Data Quality Assessment at the Rio de Janeiro Botanical Garden Herbarium Database and Considerations for Data Quality Improvement

Eduardo Couto Dalcin[1], Luís Alexandre Estevão da Silva[1], Caroline Corrêa Cabanillas[2], Marcos Gabriel Surrage M. Loures[2], Vitor Faria Monteiro[1], Geraldo Zimbrão da Silva[2], Jano Moreira de Souza[2]

[1]Instituto de Pesquisas Jardim Botânico do Rio de Janeiro

[2] COPPE/UFRJ, Universidade Federal do Rio de Janeiro

{estevao,dalcin}@jbrj.gov.br, {carolcaban,gabriemannarinol}@gmail.com, vitor@nccg.jbrj.org, {zimbrao,jano}@cos.ufrj.br

Abstract. *The herbarium of Rio de Janeiro Botanical Garden (RB) is one of the most important reference collections of Brazil's diversity. The database of the herbarium is a source of data about scientific collections, providing information about taxonomic or biogeographical research. But, for several reasons, this database, like others, has many problems associated with data quality. The present work describes the methodology adopted in order to assess data quality, identifies and classifies data quality issues and quantifies the number of records affected on each category.*

Keywords: data quality, primary biodiversity data, herbarium database.

1. Introduction

The Rio de Janeiro Botanical Garden herbarium collection, and its respective database, are the core of a Brazilian's biodiversity information system, linked with three other information systems: the official Brazilian Plants Checklist (Flora, 2012) the official checklist of endangered plants (CNCFlora, 2012) and the Virtual Herbarium of Repatriated Plants under development. Through this type of information system several uses of primary data occurrence of species (Chapman, 2005a) are permitted, specifically in the databases of herbarium; data have been collected with the aim of providing information for taxonomic or biogeographical research and those preserved botanical collections records are basically composed by taxonomic and nomenclatural data, spatial data, collection data and associated descriptive data (Sodré et al, 2012).

Having this important role, data quality is fundamental, in order to provide quality data for decision and public policy makers in conservation and sustainable use of biodiversity. The present work describes the methodology adopted in order to assess data quality, identifies and categorizes data quality issues and quantifies the number of records affected on each category.

The herbarium of Rio de Janeiro Botanical Garden (RB) is one of the most important reference collections of Brazil's plant diversity. It has about 595,000 samples, where 99% are stored in a database and accessed by means of an information system, called

JABOT[1]. JABOT was developed in house using free software, more specifically with Postgresql and PHP, between 2003 and 2005; followed by a digitalization project that occurred in two phases, the first between 2005 and 2007 (Gonzalez, 2009) and the second started in 2011.

The RB stores information from different collections; the main collections of the Botanical Garden in number of records are: **Herbarium Vouchers** with 535,192 records that are testimonies of species collected; the **Wood Collection** with 9291 and that stores information about the anatomical structure of several types of wood and contributes significantly to the recognition of trees and shrubs for taxonomic and phylogenetic research, especially when the reproductive material (flowers and fruits) is absent or scarce, being that in this context, xylothecas represent an important source of information for research, providing possibilities for identification and retrieval of data on origin, collectors, etc.; **Coleção Viva** (Arboretum) with 8626 that stores taxonomic information on the live plants cultivated in the park's arboretum, to be used.

Given the complexity of the base and the availability of resources in the area of database, a methodology was developed to enable the assessment, correction and monitoring the level of data quality. Many of the necessary corrections have been implemented through scripts developed in SQL (*Structured Query Language*) (Chamberlin et ak, 1974), by means of Expressões Regulares (Melton, J., Simon, 2002), implemented no data base Postgresql.[2]

2. Material and Methods

2.1. The database

The database uses the Darwin Core Classic Metadata Standard (2009 Darwin Core), which establishes a set of basic attributes related to taxonomy and occurrence, to facilitate the exchange of information between botanical applications. The database consists of 101 tables, 333 users, 134,378 taxa and 1384 plant families. The information relative to taxonomy and occurrence of the species are stored on tables "*tree taxon*", "*data access*", "*determination*" and "*testimonies*". This information is important not only to Access primary data of the collection but also for conservation and knowledge of biodiversity. All the other tables serve as dictionaries to these and these are tables with needs for data improvement in quality. Below, better detail of the tables.

