating interoperability and integration of resources, in order to provide digital identification and preservation of information. For example, Berners-Lee and the World Wide Web Consortium (W3C) have a more narrow and restrictive metadata definition: "*Metadata is machine understandable information about web resources or other things*" [10].

In this research work, we searched for metadata standards related to data warehousing systems. For this purpose, we mention the Common Warehouse Metamodel (CWM - ISO/IEC 19504) and ISO/IEC 11179 norms (Metadata Registry - MDR). Other contributions are also of relevance, such as the Data Documentation Initiative (DDI) and Metadata Common Vocabulary (MCV). Both of these models are complementary as they work "together" on standards and models already mentioned. However, unfortunately, these proposals also focus on the exchange and interoperability between different data sources. This means that they are restricted on the use of metadata for machines - but not for humans involved in any process of business analytics or knowledge discovery such as data cleansing.

2.2. Statistical metadata

Among the informatics community, the meaning of metadata is usually limited to formal descriptions of how data is typed and formatted with the aim of facilitating the interchange and the interoperability of information systems, as mentioned earlier. Thus, current data warehousing systems only specify some descriptions that are in fact about syntactic properties of data rather than semantic ones.

As conventional metadata do not offer specifics or the context needed to communicate meaning and analyse data, we adopt the term "statistical metadata". Nevertheless, we found that statistical metadata has focussed only on the official statistics generated by the National Statistical Institutes (NSI) in different countries.

Official statistics have been one of the main areas that have taken on board and recognized the importance of metadata. International organizations such as UNECE, EUROSTAT, OECD have promoted projects that have resulted in new models and standards for statistical metadata, such as Generic Statistical Information Model (GSIM), Statistical Data and Metadata eXchange (SDMX - ISO 17369:2013), et al. Other proposals have emerged from initiatives by National Statistical Institutes (NSI), such as the Nordic Metamodel. Unfortunately, these contributions are focused on the "metadata on statistics data" generated in NSI, as presented in the Common Metadata Framework [11]. Thus, statistical metadata is defined as "data about statistical data" [12,13].

On the other hand, we agree that statistical metadata is "*any information needed by people or systems to make proper and correct use of the real statistical data, in terms of capturing, reading, processing, interpreting, analysing and presenting the information (or any other use)*". In other words, statistical metadata is anything that might influence or control the way in which the information is used by people [14]. In addition we propose, expanding the "*statistical metadata*" beyond the domain of the NSI, so that the metadata is not limited to statistical data, but to any dataset which is of interest of the organization. With this definition in mind, data warehouse designers, managers and data analysts must pay special attention to statistical metadata.

Statistical metadata has two abstraction levels: *microdata* and *macrodata*. Microdata are data about the characteristics of units of a population, such as individuals, households or establishments, collected by census, surveys, or experiments. Macrodata, however, can be defined as data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies [13].

Statistical metadata are not limited to the form and contents of data. This kind of metadata also includes administrative facts about data, about ownership responsibility for the initial data collection and the internet access. These administrative metadata are very useful for not only searching and locating of data, but also knowing about their reliability and novelty. However, unfortunately, government official bodies or any kind of open data sites, usually do not offer all the metadata which are needed to fully undertake all the elements of the business analytic processes.

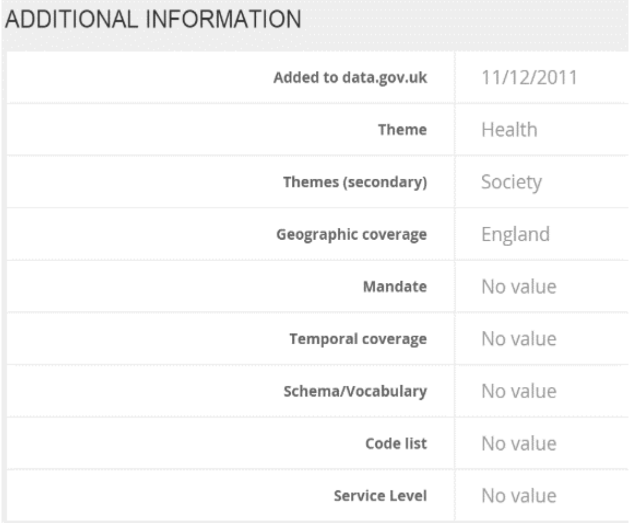
For instance, Fig. 1 shows a snapshot of some metadata for the open data web site: "DATA.GOV.UK – Opening up Government". A brief description of columns is displayed in the dataset but these columns concern only a few characteristics of administrative data and many of them have unknown values.

3. The role of statistical metadata in knowledge discovery

Before we present the conceptual model of statistical metadata, we must understand why it is necessary to enrich data using statistical metadata. For this purpose, we first present two real cases which highlight some problems that can arise due to the lack of complete metadata.

3.1. Research study 1: Innovation in Colombia

As mentioned in the introduction of this paper, the earlier knowledge discovery project regarding "Innovation in Colombia" was completed during 2009. For that work, we used several "Annual Manufacturing Surveys" provided by



ADDITIONAL INFORMATION

Added to data.gov.uk	11/12/2011
Theme	Health
Themes (secondary)	Society
Geographic coverage	England
Mandate	No value
Temporal coverage	No value
Schema/Vocabulary	No value
Code list	No value
Service Level	No value

Figure 1. Example of official metadata.
Source: [15]

the Colombian official statistical department DANE and two research databases from the Administrative Department of Science, Technology and Innovation (Colciencias) for building a data warehouse. DANE provided us with an electronic sheet and a short description of columns included in each file, corresponding to a specific year, as metadata. These annexes were crucial because most of the names of columns were numbers which referred to the questions in the surveys. Nevertheless, the researchers needed to look at several digital or paper documents to choose the data properly, which considerably slowed down the analysis of data. If anyone wanted to know, for example, the mean percentage of profits destined for research and development (R&D) of industrial establishments at Medellin city, in a specific year, first it was necessary to identify the annex corresponding to city's codes. After that search, in another electronic sheet, the researcher then needed to search the name of the column that represents R&D variable. Only after these tedious searches, knowing the code for Medellin city and the name for R&D variable, was it possible to return to the survey data for searching for the required value.

We can say that the problem originated by non-mnemonic names is general, and affects both public and private organizations. In addition, it is important that the analysts can have a good quality description of the data for each table or column in order to find the meaning of any variable and other descriptions such as unity of measure, easily and quickly.

On the other hand, inevitably, data structures change over time. This happened to the manufacturing annual surveys. Every year, the DANE department deletes some questions or adds new ones, but we did not have the metadata to allow us to create an integration process using all of the survey results. Additionally, we realized that some concepts such us "small establishment" varies over time and the data producer needed to inform DANE about these changes.

In addition to the above, the conventions used for representing missing values, the format for dates, among other features, delayed the data analysis. In the official surveys used for the "Innovation in Colombia" study, the conventions for missing values were to assign the number "-99999" but sometimes there were four or six nines. Additionally, these negative numbers affected any numeric calculation. Thus, we needed to "clean up" all the missing values before data analysis, increasing time and efforts.

In addition as well as considering the process of metadata in data integration, there are also essential data cleansing tasks. If the data are not valid, the descriptions nor the inferences made based on them are also likely to be incorrect which could lead to poor decisions that cause loss of time, money and credibility. Data mining has little or no use if applied to data that is of low quality.

