DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Marek Lovčí

Master's Thesis

Computer Science and Engineering
Information Systems
2020/2021

Supervisor:
doc. Dr. Ing. Jana Klečková

Methodology Design for Dataset Quality Assessment

Abstract

This master's thesis examines the evaluation of dataset quality. It summarizes the current standard methodologies in the field and defines the new methodology on their basis. The possibility of automatic classification of dataset quality was investigated in the following section of the work, and an algorithm that met the requirement was proposed. The methodology and classification used to evaluate the catalog's quality using COVID-19 data were demonstrated in the final section of the work.

Introduction

The primary objective of the study was to create a methodology that would allow for a universal structured procedure for evaluating datasets. The second task was to test and evaluate the possibility of automatic classification of selected data sets in terms of quality.

Background

Many studies have been conducted on the subject of methodology for data set quality analysis. These studies consider a variety of perspectives, as the requirements for methodologies and quality dimensions differ depending on the field for which the methodology is proposed.

The first factor to consider when analysing data quality is quality dimensions. These dimensions (timeliness, uniqueness, validity, etc.) must be chosen and defined in accordance with the problem's field. Although there are some dimensions definitions that can be interpreted as generally valid, this is not always the case.

The second issue to consider is determining how to evaluate the dimensions. Is it better to have a qualitative or quantitative evaluation? This question is directly related to the current state and structure of the data with that we are working with. Is the raw data available to us, or has it been processed in some way? These issues will severely limit the possibilities for analysis, dimensional quality selection, and, to a lesser extent, methodology selection for ensuring or evaluating dataset quality.

We referred to methodologies that were focused on quality evaluation regardless of field. Despite their differences, they were all very similar. As a result, these methodologies can be summarized into four steps:

- *state reconstruction*,
- *measurement*,
- *assessment* and
- *improvement*.

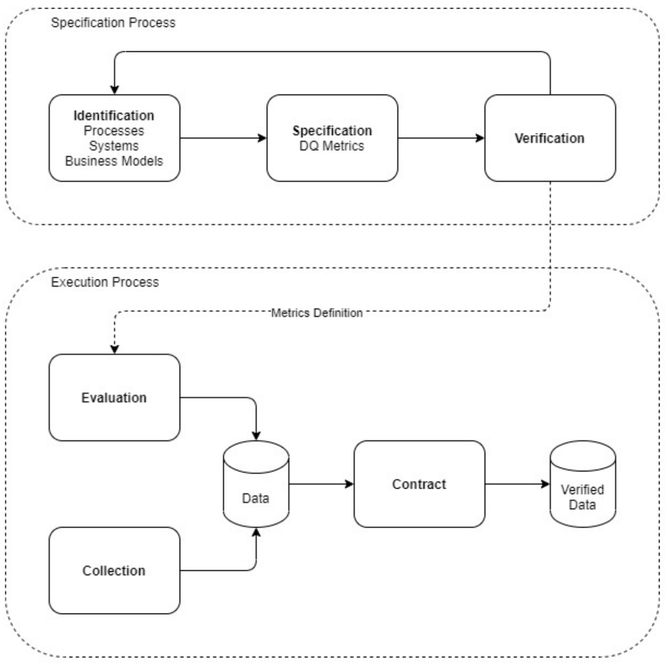
Methodology

We proposed a general methodology for evaluating the quality of data sets as part of the work's implementation. Three use cases for this methodology were proposed, depending on the degree of centralization of responsibility for data management.

These three use cases are:

- Enterprise *Information* Systém
- IoT *Cluster* and
- Open *Data* Library.

The research also included two additional quality assurance mechanisms. Proof of Constancy, which ensures that data is up to date, and Proof of Trust, which ensures that no defective or malicious data is introduced into the dataset.



Proposed methodology metamodel

In the penultimate section of the work, a system for semi-automatic dataset classification was designed.

Results

In the final chapter of the study, we used the methodology and system for dataset classification on the COVID-19 data catalog.

The data were assessed using a data quality questionnaire developed in 2004 in collaboration with the Canadian Institute for Health Information. This list was created for the purpose of evaluating datasets from the healthcare environment, so it was ideal for our situation.

The questionnaire contains approximately 90 questions that reflect 24 data quality characteristics across six dimensions (Relevance, Accuracy, Timeliness, Accessibility, Interpretability, and Coherence).

We were able to obtain a specific number expressing the qualitative status of the data catalog by applying the methodology and the questionnaire to the data in conjunction.

Conclusion

As a result of the work, we became aware of a number of issues that could arise during the evaluation of data quality. Specifically, problems relating to the nature and purpose of datasets.

However, we were eventually able to develop a structured and repeatable method for evaluating COVID-19 data, which was an unwritten requirement of the study.

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA