Assessing data quality in Open Data: A case study
Mahecha Moyano John Ferney, López Beltrán Nicolás Estefan, Velandia Vega John Alexander
Catholic University of Colombia
Bogotá, Colombia
jfmahecha77@ucatolica.edu.co
nelopez35@ucatolica.edu.co
javelandia@ucatolica.edu.co

Abstract—This article focuses on measuring the data quality extracted from the API www.datos.gov.co in the area of contracts. This platform allows public entities obey law 1712 of 2014 [2] but this does not have a validation system of minimum data quality. Three metrics are taken as a reference: completeness, traceability and compliance. The measuring of the metrics is done with a software called RapidMiner. This software allows to do data mining, in this specific case the processes for the measuring of the data stored are done and thus determine if the platform www.datos.gov.co has problems. For the programming of the prototype spiral methodology [7] is used. Calculations is created in each phase. Within the data found, serious inconsistencies are found within the platform and within the data, recalling the law 1712 of 2014, it says that the public data must be complete, which shows a breach of it.

Key words - Open data, Data Quality, Traceability, Completeness, Compliance, Software architecture and Business process model.

I. INTRODUCTION

The companies in their daily operations increasingly manage more information, this management forms a growth in the data in an uncontrolled way [1]. In Colombia is created law 1214 of 2014, this talks about transparency of the information, where public entities are required to have public information available to consult [2]. To obey law, MINTIC created a portal called www.datos.gov.co with the objective of data management, but this does not have a validation system of the information is included in the system, it is possible to upload any data set without any restriction, therefore it is not known if the data fulfill the quality standards [3]. There are metrics to calculate data quality, for this study three metrics are used: completeness, traceability and compliance, for that a prototype is developed that allows to calculate data quality obtained from the platform with the three metrics. The datasets evaluated to have related information to contracts.

II. FUNDAMENTALS

A. Open data

Open data is the data that can be used, reused and freely redistributed by anyone [8]. Open data refers not only to digital data, but also to physical data, since there is information printed in files and libraries created in the absence of technology or situations that did not apply such use [9].

According to the Open Knowledge Foundation manual, the fundamental characteristics of open data are as follows [8]:

- Availability and access: By internet preferably, it is given the opportunity to have the same information to the whole world, and preferably it is allowed to be modified. This strengthens transparency in different sectors, such as economic and political.
- Reuse: In addition to a redistribution, combined with other files.
- Universal Participation: Anyone in the world can use the information.
- Real-time processing: Normally through the cloud [9] [10].

One of the most important reasons for open data is to look for a property, interoperability, what is this? Is a skill that allows us to integrate different databases, which allows different systems and organizations to work together [8].

B. Data quality

It is said that data are of high quality when they are "fit" for use in their operational intent and decision making. Quality is also defined as the conformity of the standards that have been established [11].

There is currently a standard called ISO 8000. It gives metrics to evaluate data quality. Data quality is defined by 5 properties [11] [12]:
- Relevance
- Accuracy
- Opportunity
- Comparability
- Integrity

C. Web service

The business intelligence is responsible for the treatment of the proportioned data, however any person has to use some technology to transfer that information. REST is an architecture that uses different standards such as HTTP, XML, URL and HTML for the use of Web Services focusing on the resources of the system [13].

D. Service oriented architecture

It is a Contractual architecture that offers and consumes web services. It is make up of 3 entities: service providers who are the ones who offer the service, companies that use services who
locate them according to the corresponding interest, and service registry [14].

III. THE METHOD

A. Overview

The method used to assess the data quality on open data consists of:

- Metric selection: A searching process on indexed scientific databases and an analysis is performed to verify the viability of each metric, since some metrics did not fall within the scope of the project. After the analysis, the timeliest metrics are selected, according to the available data from the repository www.datos.gov.co.

- Metric definition: In this phase a precise definition is proposed based on the information found.

- Design and code of software prototype: Applying the technologies selected for the calculation of the metrics, followed a schedule where times are defined for the presentation of each metric. For the coding is used the spiral methodology, since the calculation of the metrics is iterative [7]. For the development of the prototype is used the following tools: MongoDB, RapidMiner, and Apache Tomcat server.

- Testing of software prototype: JMeter is used for prototype testing. For measurements stress tests and response time tests are applied.

IV. DATA EXTRACTION MODEL AND METRICS DEFINITION

The following are presented the data extraction process and the implemented metrics that allow the data quality assessment:

A. Extraction of information

The first module of the prototype is the extraction of the data. It is developed in JAVA and this consists of read sequentially one or more datasets, extracting records and information from the www.datos.gov.co and storing it in 2 collections in MongoDB. This extraction is done by means of an API called SODA. For this particular case, a validation of the fields is made because the prototype only evaluate contracts information.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject of control</td>
<td>String</td>
</tr>
<tr>
<td>Event</td>
<td>String</td>
</tr>
<tr>
<td>Type of register</td>
<td>String</td>
</tr>
<tr>
<td>Contract ID</td>
<td>String</td>
</tr>
<tr>
<td>Contractor ID</td>
<td>String</td>
</tr>
<tr>
<td>Contractor name</td>
<td>String</td>
</tr>
<tr>
<td>Project ID</td>
<td>String</td>
</tr>
<tr>
<td>Project name</td>
<td>String</td>
</tr>
<tr>
<td>Project sector</td>
<td>String</td>
</tr>
<tr>
<td>Project value</td>
<td>Number</td>
</tr>
<tr>
<td>Value executed in the project</td>
<td>Number</td>
</tr>
<tr>
<td>Contract objective</td>
<td>String</td>
</tr>
<tr>
<td>Subscription date</td>
<td>Date</td>
</tr>
<tr>
<td>Start date</td>
<td>Date</td>
</tr>
<tr>
<td>Estimated period</td>
<td>Number</td>
</tr>
<tr>
<td>Contract value</td>
<td>Number</td>
</tr>
<tr>
<td>Hiring process</td>
<td>String</td>
</tr>
<tr>
<td>Typography</td>
<td>String</td>
</tr>
</tbody>
</table>

Once the data are store in MongoDB through RapidMiner, the processes are created for do the calculations of the metrics used. The metrics and are as follows:

- **Traceability**: Traceability indicates the presence or absence of metadata associated with the process of creating and updating a dataset [4].

- **Completeness**: Completeness is understood as the measurement which the data are complete and this are of sufficient breadth and depth for the task in which they are used [5].

- **Compliance**: Compliance is the ability of data to adhere to standards, conventions or regulations in laws and similar requirements relating to functionality, reliability, usability, efficiency, maintainability, effectiveness, productivity, safety and satisfaction [6].

B. Traceability

For the development of the traceability metric 2 formulas are applied, they are the following [4]:

- \( tc = 2s + dc \): The formula measures the track of creation. It consists of verifying that there is related registered information to the creation of a dataset, such as the date of creation and source.
- \( tu = lu + du \): The formula measures the track of update. It consists of verifying that there is related registered information to the update of a dataset. It is compound of the last update date and the list of updates to the dataset.

The process for calculating the metric in RapidMiner is follows:

---

**Figure 1. Process for calculating the traceability metric.**

1. **Reading database**

The process begins with the reading of MongoDB database, the data with which the traceability metric works is the metadata associated with the dataset. In this operator the connection to the database, the collection to be used, the host and the port are
configured. The result of the operator are the records of the collection in JSON format.

2. Information conversion

The next step in the process is to convert the information collected in the previous step into a format that allows to perform operations from RapidMiner, organized by rows and columns.

3. Generate attributes

In this operator, a generation of attributes is performed for each record that is retrieved from the collection. The operation to be performed is to validate if there is a creation date and the source that created the dataset, this is done by checking if the field is empty or has the symbol ‘?’. This symbol is automatically assigned by RapidMiner to display empty fields. If the field is empty the value that takes the variable is 0, else the variable will take the value of 1. With these data it is possible to apply the first formula, in this case applies the formula for each record. The next step is validates if the last update date exists and the list of updates to the data set. The way to validate the information is the same, if the last update date exists and the list of updates, the variable takes the value of 1, otherwise the value will be 0. It is now possible to apply the second formula and obtain the 2 expected results.