Concerning integral data quality, we must consider several dimensions, such as: accuracy, completeness, consistency, and timeliness which are among the most common elements when measuring good quality of information [16]. Mistakes are frequent despite the care in the data capturing by any operational systems and the data produced by official departments. In the manufacturing annual surveys of Colombia, we found all types of errors. Thus, it was necessary to enrich the data warehouse to include the maximum and minimum values that could take each of the attributes of continuous type, in order to know if there were values outside the domain. Similarly, for discrete attributes, we used the set of possible values for checking problems of domain or referential integrity. For instance, we found the number 13 to describe types of legal entity when it was only possible to have 12 different numbers.

Furthermore, when performing data cleansing tasks, we also needed to analyse the outliers in every attribute or variable, in order to determine other possible errors. For instance, the percentage of revenues for investment in R & D values were not outside the domain which indicated erroneous values, but rather the two extreme values of 92% and 87%. These values were typing errors because no Colombian industrial establishment would allocate such a high percentage of sales in R & D. Thus, the importance of the detection of outliers is important as they can have critical effects on data mining, when they are not considered [17].

It is important to know which technique we can choose in almost predictive analysis, for example, whether the distribution of data from a variable has an acceptable fit to some normal distribution. For this, it is necessary to test the normality of data distributions and save the derived values as other relevant characteristics of metadata, specifically as macrodata, to avoid the execution of the same procedures multiple times, for different users. It is also advisable to generate and save, as another metadata, the frequency histogram or box and whisker plots. These are useful descriptive graphs which determine the shape of the distribution of the data for appropriate selection of techniques or algorithms.

For visualization and representation of knowledge, it is also essential to have an explicit statement of metadata, including units of measurement for each variable in any tabular or graphical results produced in a particular analysis. As an example, Table 1 shows the results of a query without human intervention [18]. We can see the great difficulty for understanding the meaning of any variable such as V1612V and its maximum value of 0.92. Instead of using the original short names, the meaning or description for each variable or attribute could automatically replace them by using new data warehousing systems when presenting the results.

Table 1.
Five-number Summary for Selected Variables.

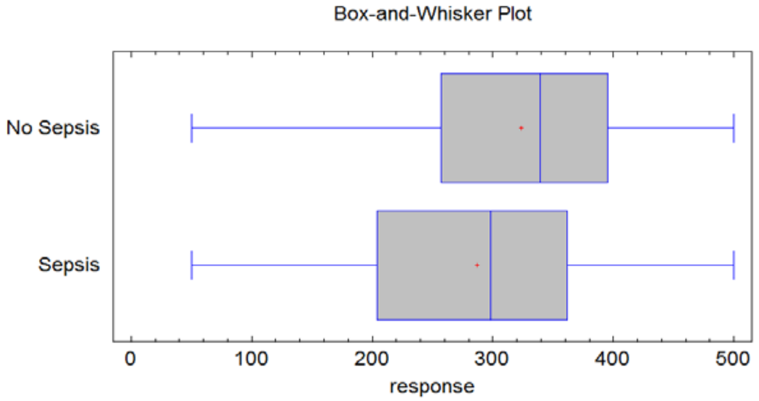
	PEC	**V1612V**	**V1613A01**	**V1613A02**
Minimum value	0	0	0	0
Quartile 1	0	0	0	0
Median	0,03	0	0	0
Quartile 2	0,06	0	0	1
Maximum value	0,76	0,92	12	70

Source: [3]

3.2. Research about sepsis in adult patients

Table 2.
Metadata Related to Attributes in Sepsis in Adults Research Project.

Full Name	***PaO_2/FiO_2* ratio**	**Short name**	**pao2_fio2**
Description	Derived from blood pressure (*PaO_2*) divided by inspired fraction of oxygen (*FiO_2*). The ratio of partial pressure arterial oxygen and fraction of inspired oxygen, sometimes called the *Carrico* index. A *PaO_2/FiO_2* ratio less than or equal to 200 mmHg is necessary for the diagnosis of acute respiratory distress.		
Source	Pacients.xls		
Variable type	Quantitative	Data type	Real positive
Measure unit	mmHg		
Missing values	27 of 589	Median	310
Mean	299	Standard Deviation	108,4
Theoretical minimal value	Unknown	Theoretical maximal value	Unknown
Observed minimal value	50	Observed maximal value	500
Normal Distribution?	Shapiro-Wilk Test (95% confidence) *P*-value = less than 0, 05. Normality on data is rejected		
Observations	A *PaO_2/FiO_2* ratio less than or equal to 333 mmHg is one of the variables in the SMART-COP risk score for intensive respiratory or vasopressor support in community-acquired pneumonia		