- ***Tree taxon*** – stores the scientific names used for the determination of the collections and is organized in a hierarchical structure where the levels represent the botanical;
- ***Data Access*** – Presents the data relative to the occurrence of the testimonies by means of control accesses, such as place of collection, coordinates, altitude and date of collection, major collector and other collectors;
- ***Testimony*** – presents the information on the materials collected from the species, such as type of collection, number of recording as historic site and the bar code in the vouchers;

[1] www.jbrj.gov.br/jabot
[1] www.postgresql.org

- ***Determination*** – presents the data relative to the background of the witnesses' statements. Storage is relevant to allow control of updates performed by taxonomists in the collection.

3. Data Quality Analysis

According Chapman, data quality and error in those data are often neglected issues with environmental databases, modeling systems (Santana, 2008), GIS, decision support systems, etc. Too often, data are used uncritically without consideration of the error contained within, and this can lead to erroneous results, misleading information, unwise environmental decisions and increased costs (Chapman, 2005b).

In 2010 a new institutional information system began to be developed and, in the process, an assessment of the main data quality issues took place. Previous data quality assessments (Sodré et al, 2012; Pipino et al, 2002; Strong, 1997) shows that the JABOT databases have major data quality problems, spread over many different data quality dimensions and categories. The most common types of errors in botanical collections related to data entry are: misclassification; typos; fields partially filled; data migration errors; lack of standardization of data and inaccurate geographical coordinates. Many of these data have drawbacks when it comes to use for species distribution studies (Chapman, 2005c).

The analysis by understanding the reasons of the errors in the database leads us to the main tool for data entry in JABOT. Since two large typing movements were performed in data entry by non specialized students in the botanical area or even by means of importing data from spreadsheets sent by researchers, it was not possible to import content with stricter control criteria. Culminating in several errors on the database.

Since there is no consensus yet or a standardized way to evaluate data quality in botany, in this work, we adopted a set of criteria based on research conducted by Dalcin (2004) and (Wang et al., 2000), in table 1 the author divides 16 dimensions into 4 categories.

Category	Definitions	Dimensions
Intrinsic	Intrinsic characteristics of the data, independent of their implementation.	Accuracy, objectivity, credibility and reputation.
Accessibility	Aspects relative to access and data security.	Accessibility and access security.
Contextual	Characteristics dependent on the context of data use.	Relevance, added value, temporal validity, Completeness and amount of data.
Representational	Characteristics derived from the way the information is presented.	Interpretability, ease of understanding, concise representation and consistent representation.

Table 1- Categories and dimensions of data quality.

Survey was carried out showing the most common errors in data entry in applications with botanical data collections. Figure 1 presents a classification in terms of importance and number of errors for the data analyzed.

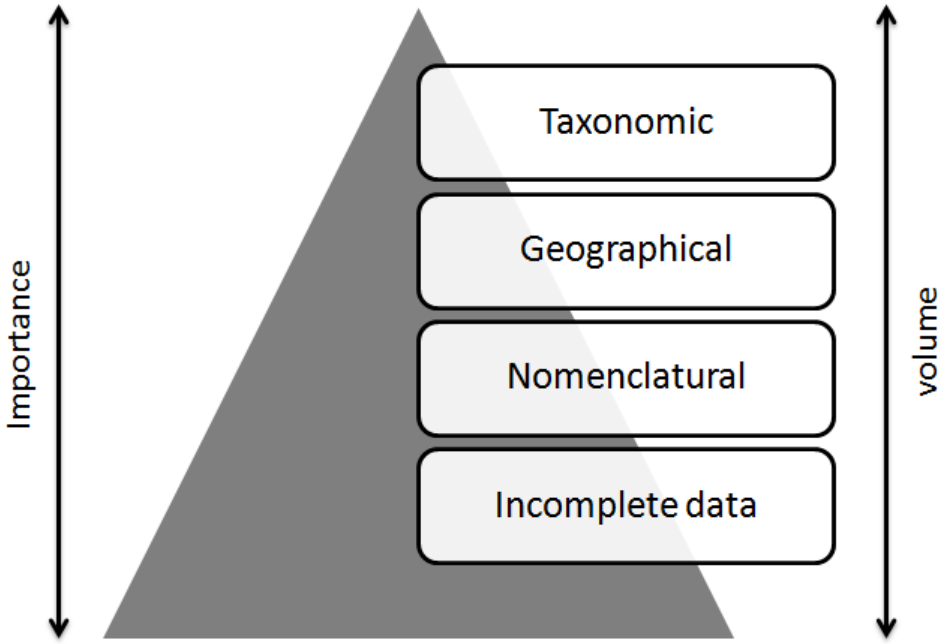

Fig. 1 - Methodology used in the evaluation process of data quality.

There were identified and classified 4 main categories of errors:

- **Taxonomic** - subdivided into *incorrect classifications* that are the errors caused by the lack of taxonomic knowledge and *typing errors* that are caused by the difficulty of reading old labels identifying some vouchers;

- **Geographical** - relative to inaccurate geographical coordinates. They are very common because many collections were made without GPS and their coordinates were added later, just based on the county where the collected was made;

- **Nomenclatural** - those occur due to lack of standardization in the names and forms of writing the same value. An example are the names of collectors and localities;

- **Incomplete data** - some attributes have been partially filled, such as the case of collection data of the plant: year, month, day, geographic coordinates.

4. Methodology

4.1 Methodology applied

The steps required for the evaluation and improvement of data quality in the data from botanical collections were divided into 5 steps, as shown in Figure 2. Thus, the methodology will consider categories and dimensions. A job of error analysis in these types of collections has been developed for the Atlas of Living Australia, in its portal of Data Quality (ALA, 2012).

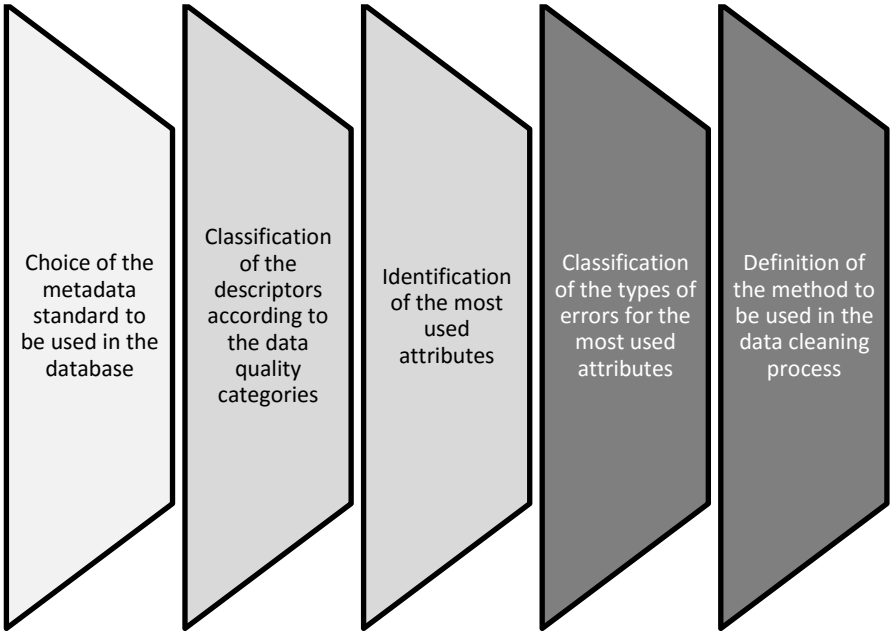

Fig. 2 - Methodology used in the evaluation process of data quality.

The starting point for implementing the steps listed in the methodology assumes that the database was developed through a project of database (Teorey, 2011) where business rules have been incorporated through the relations of the tables, restrictions on attributes and functions written in procedural language. Description of steps:

a) ***Choose the metadata standard to be used in the database*** **-** it should be checked from the standards used in the scientific area of the application, which is the most appropriate metadata. In the present work, the Darwin Core metadata (Darwin Core, 2012) was selected. The use of metadata facilitates data organization through the precise definition of attributes, facilitating subsequent quality classifications and also data exchange between applications;

b) ***Classification of descriptors according to the categories of data quality*** **-** provides a pre-assessment of the attributes more susceptible to errors. An assessment as to the accuracy criterion is presented in Table 2 below. The criterion refers to what extent the datum is correct and reliable and if the reputation corresponds to the credibility of the source organization.