4. Operation

This operator allows to perform mathematical operations established by RapidMiner, such as counting records, concatenation of values, medians, minimums, maximums, averages, variances, among others. In this case the average to the track of creation and track of update are calculated, these values will be used to obtain the percentage to facilitate the analysis of the information.

5. Second generation of attributes

In the second generation of attributes the percentage of the track of creation and track of update with the percentages obtained previously are calculated. After this the output of the process and the result of the traceability metric are gotten.

6. Select attributes

In this operator, the output attributes for the metric is selected, since the process works with all the fields of the collection retrieved at the beginning of the process and with new fields created through of the process. The percentage of the creation trace and the percentage of the update trace are selected.

C. Completeness

For the development of this metric are used these formulas [4]:

- \( \text{NCL} = \text{NR} \times \text{NC} \): Such that ncl is the number of cells, NR the number of rows and NC the number of columns. The result of this formula is used for calculating completeness.
- \( \text{PPC} = (1 - (\text{IC} / \text{NCL})) \times 100 \): Such that PPC is the percentage of complete cells, IC the number of incomplete cells and NCL the number of cells calculated with the previous formula. With this formula the percentage value associated with the completeness metric is get.

- \( \text{PCPR} = (1 - (\text{NIR} / \text{NR})) \times 100 \): Such that PCPR is the percentage of complete rows, NIR the number of incomplete rows and NR the number of rows. The result of applying the formula is a percentage value associated with the completeness metric.

The process for calculating the metric in RapidMiner is follows:

- Read MongoDB
- Convert JSON to Data
- Generate Attributes
- Operator
- Generate Attributes
- Select Attributes
- Generate Attributes
- Output

Figure 2. Process for calculating the completeness metric.

7. Reading database

The process begins with the reading of MongoDB database. The data with which the completeness metric works is the data associated with contracts. In this operator the connection to the database, the collection to be used, the host and the port are configured. The result of the operator are the records of the collection in JSON format.

8. Information conversion

The next step in the process is to convert the information collected in the previous step into a format that allows to perform operations from RapidMiner, organized by rows and columns.

9. Generate attributes

In this operator, a generation of attributes is performed for each record that is retrieved from the collection. The operation to be performed is to validate if each cell has a value, this is done by checking if each cell in the rows is empty or had the symbol ‘?’, this symbol is automatically assigned by RapidMiner to display empty fields. If the field is empty the value that takes the variable is 0, else, the variable will take the value of 1. This does it for each attribute. Finally, the value of the result of each field is added and with this is obtained by row the number of empty cells. In this operator is assigned the value 24 to another field, this is the number of columns in the datasets. This value is static, since it is defined from the beginning what would be the datasets to evaluate and they have this particular structure. Following this there is another generate attributes operator, only that it fulfills a different function. In the first generate attributes is generated a field which has the number of fields that have an empty data in that row, in this new operator is verified row by row if this field which has the number of empty cells is greater to one, if so, the value 1 is assigned to the variable, else the value is 0.
10. **Operator**

In this operator the last important variables are defined for the calculation of the metric, such as the sum of the rows with one or more empty cells, the number of rows and the number of columns.

11. **Second generation of attributes**

In this operator, the three formulas that compose the metric (number of cells, percentage of complete cells and percentage of complete rows) are calculated, using the variables generated previously.

12. **Select attributes**

This is the last operator in the calculation of the metric. It allows to filter the fields that are to be generated in the output, there are many fields that are not needed such as number of columns, incomplete columns, among others. The output is the percentage of complete cells and the percentage of complete rows.

13. **Compliance**

For the development of this metric the following formulas are used [4]:

- \( PSC = \frac{NS}{NSC} \times 100 \): Such that PSC is percentage of standardized columns, NS the number of columns with an associated standard and NSC the number of columns that meet the standard.
- \( eGMS = S + DC + C + T \): Such that eGMS is a standard that defines if the most relevant data of the information are present, S is the source, DC is the date of creation, C is the category and T is the title.

The scheme that is developed in RapidMiner for the calculation of this metric is the following:

![Process for calculating the compliance metric.](image)

14. **Reading database**

The process begins with the reading of MongoDB database, the data with which the traceability metric works is the metadata associated with the dataset. In this operator the connection to the database, the collection to be used, the host and the port are configured. The result of the operator are the records of the collection in JSON format.

15. **Generate attributes**

In this operator, a generation of attributes is performed for each record that is retrieved from the collection. The operation is to validate whether the source or author who created the dataset, the date of creation, the category of the dataset, and the title exists. This is done by checking if the field is empty or has the symbol '?', this symbol is automatically assigned by RapidMiner to display empty fields. If the field is empty the value that takes the variable is 0, else the variable will take the value of 1. With these data it is possible to apply the second formula, in this case applies the formula for each metadata record corresponding to each dataset. After is calculated the number of columns that the collection of contracts has, since the number is always constant due to the data model, the value 24 is assigned.

16. **Operator**

In the operator the last calculations of variables necessary for the application of the first formula are realized. The average number of columns that meet the associated standard is calculated. This value is obtained in the extraction of the data by means of a script that evaluates if the data input correspond with the type of data that is assigned for the column. The average of the values obtained in each row of the EGMS standard formula is also calculated.

17. **Second generation of attributes**

At this point all the variables required for the application of the formula are known, in the second generation of attributes are calculate the percentage of columns that comply the associated standard and the percentage of datasets that comply with the EGMS standard.

18. **Select attributes**

This is the last operator in the calculation of the metric. It allows to filter the fields that are to be generated in the output, there are some fields that are not needed such as number of columns. The percentage of columns that meet the associated standard and the percentage of datasets that comply with the EGMS standard is selected.

19. **Prototype**

The prototype is developed with 3 technologies, MongoDB to store the data, JAVA to create the web interface and RapidMiner for the calculation of data quality.
The prototype is the integration of MongoDB, Rapidminer and JAVA. JAVA consumes two services, the data extracted from the repository www.datos.gov.co and the embedded data provided by Rapidminer for the calculation of metrics.

VI. RESULTS AND ANALYSIS

The results obtained by the process are processed in RapidMiner server. This server offers different types of graphs and tables focused to facilitate the construction of applications based on the results of the processes generated. As the objective is to create a prototype only the graphs generated are taken to include them in the development of the prototype.

For the evaluation 6 datasets are uploaded. The identifiers for each dataset of the repository www.datos.gov.co are j3bg-66aw, whvw-q2qr, unsz-yhtr, 4v3y-ijrt, h6cd-ixp7, and jghv-pcam.

1. Traceability

The results obtained at the end of the traceability calculation process are as follows.

<table>
<thead>
<tr>
<th>TRACK OF CREATION</th>
<th>TRACK OF UPDATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>50%</td>
</tr>
</tbody>
</table>

The first formula represents the presence of metadata associated with the creation of dataset. It is found that 100% of the metadata associated with the creation of the evaluated datasets are existing, this is positive because if anyone wants to obtain the information about the dates of creation and who uploaded the information, any person can obtain it.

The second formula represents the presence of metadata associated with update of the dataset. It is found that only 50% of the metadata associated with the update of the datasets evaluated have information and 50% do not have it, the common factor that datasets have is the presence of the date of update of the datasets and the absence of the list of updates made, this is negative because if any person wants to analyze or to do a monitoring with the updates made, he or she will cannot do it.

Another process that is not possible to do to the absence of a list of updates is to do estimates according to future data with business intelligence processes. The final value of the traceability metric is 75%.

2. Completeness

The results obtained at the end of the completeness calculation process are as follows.

