DATA QUALITY ISSUES IN GEOSPATIAL DATA AND THEIR EFFECT ON GOVERNMENTAL DECISION MAKING PROCESS

BERK BAŞ

SEPTEMBER 2024

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

M.Sc. Thesis in

COMPUTER ENGINEERING

DATA QUALITY ISSUES IN GEOSPATIAL DATA AND THEIR EFFECT ON GOVERNMENTAL DECISION MAKING PROCESS

Berk BAŞ

SEPTEMBER 2024

ABSTRACT

DATA QUALITY ISSUES IN GEOSPATIAL DATA AND THEIR EFFECT ON GOVERNMENTAL DECISION MAKING PROCESS

BAŞ, BERK

M.Sc. in Computer Engineering

Supervisor: Assist. Prof. Dr. Abdül Kadir GÖRÜR

September 2024, 72 pages

Today, geographical data is used in various fields from municipalities to cartography, from cartography to cadastre, and analyzes based on geographical data are extensively used in decision-making processes so that today 78% of data using in data analysis is comprised of geospatial data and geographical data analysis is considered an indispensable part of a good management by United Nations and the World Bank

In governmental planning, that reliable and accurate decision is taken naturally depends on properly building and applying of decision making processes. There is no doubt that basis of the build process is the data that is used.

In today's World, In order to carry out data quality works properly, ISO 8000 and ISO 19157 standards have been developed as a frame and in this study it is discussed that how vector data maintaining in spatial databases can be more qualified through blending these standarts and root causes drive low quality data in the light of Turkish public institutions' dynamic

Keywords: Data Quality, ISO 8000-61, ISO 19157, Geospatial Data

ÖZET

COĞRAFİ VERİLERDE VERİ KALİTESİ PROBLEMLERİ VE BUNLARIN KAMUSAL KARAR ALMA SÜREÇLERİNDEKİ ETKİLERİ

BAŞ, BERK
Bilgisayar Mühendisliği Yüksek Lisans

Danışman: Dr.Öğr.Üyesi. Abdül Kadir GÖRÜR
Eylül 2024, 72 Sayfa

Günümüzde coğrafi veriler kamuda belediyelerden, haritacılığa, haritacılıktan kadastroya kadar çeşitli alanlarda kullanım alanı bulmakta ve karar alma süreçlerinde coğrafi verilere dayalı analizlerinden yoğun bir şekilde yararlanılmaktadır öyle ki bugün veri analizinde kullanılan verilerin %78'ini coğrafi veriler oluşturmakta, coğrafi veri analizi Birleşmiş Milletler ve Dünya Bankası gibi önemli kuruluşlarca iyi yönetişimin önemli bir bileşeni olarak nitelenmektedir.

Kamu planlamalarında güvenilir ve isabetli kararların alınması bu sayede kaynakların etkili ve verimli kullanımının sağlanması, karar alma süreçlerinin doğru bir şekilde kurgulanması ve uygulanmasına bağlıdır. Bu kurgunun temelinde ise hiç şüphesiz kullanılan veriler ve bu verilerin niteliği bulunmaktadır.

Günümüzde veri kalitesi çalışmalarının doğru ve etkili şekilde uygulanabilmesi için çerçeve niteliğindeki ISO 8000 ve ISO 19157 standartları geliştirilmiş olup bu çalışmada uzamsal veri tabanlarında tutulan vektör verilerin bu standartların harmanlanarak nasıl kaliteli hale getirilebileceği düşük veri kalitesine neden olan kök nedenler ve çözüm önerileri Türk kamu kurumlarının dinamikleri ışığında tartışılmaktadır.

Anahtar Kelimeler: Veri Kalitesi, ISO 8000-61, ISO 19157, Coğrafi Veri

ACKNOWLEDGEMENT

I would like to thank sincerely to my familiy for their endless support to gain many achievements such this. And I dedicate all success I achieved as well as I will achieve to my dearset family.

I would aslo want to thank all instructors of Cankaya University especially to my thesis supervisor Dr. Abdül Kadir GÖRÜR for everything they do.

TABLE OF CONTENTS

STATEMENT OF NONPLAGIARISM III
ABSTRACT IV
ÖZET V
ACKNOWLEDGEMENT VI
LIST OF TABLE X
LIST OF FIGURES XI
LIST OF SYMBOLS AND ABBREVIATIONS XII
CHAPTER I INTRODUCTION 1
1.1 SITUATION OF PROBLEM 3
1.2 SUB PROBLMES OF STUDY 3
1.3 AIM OF THE STUDY 4
1.4 LIMITATIONS 4
1.5 DEFINITIONS 4
CHAPTER II LITERATURE REVIEW 6
2.1 DATA CONCEPT 6
2.1.1. What is Data 6
2.1.2. What is Geospatial Data 7
2.1.3. What is Metadata 7
2.1.4. Data Life Cycle 7
2.1.5. Data Life Cycle 8
2.1.6. What is Coordinate Reference System 9
2.1.7. What is Coordinate Reference System 10
2.2 DATA QUALITY CONCEPT 10
2.2.1. General Quality Characteristics of Geospatial Data 11
2.2.2. Data Quality in Corporate Manner 11
2.2.3. Personally Identifiable Data (PII) and Organizational Data Quality 13
2.2.4. Life Cycle of Data Quality Process 13
2.2.5. Life Cycle of Data Quality Process 14

2.2.6. Importance of Metadata on Data Quality 17
2.3 DATA QUALITY MANAGEMENT 17
2.3.1. Adversities in Data Harnessing and Management 18
2.3.1.1. Complex Geospatial Data Ecosystem 18
2.3.1.2. Complex Geospatial Data Ecosystem 18
2.3.1.3. Data Quality and Data Redundancy 19
2.3.2. Human Effects on Data Quality 19
2.3.2.1. How Do Divergent Behavioral Patterns Interact on Data Quality Issues ? 22
2.3.2.2. Individuals As A Part of A Team 23
2.3.2.3. Teams in An Organization 23
2.3.3. Data Quality Management Framework 24
2.4 ISO 8000-61 DQ APPROACH 26
2.4.1. Data Quality Planning 26
2.4.2. Data Quality Planning 27
2.4.3. Data Quality Assurance 29
2.4.4. Data Quality Progress 30
2.4.5. Data Quality Progress 30
2.4.6. Resource Allocation for Data Quality Operations 31
2.5 ISO 19157 DQ APPROACH 31
2.5.1. Elements of Geospatial Data Quality 32
2.5.2. Metaquality Elements 34
2.5.3. Data Quality Evaluation Methods 34
2.5.4. Data Quality Evaluation Methods 35
2.5.5. Reporting Geospatial Data Quality 36
2.6 ISO 19157 DQ APPROACH 37
CHAPTER III STUDY'S METHOD, MODEL, DATA AND CASE STUDY ... 38
3.1 MODEL OF STUDY 38
3.2 DATA QUALITY WORK ON SAMPLE DATA SETS 38
3.2.1. Profiling and Data Quality Improvement Study 38
3.2.1.1. tblpoints Table Profiling Process 39
3.2.1.2. tblline Table Profiling Process 43

3.2.1.2. tblpolygon Table Profiling Process 46
CHAPTER IV RESULTS 50
REFERENCES 53

LIST OF TABLE

Table 1: Table of Real Estates' Taxes..... 14
Table 2: tblpoint Table's Data Statistics. 41
Table 3: tblline Table's Data Statistics. 42