Attributes	Type of errors	Criterion
CatalogNumber Year Identified Month Identified Day Identified Type Status Collector Number Day,Year, Month Collected	Invalid collection date Missing collection date Missing dateidentified Missing identifiedby Invalid of month Invalid of year Invalid of day	Accuracy
Country State Province Longitude, Latitude-degree Longitude, Latitude-min Longitude, Latitude-sec Altprof Minelevation, MaxElevation Mindepth, Maxdepth	Min_max_depth_reversed Georeference_uncertain Altitude_out_of_range Altitude_non_numeric Inverted_coordinates Negated_longitude Negated_latitude Zero_coordinates Altitude_in_feet Depth_in_feet	Accuracy
Scientific Name Family Genus Species Subspecies Scientific Name Author Identified By	Uncertainty_not_specified Identification_uncertain Missing_taxonrank Unknown_kingdom Ambiguous_name	Accuracy / Reputation

Table 2 - Type of errors found in data set of botanic.

c) ***Identification of the most used attributes*** – allows the definition of the priority to take preventive measures in order to avoid entry of errors in the database. The quantitative evaluation is also important for the choice of the method to be used in the data cleaning activity, due to the amount of records.

Attributes	Percentage %
CatalogNumber	98,4
Collector	99,67
Scientific Name	99,52
Family	99,52
Country	99,46
Year Collected	98,3
Locality	97,6
Altprof	96,61
Genus	96,21
Month Collected	93,75
Day Collected	92,83
Notes	91,51
Identified By	89,61
Species	88,96

Maximum/Minimum Elevation	85,78
Maximum Depth	85,78
Scientific Name Author	82,27
Year Identified	72,1
Latitude/Longitude - m	23,37
Latitude/ Longitude - s	23,37
Latitude /Longitude - deg	23,33

Table 3 - Principals attributes of herbarium Botanical Garden Rio de Janeiro.

d) ***Classification of error types*** – for the most used attributes, according to the classification of Wang (2000), Dalcin (2004).

e) ***Definition of the method to be used in the process of data cleaning*** – from the survey of the problems identified in the database, a detailed study was performed to define techniques to be used in the cleaning process, observing the previously checked error categories.

The major activities to be performed with such an objective are presented in Figure 3.