<table>
<thead>
<tr>
<th>PERCENTAGE OF COMPLETE CELLS</th>
<th>PERCENTAGE OF COMPLETE ROWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>89.979%</td>
<td>0.620%</td>
</tr>
</tbody>
</table>
The first formula represents the percentage of cells that have non-zero data and that provide information to the records that belong. 89.979% of cells have a value, while 10.021% are invalid empty cells. The value given by the calculation means that around 9/10 of the fields have a data, however, the fact that there are empty fields gives rise to incomplete or unreliable information for certain situations, which is a negative and for to get better.

The lack of data also generates questions about the integrity of the information provided and compliance with the quality of the information recorded in Law 1712 of 2014 [2]: “Principle of the quality of the information. All information of public interest that is produced, managed and disseminated by the obligated subject must be timely, objective, truthful, complete, reusable, usable and available in formats accessible to applicants and interested parties, taking into account the procedures of document management of the respective entity.”

The second formula represents the percentage of rows that have non-null data and provide information. Only 0.620% of rows have a value, while 98.380% are invalid or empty rows. It is a very poor value, there is not one average complete column in all datasets, which ensures that information is always incomplete and that there are fields that organizations prefer not to enter and the lack of regulation of the information entity which is being entered. Equally it is not possible to make statistics, averages or analysis with the information with the voids that present the repositories, which questions the compliance of the law.

The final result of the completeness metric is 45.296%.

3. Compliance

The results obtained at the end of the compliance calculation process are as follows.

The first formula represents the percentage of standardized columns in the datasets. None of the columns complies with the standardization of columns, which confirm that the repository www.datos.gov.co does not care about this factor and there is not control in the format with the information and the amount of empty fields.

If a person wants to make data models is difficult, this is related to that there are many empty fields, working with this information can become an impossible task. The entity that uploads this information may be incurring false or misidentified information, which can generate to legal problems.

The second formula represents the percentage of metadata records that meet the eGMS standard. The application shows that 100% of the records have fields governed by the standard. It shows that the data repository www.datos.gov.co is focused on the eGMS standard, which consists of the registration of sources, creation date of repository, category, title and description.

The final result of the compliance metric is 50%.

VII. SOFTWARE TESTING

The prototype is tested with stress tests and is obtained that the server responded well to the requests made. The initial scenario is 500 requests in 1 minute to the module of the evaluation of the metrics with the 6 datasets existing in the database. The efficiency is 500,676 requests per minute.
It can be seen the response time according to the number of requests. The response time varies between 1 and 3 milliseconds, where in 3 milliseconds it averages 8 solved requests and in 1 millisecond it averages 2 solved requests.

The structures of datasets obtained from www.datos.gov.co that contain information on contracts are not homogeneous, which makes it difficult to analyze the data and the understanding of the information reflected there. For example, the structure used by the Antioquia comptroller in the dataset j3bg-66aw is different to the structure that uses SECOP in the dataset ewm2-yzgs.

The repository www.datos.gov.co does not have a logical order or a structure in the letters and numbers that it composes, the identifiers are a random combination of these. An example of this can be the datasets j3bg-66aw and mpka-keq9. Although the datasets are of the same entity and have the same structure, there is no way to know their category or the source entity of the data set.

The repository www.datos.gov.co does not have a good information base for the public treatment, however there are more metrics for data quality evaluation. It can be appreciated that the datasets evaluated j3bg-66aw, whwv-q2qr, unsz-yht, 4v3y-ijrt, h6cd-ixp7 and jqhv-pcam, do not possess complete, reliable and reliable information, possibly studies and analyzes are erroneous and far from the reality, which means a low quality in the creation of products or provision of services that have as source this repository. For example, the dataset j3bg-66aw in the traceability evaluation obtained a result of 75%, compliance of 50% and completeness of 45.87%, with a general evaluation of the 3 metrics of 56.96%.

RapidMiner is an optimal tool for calculating these metrics, since it offers many facilities to the data analyst to do calculations, to work with columns and rows with the data using the operators or creating "custom" operators, these are developed in JAVA.

IX. REFERENCES