LIST OF FIGURES

Figure 1: Data Life Cycle 7
Figure 2: Data Quality Life Cycle 13
Figure 3: Data Quality Dimensions 17
Figure 4: Data Processing 28
Figure 5: Overview of the ISO 19157:2013 data quality elements 33
Figure 6: Geospatial Data Quality Process 35
Figure 7: Query using in analysis 38
Figure 8: Second Satellite view of Çankaya University 39
Figure 9: Analysis results on map display 40
Figure 10: Profiling tblpoint table according to all columns 40
Figure 11: tblpoint table's data profile prior to data processing 41
Figure 12: tblpoint table's data profile after data processing 42
Figure 13: tblline's low quality data on map display 43
Figure 14: tblline table's data profile prior to data processing 43
Figure 15: Usage of ST_AsText command 44
Figure 16: Result of ST_AsText command 44
Figure 17: tblline's corrected data on map display 45
Figure 18: tblline table's data profile after data processing 45
Figure 19: Sample polygon data from tblpolygon 46
Figure 20: Result of SQL query designed to find overlapping polygons 47
Figure 21: tblpolygon table's data profile prior to data processing 47
Figure 22: tblpolygon table's data profile after data processing 48

LIST OF ABBREVIATIONS

CCPA	: California Customer Protection Act
CDO	: Chief Data Officer
CRS	: Coordinate Reference System
DBMS	: Database Management System
DIGEST	: Digital Geographic Exchange Standart
EU	: European Union
GDPR	: General Data Protection Regulations
IT	: Information Technology
ISO	: International Organization for Standards
KVKK	: Kişisel Verileri Koruma Kanunu
NASA	: National Aeronautics and Space Administration
NATO	: North Atlantic Threaty Organization
NGA	: National Geospatial Intelligence Agency
NIST	: National Institute of Standards and Technology
USA	: United States of America

CHAPTER I
INTRODUCTION

Today, We are in a digital economy era where data is much valuable than ever before. The era we currently in it, data posseses same amount of importance with oil used to have in 18'th century. Data, like oil, holds great rewards for those who learns to extract and use it. Today vast amount of data being produced is geospatial data and geospatial data is the key figure of sustainability and business processes being carried out by public authorities. It is not possible to make progress without it. (Toonders 2017:1)

In today's world, volume of geospatial data increases exponentially and in one way or another numerous governmental or non-governmental organizations invest in geospatial data analysis to benefit from it at a certain level. Since it doesn't require a high budged but provide so much benefit so that today some governmental institutions of Türkiye such as Ministry of Environment, Urbanization and Climate Change or General Directorate of Mapping initiated their own geospatial data analytics programs. On the other hand, As stated in Hahmann and Burghardt's study mostly used and harnessed data type is geospatial data so that 78% of data analytics are carried out via it. In addition to this a study carried out by αlphaβeta revealed that geospatial data services provide $550b austerity per year. Geospatial data is benefited from various fields such as urban planning or reduction of travel timing, health monitoring programs, implementation of transportation infrastructure or monitoring enviromental issues

Although quality term is generally used in ISO 9000 standards to describe quality in industrial production. In same way it can be transcribed into data to ensure that data is fit for purpose. In today's World, although there doesn't exist any missing data pertaining a geolocation but it is need to be underlined that data quality currently being possesed is far belower that It is desired. (Panoho 2019)

Data quality work in geographical manner (a.k.a geospatial data quality study) was first time carried out in 1982 by American Congress of Survey and Mapping and

as a consequence of this study a frame to describe geographical data quality standartsof which elements are lineage, positional accuracy, accuracy of attributes, logical consistency and completion was put forward. But internationally accepted first standart for geospatial data quality is DIGEST (Digital Geographic Exchange Standart) which was adopted by NATO and NIST afterwards. After Eurpean version of standard is developed, ISO announced first geospatial data quality frame ISO 19113 which describes quality principles of geospatial data then sequentially ISO 19114 which describes techincal aspect of quality work and finally ISO 19157 were published. (López et al. 2020:131)

Internationally accepted quality standart for data is ISO 8000. ISO 8000 defines fundamental principles to exchange data without losing its meaning amongst different applications in an application independent manner. Introduction part of ISO 8000 briefly determines standard as both ability of generating, collecting, storing, maintaining, transfering and presenting data and capabilitiy to measure, manage and report data quality. ISO 8000 first announced in 2008 and focuses on sytax, context and master data. ISO 8000 divides data quality work to subdivisions called Data Quality Planning, Data Quality Control, Data Quality Assurance. And these subdivisions are concerned with aspects such as data accuracy, completeness, consistency, and timeliness. (López et al. 2020:130)

Geospatial data quality is mostly about positions, attributes and relationships between entites and features. Main aim of data quality works in spatial data sets is to ensure accurately visualization and analysing locational data. In case quality requirements cannot be met, Need for re-producing associated data is emerged. And this process is a financially costly and time consuming practice. (Devillers et al. 2005)

Reasons lead to low data quality can be enumerated as lack of understanding, bad planning, inconsistent development process, lack of standarts or lack of good management. Also it needs to be stated that most of organizations cannot define what is fit for purpose properly. (Henderson and Earley 2017:434)

Qualified data is useful one. In order to achieve high data quality, it needs to be cleansed from ambiguities and discrepancies. Data possesses high level of quality can be easily processed and analysed. Thus They can aid foundations to make better decisions. On the other hand, high quality data is essential for cloud migration and artificial intelligence projects. Aftermath of data quality work can solely be measured

through how accurate is analysis of associated data. (King and Schwarzenbach 2020:11)

It needs to be stressed that organisations encounter data quality problems. Since almost none of the them have an acceptable level of data management pratics. But, organizations that focus on data quality and implement data quality practices in their daily routines experience fewer problems than those that leave data quality to chance

Obtaining high quality requires high level of coordination and communication amongst departments, people and processes. Organisations need to be avare of this circumstance and take into consideration of risk for low level data quality and set plans to overcome this problem.

Data quality management is akin to other kind of quality management pratics. More clearly. It includes determining quality standarts and a life cycle management including creation, manipulation and storage of data. (Henderson and Earley 2017:182)

In our country, Ministry of Environment, Urbanization and Climate Change is responsible to ensure uniformity in data quality for institutions that maintaining geospatial data. As these institutions are to subscribe TUCBS (Turkish National Geographical Information System – TR: Türkiye Ulusal Coğrafi Bilgi Sistemi) which is created based on EU's INSPIRE Directive and fulfill necessary requirements. (Alan 2022:37)

1.1 SITUATION OF PROBLEM

In this study, description of geospatial data quality and likely results of low data quality level in geographic analysis is scrutinized. In experimental phase, a sample anonymized data set's quality problems are determined and fixed

1.2 SUB PROBLMES OF STUDY

1- In case a data quality process is determined to be applied. Whether governmental organisations would adopt this pratics.

2- Whether there is a correlation between quality level of data in data silos of governmental organisations and degree of being familiar with concept of data quality for employee of it.

1.3 AIM OF THE STUDY

Aim of this study is to highlight and encourage Governmental Institutions of Türkiye to adopt data quality pratics and some emerging technologies such as DataOps.

1.4 LIMITATIONS

In case study and literature review of this thesis is limited with geographical vector data. Thus, raster data is not examined in this study. In case study section A sample data set is examined to determine and fix data quality issues. Prior and afterwards data quality process data quality is measured via data quality dimensions.

1.5 DEFINITIONS

Accuracy: It expresses degree of matching between a data and entity that being represented by data.(McGilvray 2021:15)

Attribute: Data field storing attributive specifications of data (ISO/TS 21089:2018)

Chief Data Officer: A responsible person at senior level for effectively acquiring, classifying, harnessing and management of data.(Federal CDO Council, 2023)

Conformance: It states fullfilment of a requirement. (ISO 19157:2023)

Conformance Quality Level: "*It states threshold value or a set of threshold values for data quality results used to determine how well a dataset meets the criteria set forth in its data product specification or user requirements*" (ISO 19157:2023)

Consistency: It implies that data resides in different data sets should be matched in terms of format and characterization.(Cichy and Rass 2019)

Data Governance: Design and implementation of data management policies. (ISO 8000-2:2020)

Data Management: All activities using in one or more than one IT systems for description, creation, storing, maintenance and access of data (ISO/IEC TR 10032:2003)

Data Owner: It implies the people responsible for a data set.(King and Schwarzenbach 2020:8)

Data Quality Criteria: Kalitelerini anlamak için verilere uygulanan belli başlı testlerdir. Bunlar ayrıca kalite değerlendirmesinde kullanılan metotları da içerebilirler.(King and Schwarzenbach 2020:8)

Data Set: Data set logically meaningful (ISO 8002:2020)

Data Steward: Staff whom is attained a data source's or data set's responsibility and management. (ISO 8000-2:2020)

Feature: It is stated as abstraction of real world phenomena (ISO 19157:2023)

Feature Attribute: It is stated as characteristics of features. (ISO 19157:2023)

Lineage: It is stated that production process of a data set. (ISO 19157:2023)

Item: It is stated as anything can be described and considered separately (ISO 19157:2023)

Metadata: Data describes other data. (ISO / IEC 11179-3:2013 – changed)

Metaquality: Information describes quality of data quality process. (ISO 19157:2023)

Precision: A data's certainity level. (ISO/IEC 11179-3:2013 - changed)

Structured Data: Data in a data set which fullfils some explicit and pre-determined regulations. For instance, tables, rows or primary/foreign keys in a relational database's table. (King and Schwarzenbach 2020)

Register: Set of files possessing descriptive information pertaining to items (ISO 19157:2023)

Timeliness: Momentarily availability level of a data (Foote 2023)

Uniqueness: Whether there exists a single data entry in a data set (Foote 2023)

Unstructured Data: It implies data sets don't comply any determined data set rules. Such as lidar data files.

Universe of Discourse: It is stated as view of the real or hypothetical world that includes everything of interest (ISO 19157:2023)

Validity: It implies that data is validated through some syntaxial and formal rules. (Foote 2023)

CHAPTER II
LITERATURE REVIEW

2.1 DATA CONCEPT

2.1.1. What is Data

Data is a concept expresses information which is genuine and wasn't processed. Data is acquired or gained through experimental, quantative or qualitative research methods. (Açiler 2020)

Data is essential part of information. Data is transformed into information through classification, summarization and interpretation. Unlike data, information includes interpretations and data consumer integrates information with his or her experiences to convert it into knowledge. The factors make information valuable are information's timeliness, accuracy and being fit for purpose. To be more explicit the example can be given as follows:

- Data: 33
- Context that data can be used within: Author is 33 years old.
- Information that can be elicited from data: He is old enough to get a driving license.(Dülge 2009:4; Talend Team 2024)

Today, Although private institutions and governmental organisations consider their data infrastructure as an expenditure item. As previously stated, data has same amount of worth with oil. For instance: According to a research conducted in Harvard School of Medicine, human pathologs and machine learning systems are compared in terms of diagnosing breast cancer accurately. At first, human pathologs diagnosed 96% of accuracy rate whereas ML systems diagnosed with 92% accuracy. But afterwards, outcomes of human pathologs researchs are given to ML systems as training data and ML systems' accuracy rate increased to 99,5%. (Bhageshpur 2019)

2.1.2. What is Geospatial Data

Geospatial data is a kind of data model that presents location information pertaining to any location on the Earth in terms of longitute and latitude. There exists two types of geospatial data:

- **Vector Data:** It implies that points, lines or polygons placing on a map
- **Raster Data:** It implies that data consisting of pixels a.k.a base map.

Vector data can be represented through divergent projections such as EPSG (European Petroleum Search Group Code): 4326 or EPSG:3857 in divergent data types for spatial databases such as geography or geometry. (CGIAR 2021)

2.1.3. What is Metadata

Metadata is basicly data that describes data. (Ahronheim 1998; Coyle 2005) Metadata provides abudant information pertaining to specifications of data. Thus metadata is an crucial data type allows us to figure out actual context of data.

Along with developing technology amount and volume of data is increased. There exists significant amount of useless and useful data in this large volume of data. Thus the problem to access desired information among these data is arose. Thanks to metada, users can determine whether data they selected is fit for their purpose as well as gain information about context of data and this let them to save time. (Araz 2013)

Metadata is a key concept to carry out a effective data quality management. Quality of data depends on how much data is fit for users' purpose. Metadata describes what data actually represents. Thus having a robust data quality management process is directly related with metadata management.(Henderson and Earley 2017:46)

2.1.4. Data Life Cycle

Data has a life cycle such as other entities in an organisations. Data is acquired, harnessed and after it reaches end of its useful life either it is removed or archieved.

Data quality needs to be managed in each phase of data life cycle from data acquiring to removal of data and when data is in circulation between systems, It is a must to guarantee the maximum level of data quality.

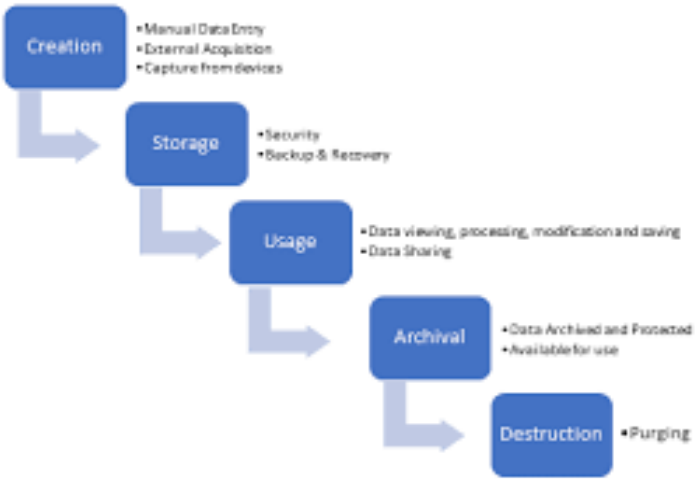


Figure 1: Data Life Cycle (DataWorks 2024)

Creation: It implies either manually or automatically generation of data. Creation phase can be done through humans as well as sensors.

Storage: Created data is stored in data silos of organisations. Organisations need to also ensure security of data at rest and take necessary precautions to avoid data loss.

Usage: Usage phase implies harnessing data to generate visual reports and make analytics along with sharing it with other parties.

Removal: Data has no benefit for organization and out of regulations that mandates data to be stored is deleted in order to reduce hardware expenditures. (DataWorks 2024)

2.1.5. Data Life Cycle

Data is generated in an unprecedented pace and it is much more valuable than ever before. Thus, the World is becoming a data driven place gradually day by day, states and organizations sovereign to data gain much more competitive advantage to those don't.

Producing large volume of data brings organizations new and divergent challenges. Most notable one of these problems is that the data has proper quality level as well as data is accessible by authorized users to obtain qualified organizational outcomes and complies with associated regulatory requirements.(Galdies 2014; Gregory 2011)

At this context approaching data as an organizational asset is an important fact to overcome this problem.

Data needs to be governed and needs to be reached as if they are fiscal assets Thus: Attributes of data can be summarized as

- Data has high value for organisations.
- Data can be subject to quality evaluation.
- Data can be a key attribute to boost organisational performance and aid to make better decisions.
- Data has a life cycle (From obtaining to harness and deletion)
- Proper management of data can reduce business cost

Unlike pyhsical assests data takes role in strategic decision making processes. So misunderstanding of data may lead wrong decision to be taken. (Syafik 2023)

2.1.6. What is Coordinate Reference System

Coordinate reference systems are used to describe a real world location in a coordinate system. They are selected according to data collection method and extent of data.

Coordinate systems can be examined under two categories. These are geographic coordinate system and projected coordinate system.

Geographic coordinate system uses longitude and latitude while stating a location whereas projected coordinate systems use x and y values (cartesian coordinates) Since it transforms 3D Earth sphere into 2D coordinate plane.

Each coordinate system is designed based on an datum. Datum defines the shape and size of the Earth using for accurate positioning and mapping purposes. Each datum has a origin point for instance for WGS 84 datum origin point is Greenwich. (GIS-Geography 2024)

Generally EPSG (European Petroleum Survey Group) codes are used to imply coordinate reference systems. Widely used EPSG codes are EPSG:4326 and EPSG:3857 and EPSG:4326 uses geographic coordinate system and WGS 84 datum. EPSG:4326 is also used by Google Earth and GPS devices whereas EPSG:3857 uses projection coordinate system and it uses metric coordinates. EPSG:3857 is mainly used by OpenStreetMaps project. EPSG:4326 is used when precision and geographic

correctness are crucial, but if performance and compatibility with web mapping tools are important, EPSG:3857 is prefered. (Linz 2024)

2.1.7. What is Coordinate Reference System

DataOPS is a novel phenomenon which is generated on same basis with DevOPS. Briefly DataOPS integrates data engineering, data integration, data management and data quality practices to ensure reliability of data analytics. DataOPS is an automatized process starts with generation of data then continuing with provision of data quality. And transformation of data to analyzable format for ML or Data Analytics projects.

DataOPS provides some benefits as follows:

- DataOPS provides inter-teams communication within organization
- DataOPS provides automation and aids to decrease rate of low quality data
- DataOPS practices aid to establish security and to comply regulations such as GDPR, CCPA or KVKK (Komtas 2024)

2.2 DATA QUALITY CONCEPT

In geospatial data silos, data is collected from various data sources with different techniques as well as these divergent sources may have different data standarts such as EPSG:4326 or EPSG:3857 etc. So this circumstance may lead biases in data analytics and reduce reliability to them. Thus It is expected to implement a quality improvement process prior to use the data for analysis. (Coetzee et al.2020)

In ISO-9000 quality management framework, quality term implies that evaluation of whether an object or a process is fulfilled what is expected from itself. Same logic can be transcribed into term data quality.

Since Data is an asset and has an organisational context and it is often used to support numerous business processes via data analysis. There should be some requirements that data is supposed to meet.

Thus, for data quality, it can be said that measurement of how well data sets pertaining to an organisation fit for its purpose. In same manner data quality can be described as health status of data in a given time of its life time. (Koçak 2023:24)

Geographical data quality is described in ISO 19157:2013 along with principles and evaluation methods and thus geogprahical data quality can be described

as making geographical data suitable for using in temporal or cartographical systems (Alan 2022:6)

The main impact of data quality is to ensure that the right data is available to the right people at the right time to make the right decisions and get the right results. This statement can also be expanded to say that good quality data is data that is secure, legal and processed correctly.

Data quality is a subjective concept. Quality measurement of data can vary from person to person. For instance unless giving details, if one indicated that weather is bad, would have a narrow meaning. On the other hand for anyone else weather may be good. So if we start from these examples, without considering data's context, it would be hard to interpret data quality only regarding someone's declaration on how low data quality is. That's why in order to measure data quality in absolute manner, data quality dimensions are put forward. (Kumar 1998:56)

2.2.1. General Quality Characteristics of Geospatial Data

There are various studies to determine quality characteristics of data. However, none of these studies encompasses all characteristics. Another aspect of the adversity is that people have divergent expertise can evaluate data requirements from seperate point of views. For instance end users intrest in ultime outcome whereas data designers focus on mandatory and optional attributes of data.

All these point of views are put together in ISO 19157 that generates an integrated framework for requirements and characteristics of geospatial data. In addition to this ISO 8000 and INSPIRE Directive are important frameworks can be used for evaluation of data quality.

In order to understand these characteristics more clearly, data quality dimensions were put forward. These dimensions are accuracy (positional), timeliness, consistency, completeness, validity, uniqueness and precision(geometric and semantic precision) (Alan 2022:6)

2.2.2. Data Quality in Corporate Manner

Since it was started using as a governmental activity, IT investment in Turkish government has made significant progress. At the very top of this progress, there is digital transformation of Turkish Government and one of the latest phases of this

digitalization includes big data, data analytics and GIS(Geographic Information Systems).

Inside these each broad category, technologies and approaches are constantly evolving. Each step taken is overcome shortcomings of previous technology or approach.

During all these technologic progress, data appears as a constant factor. It is essential that data mustn't lose its semantic(meaning) during transition from old technology to new one or data transfer projects. However, geospatial data acquiring processes have always included risks and in terms of quality of data. For instance because of a different standarts used, notable amount of data can be dropped and this causes time cost.

In this case, It won't be wrong to claim that data migration projects or data aquiring operations may cause oftenly data quality problems. And after this kind of projects, data quality practices need to be carried out.

On the other hand, since governmental institutions of Türkiye don't welcome cloud computing, data quality problems generally arise from database migrations or human errors. (King and Schwarzenbach 2020:16)

Governmental institutions need reliable data to serve better, make more accurate decisions and create effective politics. Low quality of data causes governmental services to be unsuccesful, problematic decision making processes and unable to produce resolutions for problems.(Government Data Quality Hub 2021)

In our country ,Türkiye, most of governmental organisations don't focus on data quality issue as it takes. For this reason, public institutions need to take a more structured approach to this phenomenon.

Official data quality management resembles ISO 9000 quality management in same manners. Data quality management requires to determine standards and it is ensured that data meets these standards during its life cycle thus qualified and reliable data is generated and this data can be used for strategic goals of organization. Reliable data on the other hand increase efficiency among employees since they spend less effort because when they don't spend time to fix data. (Henderson and Earley 2017:47)

2.2.3. Personally Identifiable Data (PII) and Organizational Data Quality

Another pillar of organizational data quality works is constituted by personally identifiable data and legal regulations covering this data. Since geospatial data may include PII data as attribute, It needs to be examined in terms of PII regulations.

General Data Protection Regulations (GDPR), known as the EU's personal data protection law, have come into force since May 2018. GDPR does not consist of just a set of rules to be followed, it is a measurement and comparison criterion (benchmark) that covers all consumer data and data security approaches and determines how we will approach personal data in the future.(Chasinov 2018)

The emergence of GDPR in 2016 created a snowball effect in data privacy laws around the world. The California Consumer Privacy Act (CCPA) emerged in the US state of California in 2018, right after the GDPR, and came into force in 2020.

GDPR is a package of laws that determine how and why personal data of individuals living in or citizens of the 28 EU member states can be collected. Within the scope of GDPR, companies both in EU member states and outside the EU are updating their privacy policies and data collection practices. (Early 2021)

As a matter of fact, in Article 5 of the GDPR, “All reasonable steps should be taken regarding inconsistent data, and after evaluating the purpose for which it is processed, this data should be deleted or corrected without delay.” (Chasinov 2018).

This data organization also dictates that this information be organized into a consistent and holistic view, as the same type of information may be fragmented in many places or systems within the organization.(Talend Team nd.)

2.2.4. Life Cycle of Data Quality Process

The data quality life cycle consists of five stages. In the first phase, practitioners investigate how desired business processes are affected by poor data quality. This stage is described in Loshin’s book as integration of two approaches called bottom up and top down. Activites in bottom up approach are that what data related activities are carried out by business users and document things they regard notable. Top down approach is to detect potential data anomalies that can be realized by experts of associated field data belongs through statistical methods and data analysis tools.

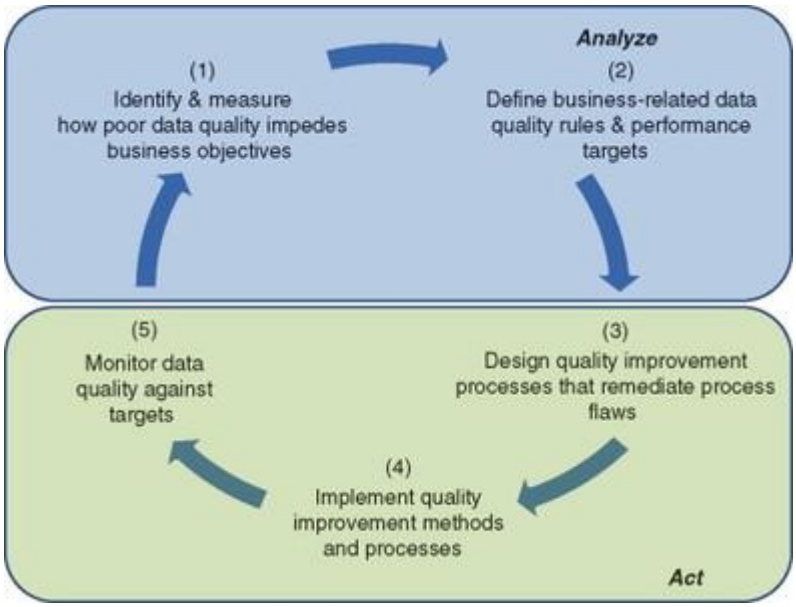


Figure 2: Data Quality Life Cycle (Loshin 2011:18)

In the second phase, data quality analysts synthesize the output of transactional results from both top-down and bottom-up behavioral models and focus on data elements evaluated as critical based on the needs of selected business users. Ultimately, this empirical analysis allows us to measure data quality within a specific business context. During this phase, data quality rules, acceptable threshold values and measurement methods against these rules are also developed.

In the third phase, data quality problems are prioritized according to their degree and possibility of correction. At this point, an action plan is developed and the selected solutions are added to the plan.

As soon as the solutions are determined, the fourth level is applied. Correction procedures at this stage; is data correction and may also include discovery and disclosure activities.

At level five, data quality analysts observe whether the data is at an acceptable level. (Loshin 2011:18)

2.2.5. Life Cycle of Data Quality Process

Data quality dimensions are measurable attribute or characteristic of data. Data quality dimensions are attributes in order to describe quality of data in the frame of given requirements. In order to measure the quality of the data, the relevant data must be both measurable and measurable for the business processes of the organization that owns that data. Therefore, it can be said that data quality dimensions form a basis for

measurable rules that should be directly linked to potential risks in critical processes. For example, suppose that for some land registry records property owner's mobile phone number field is missing thus we cannot send information to the relevant persons via SMS. And as a result of this situation, the citizen may suffer material damage. So for this scenario we are supposed to fill 100% of mobile phone numbers in order to reach desired data quality level. (Henderson and Earley 2017:445)

Numerous researchers studying data quality published their own data quality models but most notable ones among them are Strong-Wang Framework, Thomas Redman Model and Larry English Model.

For instance let's scrutinize data quality dimensions suitable for table of real estates' taxes as follows:

Table 1: Table of Real Estates' Taxes

ID	Geom	Length	Width	Height	Building Color	Num of Floors	Built Date	Estate Tax
10	0A242AF2D2..1	59.5	29.0	29.0	Yellow	-		
12	E24D4DE251..4	59.5	29.6		Blue	N/A	1.09.2001	TRY30
14	DE24214A..21	79.8		9	Brown	10		
15			15	11		4	12.23.91	TRY16
44		47.8	7	9			27.04.2014	TRY7.12
45		60.0	29.4	28.5	Pink		15.07.2015	TRY4.21

Accuracy: Whether data reflects the real word object that it represents. For instance, let's pick ID 010 from table, according to information on the table, estate with ID 045 is yellow and with 60x29x29 dimensons. But if it is actually red and with different dimensions, It is said that this data is not accurate. (Foote 2023:1)

When accuracy is scrutinized in terms of location. We examine x,y,z values that creates Geom (geometry) field to examine whether it points true location.

In addition to this thematic accuracy can be divided into two parts as qualitative and quantitative thematic accuracy. Qualitative thematic accuracy states high value of geographical data. Whereas quantitative thematic accuracy is related with raster data.(Alan 2022:10)

Completeness: Whether all attributes pertaining to an object is filled. For instance for building ID 010, attributes are not completely filled. Similarly, if ID 017 is expected in the table but it is missing, We say table is not complete. Thus we can say that completeness can be measured at table or record level. In order to determine completeness We ask questions of Does the data set contain all expected records? Are the records filled in correctly? (Records with different statuses may also have different

expectations for completeness.) Are the columns filled as expected? (Some columns are required. Optional columns are filled only under certain conditions.)(Buzzelli 2023:15)

Consistency: Geographical data needs to be consistent in terms of time, location and topology which means location used in a Geographical Information System should be the exact location of an entity or phenomenon exist in real world for a given time. If in another GIS, there exists an entity or phenomenon different for a given time, We say consistency is violated. Also consistency needs to be satisfied in terms of edge matching namely topologically. In other words, vector data adjacent to each other such as line, multiline or polygon is expected to touch each other and there shouldn't be any gap between them. On the other hand, attributes of geographical data needs to satisfy consistency as well. So For instance building 012's built date is entered 01.09.2001 in the table. But in population and citizenship system it may be entered as different date. In this circumstance We say data is inconsistent. Additionally, this dimension can also refer to the size and composition of datasets across systems or over time. (Alan 2022:11)

If consistency is sought between two or more than two data in a record (a.k.a. table) that describe same entity this is called consistency at record level. If it is sought between divergent records it is called consistency at inter-records level. Consistency can also be attributed format consistency and It needs to be underlined that consistency isn't truth. (Henderson and Earley 2017:442)

Validity: It expresses whether data is suitable for given format and is measurable via metrics using in evaulation. For instance real estate taxes table includes numerous data formats for date section. This situation violates validity. (Doger 2021:21)

Timeliness: It states that data needs to always be up to date and available. Thus, Timeliness is measurement of data whether it is the most updated version of it.

Intermediate or static data such as code tables or reference tables can stay updated for a long time. However, operational data can stay updated for a relatively short time. For instance stock prices in markets. They are expected to change many times in a day. And if it is updated lately, some investors may be financially damaged. (Henderson and Earley 2017:443)

Uniqueness: Whether there exists only one single representation for each physical entity. Uniqueness can be provided via keys in a database table. As primary

key structure ensures that a record can be recorded only once. If we consider real estate taxes table given before, there doesn't exist any dublicate id number or any data with different ID number qualifying each other so it means all data is unique.

This sample analysis is a starting point of quality works. That's why in order to ensure data to fit for purpose another processes that provide an integrated approach need to be carried out. And in order to capture all necessary requirements to implement a good data quality management these processes need to be implemented from simple to complex one. Uniquness is also crucial for success of AI projects.(McGilvray 2021:38)

Dimension name	Sub-Category	Definition
Accuracy	Syntactic Semantic	Distance between v (the correct value) and v' (the incorrect value)
Completeness		Degree to which all values are present in a data collection
Time-related aspects	Currency	Degree to which the data is up-to-date
	Volatility	Frequency with which data vary with time
	Timeliness	How current the data is for the task at hand
Consistency		Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules
Interpretability		Concerns the documentation including the data model, and other metadata that are available to correctly interpret the meaning of data
Accessibility		Data is accessible for those needing access to the data in a format that can be understood
Quality of information source	Beliveability	Is the data provided true, real and credible?
	Reputation	Is the source normally credible?
	Objectivity	Is the source believed to be objective?

Figure 3: Data Quality Dimensions(Research Gate 2024)

2.2.6. Importance of Metadata on Data Quality

Metadata is a critical phenomenon for data quality management. Since data quality can only be measured through how well data meets requirements of business. Metadata has a critical role in meaning of data due to It describes actual data. Thus a good metadata management is crucial for data quality. (Gupta 2021)

2.3 DATA QUALITY MANAGEMENT

According to ISO-8000-2 and ISO 19115: 2014 concept of data quality implies the integrated activities that directs and controls the organizations.

Data quality management is not about managing quality of data. This activity starts with estimating why data is erroneous. In other words, if a data cleansing operation is carried out without figuring out root causes. It leads that the operation is repeated in given intervals.

Data quality management is also not about trying to achieve perfect data sets. As stated before, it is not possible to achieve this in terms of cost, time and effort. Therefore, it is necessary to build data quality management on balancing the existing data quality with the required data quality and benefiting from this improvement.(Jugulum 2014:4)

2.3.1. Adversities in Data Harnessing and Management

Managing data quality is not a simple task As In organizational context, there exists many factors that determine data quality. One of the adversities in data quality management is that quality expectations are not always previously well determined. Because sometimes stakeholders may not put into words their requests and even management staff may skip to ask. However, if data is supposed to be reliable and secure, data management staff needs to determine data and business requirements as well as how to measure them. (Henderson and Earley 2017:475)

2.3.1.1. Complex Geospatial Data Ecosystem

Data ecosystem can be described as an organization's all data asset. (Chisholm, 2011) If this concept is transcribed into geospatial data, It can be said that geospatial data ecosystem states all geospatial data governed by an institution.

Except smallest ones, all organizations or institutions maintain numerous Geogprahical Information Systems, data analytics tools or spatial databases. In addition to this, if we consider manually generated maps including vector data, We can say that geospatial data ecosystem of organizations are so complex and along with developing technology they will become more complex. . (King and Schwarzenbach 2020:15)

2.3.1.2. Complex Geospatial Data Ecosystem

Data ownership can be defined as determining how and by whom data can be accessed and the distribution policy of the data. In addition, these people are also responsible for ensuring that the stored data is up-to-date, accurate and usable.

Although the common belief is that these people are the people or organizations that produce the data, the issue of data ownership has different dimensions.

Organizations often worry about assigning ownership of data sets. Because these data can be changed or new ones can be created as a result of business rules. For example, personnel data, which can be described as master data, may be updated as a result of transactions carried out in the organization.

Assigning data ownership and describing explicit descriptions for data can be preferred to keep under control data quality. However, legislating these definitions allows compliance with them to be evaluated objectively

On the other hand, It would be beneficial to state that managing data storage is not data ownership. Data owners the ones who decides what data is suitable for business requirements.(Guess 2013; Kerry and Morris 2019)

2.3.1.3. Data Quality and Data Redundancy

All well managed IT systems have an acceptable redundancy policy. In case a big adversity that leads severe hardware or software problems, organizations are expected to restore the most recent well-being state prior to error occurred.

Redundancy types can be classified according to the time that it is taken. Such as hourly, daily, weekly or monthly.

In addition to redundancy systems, high availability systems such as failover mechanisms or disaster recovery centers can used in disaster recover scenarios without losing data.

Data or system redundancies' importance in data quality management are close to zero Because, If for a long time low quality data has been produced without noticed. It is nearly impossible to determine exact time of low quality data entry and to restore prior to that process. On the other hand, it causes a vast amount of data to be removed. (King and Schwarzenbach 2020:17)

2.3.2. Human Effects on Data Quality

Employees in an organization have common behavioral patterns that they exhibit against to new works, rules or standards. As a result of this phenomenon, these behavioral patterns may affect data quality of an organization in case pattern practitioner take role in a team that generates or manipulate data as well. Thus It won't be wrong to claim that main reason leads low quality of data is the human being. To

exempfly some behavioral patterns, It can be said that the reason leads to low quality data is erroneous processes carried out by staff which is authorized to enter or edit data manually or issues in impelementation of system configurations' change in an organization. Although this process can be regarded at first as a software or hardware problem. Yet, it shouldn't be disregarded that there exists human factor behind the curtain for determination of configuration settings.

Another one of the biggest issues in data quality management that there's not a feed back cycle between mistakes made and feeling aftermaths of the mistakes. For instance let's consider a clerk officer in an governmental organization where officer wasn't able to completely enter data in database or purposfully enter incomplete. If an administrative staff doesn't observe the record process, It can be slided over by employee and it would lead errenous data anaylsis afterwards. Thus adminisitrative staff that prepare strategic planning for policy makers suffer from this circumstance.

On the other hand, one of the most important things to keep in mind is that people's approach to data can be compared with their approach to work activities. For example, answers of some questions pertaining to daily routines of the officers can explain so much details about this phenomenon. Such as:

- Do the staff obey rules or tend to bend them ?
- Do the staff leave their desks tidy ?
- Is there any team showing low performance when they are compared with their peers in terms of security and quality ?
- Do employees have a tendency to find shortcuts to get around work rules ?
- Is there a lack of trust against to the systems or the employee ?

Answers of questions like these are gaining importance. The factors above may indicate the existence of individuals or teams whose behavior and approach may cause data quality problems.

For example, while an officer who is conscious and has a good command of details can help create good quality data, officers, who are creative enough to create execuses to leave their shift without doing work, may not display sufficient effort in contrubiting to boost data quality or enter qualified data. Additionally, a messy workspace may be an indication that the data that will emerge in that workspace will not be more organized than the workspace itself.

As a result, instead of trying to understand different people how interact with data in different ways. We can use three dimensions to analyze human behavior on data.

Accomodation: Measurement of how does staff obey verbal or writen rules within an organization

Effect: Measurement of how much effect does an individual staff create

Strategic Awareness: Measurement of how much a staff is aware of organization's goals and his or her contribution to realize these goals. (Tim King and Julian Schwarzenbach 2020:40)

In governmental organizations of Türkiye, there exists prevalent behavioral patterns. And mostly these patterns don't vary organization to organization very much, These patterns can be counted as low level accomodated behaviours, medium level accomodated behaviours and highly motivated or highly accomodated behaviours. Through coming together, these behavioral patterns constitute for organizational culture. And culture of an organization is directly related to quality of data it generates.

The staff exhibit low level accommodative behaviours can be classified in 3 topics. These are complainant staffs, staff who has lack of confidence and careless staffs. (King and Schwarzenbach 2020:31)

Complainant staffs generally tend to complaint every advancement within organization. Since they are indolent people they are barely shown up and participate projects. When it comes to data quality work just like in other type of Works, They complains about data quality works and object it. These people are lack of foresee. Thus They cannot understand importance of quality work and play an obstructive role in organizations and reduce other's motivation.

On the other hand, there exists some positive sides of these staffs. Since this kind of employee generally consists of talkative type people. They should be listened carefully in order to spot root causes of problems and weak point.

People who lack confidence tend to constantly feel insecure in their roles in the organization and tend to hide areas of responsibility in which they are consistently unsure or unable to use technology effectively. Additionally, these individuals tend to be hardworking employees. Their business input is fine as long as they are confident in what they are doing.

Careless staffs are generally hard working types of employees and they are generally benign people. Although they intended to benefit organization. They may

harm organizational goals since they often forget somethings to take into account. (King and Schwarzenbach 2020:32; Luan 2022)

In medium accommodative behavior pattern, While the staff are doing what they are expected to do, they exhibit low level motivation. This kind of employees require too much endeavour to comply standarts so they are tend to disobey standarts and If they find a chance, they generally conceal problems..

Some of the employees belonging to this group are the types who put forth a reasonable amount of effort but need to be convinced before a new approach or standard is adopted.

Another type of employees belonging to this group object everything that they don't make sense. This type of people are very resistant to change and tends to maintain their work practices. (Luan 2022)

Highly adaptive type staff is known with their sense of dicipline and if they encounter any kind of obscurity in their Works, they hesitate to go on process. Since they deeply obey rules.

This kind of people are abudantly shown up in teams or organizations that have heavy workloads. On the other hand this kind of employees can be called as people of advice. They are generally innovative people and responds requirements in an unorthodox way and fastly. (Güneş 2022:90)

2.3.2.1. How Do Divergent Behavioral Patterns Interact on Data Quality Issues ?

In most of organizations there exists numerous behavioral patterns being exhibited against data quality works. Some of these attitudes are positive some are not. There exists various theories on how divergent behavioral patterns affects each other. Yet, most notable one amongst them is Belbin's Team Inventory.

In the 1970s, Dr.Meredith BELBIN and her team at Henley Management College began researching team behavior, particularly what makes a team successful. Belbin and her team defined the role of a team as "acting together with others in a certain way, contributing" (Belbin, 1991) and stated that people with different behavioral patterns (positive and negative) should be in the same team. He argued that in this way, these people can be used depending on the purpose and a balance of roles will emerge (Monsalves 2023; King and Schwarzenbach 2020:41)

2.3.2.2. Individuals As A Part of A Team

In work life of a governmental institute, most of individuals work as a part of team or teams. Teams can be created based on various philosophies such as ones comprised of employees having similar talents and where they take on similar tasks. Or ones whose each of employe has unique talent and each employee takes on different tasks. Data is an indispensable factor for work of these teams.

On the other hand, according to social identity theory, human beings are innately inclined to belong a group and they are tend to obey group norms such as behavioral patterns or social pratics of associated group. In same way, It can be said being a member of successful team makes individual's performance boost and acquire talents. Whereas being member of a team of which members have negative attitude and inadequate manners gets individual to adopt inadequate practices. (King and Schwarzenbach 2020:41; Syah 2024:152)

2.3.2.3. Teams in An Organization

Just as individuals effect behaviour patterns of teams that they belong, teams in an organization effects organization's policies and actions. In data quality for organizational context, interdependence between teams is one of the a key factors. More clearly, There exists various tasks classes that effect quality of data in proportion to its size or its level of critically. For instance:

Tasks carried out positions such as ordinary clerk officer in a provincial branch of a ministry requires low level of coordination amongst employees. Since this kind of employees are generally responsible to make phone calls or answer incoming calls. Thus their effect on low level of data quality is quite limited.

Tasks carried out by IT staffs in a ministry delivering a medium size or a large size project require medium level of coordination. As there exists work delivery amongst team. Thus in case any negligence driving low level data quality doing by delivering team, low level of data quality can be spotted by taking over team.

In tasks require high level coordination, team handing over output can be taking over team in another process. Thus there exists interdependency in this kind of situations and spotting factors leading low level data quality is difficult and low data quality effects entire organization.

In addition to this, administrative staffs' stance in a governmental institution does matter. Regardless of an organization's written rules, its essence and culture

(general dominant behavior patterns and way of working) are affected by the actions of its leaders. Thus quality of data in a governmental institute is directly related with decisions of administrative staffs' decisions. For instance if in a govermental institution administrative is act behaviours as follows:

- If decisions are made with meager analysis or none
- If data which is source of decision is not quesitoned
- If it is not objected when data quality issues are arose
- If performance is evaluated without suitable quantative methods
- If local data stores are tolerated

It can be indicative of that the staff having a good approach against data are not valued in the organization, or it can be claimed that completing work is accepted without considering good or bad practices are used while doing it. (Torsten et al. 2017)

2.3.3. Data Quality Management Framework

Data can be regarded as blood that circulates within veins of organizations. Thus managing data quality requires a corporate approach. Data quality implies a series of actions ultimately carried out and necessary approaches that ensures data is fit for organization's purpose. ISO 8000-61 handles data quality issue within 20 sub-topic including setting policy and accountability issues independent from technology as follows:

- Requirement Management
- Strategy Management Data Quality
- Data Quality Policy, Standarts and Procedures Management
- Plan Development for Data Quality Implementation
- Acquiring Data Descriptions and Business Requirements
- Data Processing
- Monitoring and Controlling Data Quality
- Reviewing Data Quality Issues
- Provision of Measurement Criteria
- Measurement of Data Quality and Work Performance
- Evaluation of Measurements
- Root Cause Analysis and Resolution Development
- Data Cleansing

- Developing Process to Avoid Data Nonconfirmities
- Data Architecture Management
- Data Transfer Management
- Data Operations Management
- Data Security
- Data Quality Organization Management
- Human Resources Management(King and Schwarzenbach 2020:57)

Tim King and Julian Schwarzenbach mention about five key concepts for data quality management in their book "Managing Data Quality - A Practical Guide".

1- Carrying Out an Implementation Centeric Approach: In order to carry out a systematic data quality implementation, monitoring, controlling, assurance and data development phases must be carefully planned.

2- A Progressive Maturity Level needs to be aimed: The best solution is the 'plan, do, check, act' cycle. However, it is not possible to find the best solution immediately. To achieve this, it is necessary to take many gradual steps. Each step requires a plan to understand the current maturity level, set a target level to achieve, and implement a series of practical changes to achieve that target.

3- Creating Explicit Data Definitions: This is the step that will ensure that all data providers understand what type of data they will provide and use, and what type of data they can interpret correctly.

4- Determining Roles and Responsibilities: Because data is so prevalent in modern organizations and every employee has a potential impact on data quality. Institutions should plan and take a comprehensive approach to defining data-related roles and responsibilities. This approach also includes planning and executing appropriate training activities to ensure that each individual understands how to perform assigned tasks and their impact on the desired level of data quality.

5- Effective DQ Program is not only IT Staffs' Business: There should be chief data officers or equivalent personnel who understand the business value of data in the organization, know how the data will be used, and investigate how data can be better shared and used in the organization and encourage implementation, but do not develop actions directly for data processing. However, most organizations still do not have a chief data officer,

and someone needs to handle data management in the short term. And while doing this, we should stay away from focusing on the role of technologies in work.(King and Schwarzenbach 2020:112)

2.4 ISO 8000-61 DQ APPROACH

ISO 8000, the international standard for data quality, was created to ensure that the quality of complex data is improved in a technology-independent manner. “In the introduction to the standard, ISO asserts: “The ability to create, collect, store, maintain, transfer, process and present data to support business processes in a timely and cost-effective manner requires both an understanding of the characteristics of data and the ability to measure, manage and determines reporting ability.” ISO defines quality data as "portable data that meets specified requirements." ” The data quality standard is a product of ISO's general concern for data portability and protection. Data is considered 'portable' if it can be separated from a software application.

“ISO 8000 - Part 61 Information and data quality management process reference model is under development and this standard defines the structure and organization of data quality management, including the following.”(Henderson and Earley 2017:446)

2.4.1. Data Quality Planning

In an governmental organization in order to reach desired maturity level in data quality it implies that taking of all requirements, developing all necessary plans and setting ultimate goals. This pratic includes works as follows:

Requirement Management: It states that data related requirements are determined, described and prioritized amongst themselves in an organization.

Strategy Development for Data Quality: It implies that setting, evaluating and development of a data quality strategy for organization. Purpose of creating a data quality strategy is to provide support of administrative staffs for data quality management work and to set a road map for describing how to improve data quality level gradually. The management of the data quality strategy aims to implement the requirements holistically by collaborating and evaluating them and then transforming them into a strategy. This strategy is also printed as a small document, accessible to people in important positions in the organization.

Data Quality Policy Standarts and Procedures Management: The purpose of data quality implementation is to determine how and by whom policies, standards and procedures will be implemented. The outputs of this process are; These are plans prepared for the implementation of data quality management, including the implementation schedule, determination of roles and responsibilities, and assignment of authorities. In addition, establishing structures that provide supervision and management of relevant process implementation activities is also within the scope of this title

Plan Development for Data Quality Implementation: The role includes responsibility, funding and planning to ensure the establishment of relevant technologies and implement activities related to data quality management. (King and Schwarzenbach 2020:113)

2.4.2. Data Quality Planning

It implies a routine description and control process carried out to ensure that generated or updated data meets requirements.

Acquiring Data Descriptions and Business Requirements: The purpose of this process is to create the basis on which the organization's primary data processing outputs and approaches can be compared with the desired outputs and approaches.