Source: [19]

More recently, we have conducted research about Sepsis in adult patients that led us to enrich the conceptual metadata model in order to facilitate the data cleaning process and avoid repeating the same tasks in further analyses. For that research, we designed formats for registering the statistical metadata to discover the data quality and also to ascertain the data distribution for selecting the proper techniques required in comparisons. As an example, Table 2 displays the format for documenting each attribute or variable in the dataset. The table includes a boxplot to visually describe the data distribution and detect possible differences in the response variable: the category of adult patients (sepsis and non-sepsis patients) [19].

All the statistical metadata collected for the Sepsis research project such as that presented in Table 2, should be stored as metadata in order to detect changes in variables and to formulate hypothesis about which variables we must consider for deeper analyses. For instance, in the Sepsis database, the variable with short name *pao2_fio2* is the ratio of partial pressure arterial oxygen and fraction of inspired oxygen, sometimes called the Carrico index. According to [20] a *PaO2/FiO2* ratio less than or equal to 333 mmHg is one of the variables in the SMART-COP risk score for intensive respiratory or vasopressor support in community-acquired pneumonia. Thus, it is important to know this information to compare the groups in clustering analysis. For example, the distribution of adults who developed sepsis has a lower *PaO2/FiO2* ratio, as shown in Table 2: The median was 300 mmHg indicating that the 50% of that group had pneumonia. In addition, almost 75% of these patients are at high risk. However, we aimed to validate our findings using a statistical approach and discovered that the Carrico index did not fulfil the normality assumption for the distribution; we needed some non-parametric test to show median differences. Thus, the macrodata included this as part of our results in order that further studies would be able to reduce any unnecessary additional time and effort.

4. Materials and methods

The definition of the conceptual model and its validation constituted an iterative cycle of refinement. The approach chosen in this research for defining the conceptual model was mainly a top-down approach. In this approach, the process started with specifying the global requirements and assuming that each component has global knowledge of the system. We concluded with a model of the solution by enriching the initial model with details.

In the conception of our model, we consider the points of view raised by Grossmann (2002) that indicates that the "*Statistical structures are defined by a number of statistical objects, necessary to model statistical units, populations, statistical variables and statistical data. These objects are defined by their properties, their relations, and the statistical methods which can be applied to such statistical objects in order to produce new statistical objects*" [21].

The conceptual model presented in the next section as a result of our research work is presented using the Unified Modelling Language™ (UML®) Version 2.5 standard [22]. We propose the use of UML because this is an object oriented visual modelling language that has been widely accepted and is easy to understand for users.

The software tool used for drawing the diagram using the UML language was Visual Paradigm Community Edition. This tool supports the latest standards of UML notation.

5. Conceptual model for statistical metadata

The problems outlined in the previous examples forces users to waste efforts that can be avoided if the metadata were available and complete. Perhaps, its importance should be stated more clearly: without metadata, the data is totally meaningless, the content of the information system becomes inaccessible or is of little use [23].

In order to have enough statistical metadata to enrich digital information and create knowledge, we have conceived and designed the conceptual model presented in Fig. 2. This model presents the crucial parts of any data repository. The data source class represents the basic requirements for knowledge discovery. For each instance of data source, such as a text file or some relation in a relational database, is necessary to know some details about its context and keep all administrative metadata for future acquisitions e.g. the format or the URL. Furthermore, to some extent, this administrative metadata also helps to determine the reliability of data.