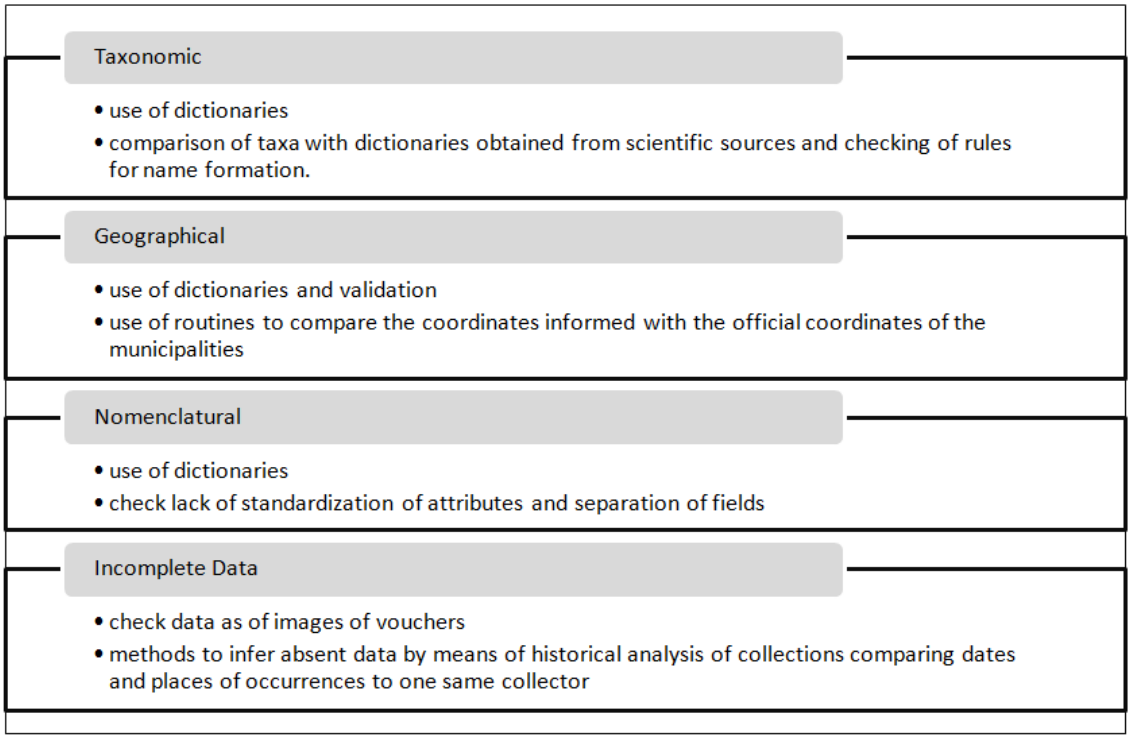

Fig. 3 – Study of methods used in the activity of data cleaning.

5. Results and Solutions Adopted

For data duplicity problems, there were developed several scripts to consult the database; for instance, checking duplicities at the several taxonomic levels, with variation only in the name of the author. Another example is that of a function that

replaces the taxa having duplications with the names of authors derived from the Brazilian Flora List (Flora, 2012), which is adopted as taxonomic standard.

Among the methods evaluated, Expressões Regulares (regex) was responsible for most of the operations to identify errors and correct them, using its resources to build complex queries. Allows regular research of characters in chain through agreed syntax and a set of metacharacters defined in POSIX 1003.2 (POSIX, 1994) and is widely used in applications in various areas of research, particularly in database.

Many errors are due to authors' names written differently to the same author, leading to duplication of the taxon, differing only in the author. The diversity of writings of author names was one of the most difficult problems to solve. An example of this case is as follows: Abarema laeta (Benth.) Barneby & J.W.Grimes and Abarema laeta (Benth.) Barneby & J.W. Grimes.

According to the taxonomic rules, an author's name with a dot should not have a space after it. Therefore, the correct name for the above cases is the second one. The lack of standardization also applies to the collector. These types of errors can be identified according to taxonomic rules. To reduce this problem it was developed a regular expression to validate names of authors, according to the rules of taxonomy.

Table 4 presents errors, attributes and solutions adopted to errors taxonomic.

- Category of the error: accuracy. There is a total of 57,535 taxa in the base, with 92,291 changes.

Error Found	Attribute	Solution Adopted
• Duplicity	• Family • Genus • Species • Subspecies	• Check if it is from the same family, genus, species
• Duplicity	• State Province • Country	• Removal of duplicated names
• Lack of standardization of spaces between names and parenthesis • Lack of punctuation in abbreviations • Lack of standard in capital and small letters. • Name order • Duplicated names	• Scientific Name Author • Identified By • Collector	• Standardization of spaces • Standardization of punctuation • Change all characters to small letters and then change only due characters to capital letters • Prioritize abbreviations. • Remove duplicated names.

Table 4 - Tables with taxonomic data results.

In Table 5 presents errors, attributes and solutions adopted to errors occurrence.

- Category of error: accuracy and reputation. Total corrections performed was 113,569 registrations.

Error Found	Attribute	Solution Adopted
• Yearcollect: invalid date • Daycollect: invalid date (Day greater than 31) • Monthcollect: invalid dates (month greater than 12) • Invalid dates (Non numeric characters).	• DayCollected • YearCollected • MonthCollected	• Leave the Day Field blank. • Carry out the check by the image of the testimony • Numerals and texts were converted to numbers • They followed the standardization to format dd/mm/yyyy
• Numeric and text characters in the same field • No measurement unit • Invalid measurement unit. Very high values (>2000m)	• AltProf: altitude and depth	• Separation of characters • Standardization of all to meter • The measures that had **f** or **t**, were converted to *feet* and the rest to meter • Check collections in the same location and find correct altitude.

Table 5 - Table with occurrence data results.

6. Conclusions

Database of scientific collections has special features that facilitate the occurrence of errors, by the very nature of names, historical data, which may further increase if used in the process of adding data users without specific knowledge of the area. Despite the difficulties, control of data quality must be continuously monitored by specialists to encourage the safe production of research from these data.

This study evaluated the level of data quality and proposed a methodology to improve data quality in the database of the Botanical Garden of Rio de Janeiro. As suggested methodology we evaluated that any work on data quality should be developed after a database project and the definition of the standard metadata to be used, as this facilitates the definition and implementation of rules and the location of errors. The other steps, more related to the control of data quality, are characterized by assessing the most used attributes and the development of routines needed for identification of key errors.

Another part of the methodology, still under study, evaluates the implementation of routines for the identification of errors in large volumes of data; for example, incorrect coordinates for a particular location. At this point, for the identification of outliers the use of algorithms available in data mining may be most suitable.

Contributions to the work as much as possible; we can highlight the promotion of a discussion from the moment that there is experience in databases of scientific collections is shared with other institutions. This research focuses primarily on issues related to data quality bases in the flora and the impact of the presence of errors in conservation work.

7. References

Atlas of Living Australia - Data Quality Portal. (2012). Retrieved from http://code.google.com/p/ala-dataquality/

Chamberlin, Donald D; Boyce, R. F. (1974). SEQUEL: A Structured English Query Language. *Proceedings of the 1974 ACM SIGFIDET Workshop on Data Description*, 249–64.

Chapman, A. D. (2005a). *Uses of Primary Species-Occurrence Data, version 1.0.* Copenhagen. Retrieved from http://www.gbif.org/orc/?doc_id=1300

Chapman, A. D. (2005b). *Principles and Methods of Data Cleaning – Primary Species and Species*. Copenhagen. Retrieved from http://www.gbif.org/orc/?doc_id=1229

Chapman, A. D. (2005c). *Principles of Data Quality*. Retrieved from http://www.gbif.org/orc/?doc_id=1229

CNCFlora. (2012). The Official Brazilian Plants Checklist. Retrieved from http://cncflora.jbrj.gov.br/?q=lista_vermelha/redlisting

Dalcin, E. C. (2004). *Quality Concepts and Techniques Applied to Taxonomic Databases*. University of Southampton. Retrieved from http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf

DC. (2009). Darwin Core. *Darwin Core Task Group*. Retrieved from http://rs.tdwg.org/dwc/

Flora. (2012). Lista de Espécies da Flora do Brasil. Retrieved from http://floradobrasil.jbrj.gov.br/2012

Gonzalez, M. S. (2009). Quantificação de Custo e Tempo no Processo de Informatização das Coleções Biológicas Brasileiras. *Rodriguesia*, *60*, 711–721. Retrieved from http://rodriguesia.jbrj.gov.br/FASCICULOS/rodrig60_3/014-09a.pdf

IEEE Standard for Information Technology-Portable Operating System Interface. (1994). Retrieved from http://standards.ieee.org/findstds/standard/1003.2d-1994.html

Melton, J., Simon, A. R. (2002). *SQL:1999: Understanding Relational Language Components*. Morgan Kaufmann.

Pipino, L. L., Lee, Y.W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*, 211–218. Retrieved from http://dwquality.com/DQAssessment.pdf

Santana, F. S., Siqueira, M. F., Saraiva, A. M., & Correa, P. L. P. (2008). A reference business process for ecological niche modelling. *Ecological Informatics*, *3*(1), 75–86. doi:10.1016/j.ecoinf.2007.12.003

Silva, L. A. E, Barros, R. O., Dalcin, E. C., Zimbrão, G. S., Souza, J. M. (2010). Abordagem Colaborativa para a Melhoria da Qualidade de Dados em Bases de Dados Botânicas. *II Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, Belo Horizonte. XXX Congresso da Sociedade Brasileira de Computação - Computação Verde: Desafios Científicos e Tecnológicos*. Belo Horizonte.

Sodré, F., Fernandes, R. A., Moraes, M. A., Pougy, N.M., Caram, J. S., Dalcin, E. C., Martinelli, G. (2012). Spatial Data Quality of Herbarium Datasets and Implications for Decision-Making on Biodiversity Conservation in Brazil. *Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Florianópolis. Retrieved from http://www.spatial-accuracy.org/BarrosAccuracy2012

Strong, D. M., Y. W. Lee, R. Y. W. (1997). Data Quality in Context. *Communications of ACM*, *40*, 103–110. doi:10.1145/253769.253804

Teorey, T. J, Lightstone, S. S., Nadeau, T., Jagadish, H. V. (2011). *Database Modeling and Design: Logical Design* (p. 352). Morgan Kaufmann.

Wang, R.; Ziad, M.; Lee, Y. W. (2000). *Data Quality*. Kluwer Academic Publishers.