However, we also know that some data sources are an aggregation of families or groups of variables or attributes. For instance, a survey can have sections to include both demographic aspects and behaviour aspects. Thus, in our conceptual model, a data source can be composed either by a group of variables or directly by individual variables. For each variable of a dataset, is important to collect the full name besides its short name (usually the name in original file) and the other characteristics depicted in Fig 2. The statistics vary depending on the type a variable. Thus, for quantitative variables we propose to store the mean, the median and the mode as estimations of central tendency and the standard deviation and absolute deviation mean as estimations of the dispersion, as well as the maximum and minimum values. For each categorical variable, it is important to record the mode or modes and the more infrequent values. The diagrams about data summaries are also different for quantitative and qualitative variables and should be stored in the data warehouse as macrodata. This kind of metadata, macrodata, about groups or collections is very useful in data comprehension and the selection of data mining techniques.

The values or measurements for variables like quantities or prices constitute the class of main interest in any analysis. These data are termed as "*Facts*" or *Data Item class* (see Fig. 2). In the case of our research about innovation, for instance, we considered the *Fact Class* has all the quantitative characteristics about industrial establishments such as the quantity of qualified workers and their annual revenues. Any fact table has several dimensions, such as the economic activity and the size of the establishment for our example. These dimensions derive from qualitative variables and are useful for multidimensional analyses. Because data warehouses store historical data, the dimension *Time Class* always appears in the model and we present this dimension as an instantance of *Dimension.*

Considering changes can occur in some dimensions, we created the *Validity Period* class in order to register the dates when one instance was valid and the final values before the change. In this way, it will be easier for all organizational members and researchers to consider the correct values for dimensions. In the same way, in order to manage the changes in metadata without significantly affecting the simplicity of the model, we created a relationship between *Validity Period* and the entity *Variable/Attribute*. This relationship is exclusive with respect to the existing link between *Validity Period* and *Dimension* classes, because one change is for one dimension or for some of the statistical metadata, but not for both simultaneously.

In order to create summaries or model fitting it is usually necessary to have different views of data and we therefore suggest a procedure to register the views generated in our conceptual model for metadata as the data source of all the academic production resulting from research works. The reports or models must be related with the techniques or methods used for that purpose. Finally, all the academic production is very useful as future points of reference.

6. Discussion

We present in this document a proposal of metadata which is useful in all the phases of knowledge discovery including data profiling, data analysis and data visualization. This will allow increased understanding and comprehension of the domain and improve communication and comparison of results of future research studies.

We conceived an extension of statistical metadata in a data warehouse applicable to any type of organization, not only for use in official institutions. Descriptive statistics such as minimum, maximum, mean, mode, standard deviation, frequency, aggregates such as count and sum, and additional metadata such as data type, length, discrete values, uniqueness and other restrictions are required for discovering

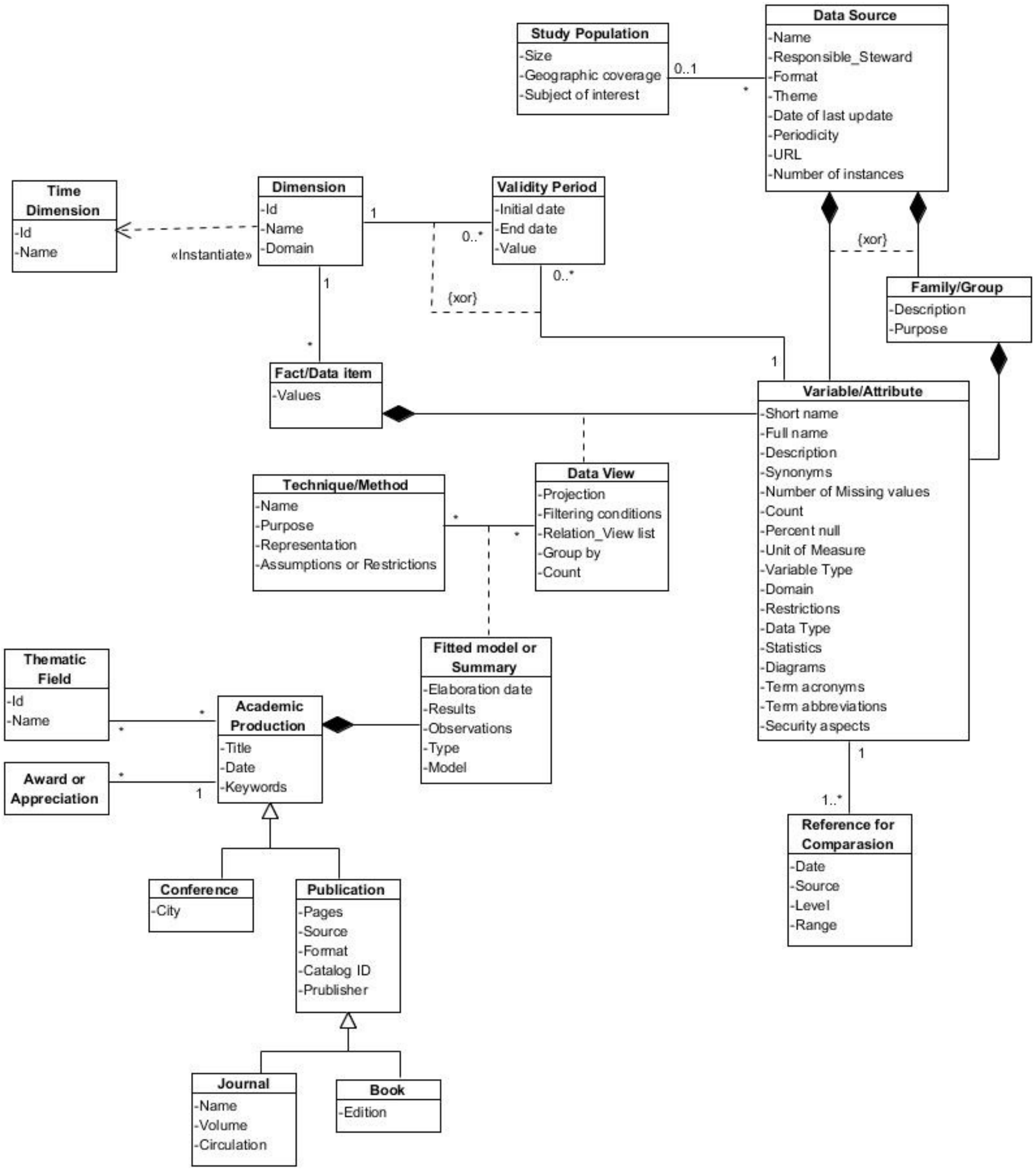


Figure 2. Conceptual model: Collecting data in Data Warehouses; New forms of statistical metadata
Source: The authors

problems such as illegal values, misspellings, missing values, and duplicates in all the contexts. Statistical metadata is vital for the processes of producing and interpreting knowledge. However, defining statistical metadata requires knowledge of the potential users and usages and, thus, this is a complex task. The breadth of meaning is such that the metadata producer must address its production in a manner similar to that used for producing the data itself. Therefore, metadata management must be seen as an integrated part of data production. In addition, as the access of metadata is frequently combined with the data, we propose to structure both of them together in the warehouse.

7. Conclusions

Metadata has several roles to play in addition to the conventional use of interchange and interoperability between information systems. Metadata should provide the basics for data quality validation, data preparation, and for selecting the proper techniques in analytics processes and data comprehension and dissemination. Consequently, metadata should support any knowledge discovery in databases and analysts should use them in order to be increase the efficiency and effectiveness of business processes.

We adopt and generalize the term statistical metadata to emphasize the necessity to design and represent administrative metadata, microdata as well as macrodata with the aim to have an explicit and complete statement of the information in any data warehouse. Administrative metadata are necessary because currently the data in a data warehouse arises from multiple sources and these sources not only are relevant for future acquisitions, but they also determine in part the data quality. In this way, we propose an enrichment of metadata, specifically in macrodata, in order to improve the accessibility, generation and dissemination of current and future research.

References

[1] Loshin, D., Business intelligence and information exploitation, in Business Intelligence: The Savvy Manager's Guide, 2a ed., Ed. MA, USA: Morgan Kaufmann, 2013, pp. 1-13. DOI: 10.1016/B978-0-12-385889-4.00001-6

[2] ESSnet-Data warehouse, Metadata framework for statistical data warehousing, MEETS ESSnet projects-European Commission, 2013. [Online]. Available at: http://ec.europa.eu/eurostat/cros/content/dwh-sga2-wp1-11-metadata-framework-statistical-data-warehousing-v112-final_en.

[3] Robledo, J., Malaver, F. y Vargas, M., Encuestas, datos y descubrimiento de conocimiento sobre la innovación en Colombia, 1st ed., Bogotá: Javegraf, 2009.

[4] Dippo, C.S. and Sundgren, B., The role of metadata in statistics, International Conference on Establishment Surveys II, [online]. 2000, pp. 1-12. Available at: http://www.bls.gov/ore/abstract/st/st000040.htm

[5] Gartner, Magic quadrant for metadata management solutions, [online]. 2017. Available at: https://www.gartner.com/doc/3778891/magic-quadrant-metadata-management-solutions

[6] Sigmon, P.W., Getting started with information governance: The glossary approach, IBM Data Management Magazine, [online]. 2013. Available at: http://www.ibmbigdatahub.com/blog/getting-started-information-governance-glossary-approach

[7] Fryman, L., Lampshire, G. and Meers, D., Aligning the language of business: The Business glossary, in The Data and Analytics Playbook, Eds. Boston: Morgan Kaufmann, 2017, pp. 137-157. DOI: 10.1016/B978-0-12-802307-5.00005-8

[8] Inmon, W.H., Strauss, D. and Neushloss, G., DW 2.0: The architecture for the next generation of data warehousing. Morgan Kaufman Series in Data Management Systems, 2008. DOI: 10.1016/B978-0-12-374319-0.00002-6

[9] NISO, Understanding Metadata, 1st ed. MD, USA: National Information Standards Organization, [online]. 2004. Available at: http://hdl.handle.net/10150/105486

[10] Berners-Lee, T., Web architecture: Metadata, 1997. [Online]. Available at: https://www.w3.org/DesignIssues/Metadata.html. [Accessed: 15-Jul-2017].

[11] UNECE, CMF Part A - Statistical metadata in a corporate context: A guide for managers, Geneva: United Nations, [online]. 2009. Available at: https://statswiki.unece.org/display/metis/Part+A+-+Statistical+Metadata+in+a+Corporate+Context

[12] Sundgren, B., Guidelines for the modeling of statistical data and metadata, Geneva: United Nations Statistical Commission and Economic Commission for Europe, 1995. [Online] Available at: http://www.unece.org/fileadmin/DAM/stats/publications/metadatamodeling.pdf. [Accessed: 15-Jul-2017].

[13] OECD, OECD Glossary of statistical terms. Paris: OECD Publishing, 2008. DOI: 10.1787/9789264055087-en

[14] Westlake, A., Models and metadata, in Proceedings of the Final MetaNet Conference, Ed. Greece: University of Athens, 2003, pp. 108-117. [Online] Available at: http://www.data-archive.ac.uk/media/1689/METANET_proceedings_finalreport.pdf. [Accessed: 15-Jul-2017].

[15] DATA.GOV.UK, Statistics on Obesity, Physical Activity and Diet, England - Datasets, Datasets, 2015. [Online]. [Accessed: 1-Dec-2016]. Available at: https://data.gov.uk/dataset/statistics_on_obesity_physical_activity_and_diet_england.

[16] Batini, C. and Scannapieco, M., Data quality: Concepts, methodologies and techniques, 1st ed., NY, USA: Springer Berlin Heidelberg, 2006. DOI: 10.1007/3-540-33173-5

[17] Weisberg, H.I., Bias and causation: Models and judgment for valid comparisons. Wiley Blackwell, 2010. DOI: 10.1002/9780470631102

[18] Jiménez, C., Villa, F. Rico, M., Metadatos de una bodega de datos para descubrir conocimiento, en: Encuestas, datos y descubrimiento de conocimiento sobre la innovación en Colombia, 1st ed., Javegraf, 2009, pp. 33-51.

[19] Rodríguez, A., Soporte para el diagnóstico de sepsis en adultos, usando técnicas de minería de datos supervisadas, MSc. Thesis, Universidad Nacional de Colombia, Medellín, Colombia, 2015.

[20] Pereira, J.M., Paiva, J.A. and Rello, J., Severe sepsis in community-acquired pneumonia — Early recognition and treatment. Eur. J. Intern. Med., 23(5), pp. 412-419, 2012. DOI: 10.1016/j.ejim.2012.04.016

[21] Grossmann, W., Structures for metadata, in metanet work package 1: Methodology and Tools, Ed. The MetaNet Project, [online]. 2002, pp. 11-28. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.203.5184&rep=rep1&type=pdf. [Accessed: 15-Jul-2017].

[22] Object Management Group, OMG Unified Modeling Language (OMG UML) - version 2.5, [online]. 2015. Available at: http://www.omg.org/spec/UML/2.5/. [Accessed: 15-Jul-2017].

[23] Inmon, W.H., O'Neil, B. and Fryman, L., Business metadata: Capturing enterprise knowledge, San Francisco, USA: Morgan Kaufmann Publishers Inc., [online]. 2008. Available at: https://www.elsevier.com/books/business-metadata-capturing-enterprise-knowledge/inmon/978-0-12-373726-7

C. Jiménez, received the BSc. title in Statistics by the Universidad de Medellin in 1983, a Sp. degree in Programming and Databases Technologies in 1992, MSc. degree in Systems Engineering in 1999, and the PhD degree in Systems and Informatics Engineering in 2008, at the Universidad Nacional de Colombia, Medellin, Colombia, awarded all postgraduate titles. From 1994 to current date, she is a full professor in the Computing and Decision Sciences Department, Facultad de Minas, Universidad Nacional de Colombia. Her research interests include data mining, knowledge management and discovery using statistics and computational intelligence techniques.
ORCID: 0000-0002-0786-9513.

M.E. Burke, is head of research and knowledge exchange, Faculty of Business, Law and Sport at the University of Winchester, Hampshire, UK. She is professor of management and researches within the context of Information Systems. Her main areas of research expertise concern the application of new digital technology to economic, environmental and social systems. Research publications include several books, over 100 journal papers, international conference papers and book chapters. Professor Burke was awarded her PhD by the University of Salford in 2005. She is a member of the British Association of Management, and a Fellow of the Royal Society of Arts.
ORCID: 0000-0002-2578-7714

I. Rodríguez-Flores, received the BSc. Eng in Systems and Computing Engineering from Escuela Politécnica del Ejército, Ecuador, MSc. degree in Applied Computation from Escuela Superior Politécnica de Chimborazo, Ecuador, in 2005. Current PhD candidate in Systems and Informatics Engineering at the Universidad Nacional de Colombia, Medellin, Colombia. From 1993 to current date, she is a full professor in the Facultad de Informática y Electrónica, Espoch, Ecuador. Her research interests include business intelligence, data warehousing, data quality, metadata and information management, and data governance.
ORCID: 0000-0002-7788-2609