The outputs of this process are the set of rules that provide a clear expression of the requirements to implement the data processing process and the data that results from this process. Business requirements include as follows:

- Purpose for data processing
- Planning for data processing
- Roles and responsibilities for data processing
- Assign staff to the related roles
- Resources and information required for data processing
- Creating necessary documentation
- As defined in ISO 8000-8, effective data identification covers the following topics:
- Syntax of data
- Meaning or context of data

Pragmatic estimations for data

Pragmatic predictions arise from decisions made by the end user based on data. One of the most important of these estimates is the scope of the data set. All data sets are incomplete. That is, no data set can contain all possible information about the subject. Whereas some data can serve organization's strategic decision making process, some data can be classified redundant. Fort his reason, when data marts are created, redundant data should be ignored otherwise resource waste is occured. For instance to determine buildings possess collapse risk because of ground liquefaction in an earthquake, There's no reason to maintain and regard which subcontractor provided construction equipment. Because this data is specific to contractor and can be evaluated as trade secret.

Another pragmatic prediction is timeliness. When timeliness is regarded context of data is not taken into consideration as an evaluation criterion rather bandwidth of line and speed are regarded as criteria. For instance, displaying earthquakes' geographical locations momentarly on a web site without any delay.

In order to business requirements and data definitions are effective. Field experts need to participate DQ work within the scopes as follows:

- Business processes are carried out in a sequential manner
- Interval of probable scenarios and cases used by pre-determined processes need to be determined.
- All relevant data attributes must be defined and the correct units of measurement, precision and allowed value ranges must be defined for these attributes.

In order to create a data definition, We first handle decisions made by end users and for each decision we examine data that can be used to support that decision. After then, We need to determine the characteristics that associated data is expected to have. Scope and characteristics of data determines basis of data definitions.

Data Processing: The purpose of data processing is to meet the information needs of defined transactions carried out by a defined set of users.

Output of this process is data context in suitable format. Hence, data can be used to support decision making process carried out by end users.

Figure 4: Data Processing (tutorialspoint.com 2024)

Data processing includes various types of sub-processes such as data generation, data transform, elimination and dissemination. This processes can be carried out either manually or via an automatic tool. Prior and after to this process definitely data quality assessment needs to be carried out.

Monitoring and Controlling Data Quality: Data Quality controlling and monitoring process is used to determine points where data processing fails. Aim of this process is to reduce failure probability of overall data quality

This process mandates to create tests to validate data definitions and spot data quality issues that could't be fixed through data processing. Key factor that needs to be kept in mind about this process is that during creation phase of processes and test tools to stick internationally accepted standards.

On the other hand, data definitions should be described computer-processable. In this way, it will be possible to evaluate these specifications via a pass/fail test that will be automatically generated by the computer. Thus, large amounts of data are processed and subject to rules without human intervention. This also provides confidence in the rules. (King and Schwarzenbach 2020:73)

2.4.3. Data Quality Assurance

It is the process of understanding the current data quality before evaluating what actions to take. Data quality assurance involves a set of processes that support data quality management.

Reviewing Data Quality Issues: In order to carry out root cause analysis and fix them, data quality problems exist in data are tried to be found out.

Provision of Measurement Criteria: In order to understand effect of quality issues exist in data for organizational goals, we determine a set of criteria to evaluate quality issues objectively.

Measurement of Data Quality and Work Performance: It is to measure the data quality level and the performance of data quality management processes.
Evaluation of Measurements: It is analysis of results of measurement criteria previously determined.

2.4.4. Data Quality Progress

In order to fix data quality problems permanently, the methods that authorized staffs need to follow.

Root Cause Analysis and Resolution Development: To diagnose the root cause of data quality problems and offer solutions to prevent the recurrence of these problems
Developing Process to Avoid Data Nonconfirmities: Implementing solution suggestions that will reduce or eliminate the root causes is to prevent the recurrence of data and process nonconformities.

2.4.5. Data Quality Progress

Some activities support data quality management are specific to technology or management style. These are as follows;

Data Architecture Management: It states that understanding data architecture of an organization along with data management practices. As a result of an effective architectural governance data is stored consistently and far away from quality problems across organization's data eco system. (King and Schwarzenbach 2020:77)
Data Transfer Management: Data transfer management can be stated as development of associated tools or scripts to transfer data from source system to destination system. During this process a series of processes are caried out such as type or format transformations.

Aim of data transfer management is to transfer data without leaving any obscurity may lead to data quality problems. Thus output of this process is directly related with understanding context of data and mechanism that transforms and transfers data.

As a result of this process;

- Data is transformed into required format
- Data is transferred into reciever system.
- Reciever system make data available for authorized staffs.

Key point needs to be kept in mind in this process is that the process needs to be implemented via automatic tools rather than carrying out manually. And at the end of process, validation for transferred data must be done.

Data Operations Management: It implies effective governance of data ecosystem of an organization through suitable technologies and methods.

Data Security: Determining data security policies and implementing them across data ecosystem of an organization.

2.4.6. Resource Allocation for Data Quality Operations

At the organizational level, it is how resources are allocated to data management and data-related skill development. It includes the following topics:

Data Quality Organization Management: It can be described as preparing an organization to carry out data quality work through setting rules, motiving the staff and informing stakeholders.

Human Resources Management: It is the development of data quality experience and skills within the workforce. This process also includes teaching best practices and skills on an institutional basis.(Henderson and Earley 2017:520; King and Schwarzenbach 2020:57)

2.5 ISO 19157 DQ APPROACH

ISO 1957 is developed to measure and evaluate geographical data quality in order to ease determination process for data sets to whether they satisfy organizational requirements. Data quality reports are essential for geographical data to be shared or interchanged as well as aid end users to assess performance of GIS application or analysis tool they use. ISO 19157 standard is prepared by ISO's ISO/TC 211 Technical Comitee through collaborating with European Committee for Standardization's (CEN) technical committee under Vienna Agreement. After that first edition of standard (ISO-19157-1) is revised to ISO 19157-3. ISO 19157 can considered as integrated version of ISO 19013, 19109 and ISO 19115 standarts. This framework provide a series of instruction to aid data users to evaluate quality, to prepare a conceptual model to handle quality info and a sketch to prepare quality report. (López et al. 2020)

2.5.1. Elements of Geospatial Data Quality

Data quality elements are data quality components in essence which describes a given aspect of geospatial data quality. There exists 21 spatial data quality elements or sub-elements. (Fischer et al. 2023)

Completeness: Completeness is degree of a data set or more than one data sets representing a feature possess all mandatory values and related entity instances at the time at which the dataset or datasets has been created (ISO 8000-2; ISO/IEC 25012). In other words, term completeness is used to state whether features exists or not along with their attributes and relationships. According to the standard, two types of elements define completeness:

— **commission:** Much data than desired present in a dataset;

— **omission:** Data absent from a dataset. (ISO 19157:2023)

Logical Consistency: Logical consistency is defined degree of how well obeying to logical rules of data structure, attributes and relationships. ISO 19157 standard specifies logical consistency along with data quality elements and their associated measures:

— **conceptual consistency:** matching degree to rules for reference universe

— **domain consistency:** obeying degree of value rules for domains set in data model

— **format consistency:** degree of how suitable data stored in accordance with the physical structure of the dataset (Porfirio et al. 2020)

— **topological consistency:** degree of correctness for topological characteristics of a given dataset

Positional Accuracy: Positional accuracy is in essence measurement accuracy (ISO/IEC Guide 98-3; ISO/IEC Guide 99) thus it is said that positional accuracy states how close a measured position of feature to a position within a spatial reference system each other.

Depending on the scope and type of the reference system, the standard specifies three spatial data quality elements to describe positional accuracy:

— **absolute / external accuracy:** degree of how close given coordinate values to values regarded as correct in a standard coordinate reference system

— r**elative / internal accuracy:** *"closeness of the relative positions of features in a related dataset to their respective relative positions accepted as true in a local coordinate reference system"* (ISO 19157:2023)

— **gridded data positional accuracy:** measure of how close gridded data spatial position values to values regarded correct. Positional accuracy is estimation of uncertainty for measurement results (ISO 19116)

Temporal Quality: Temporal quality is defined as the quality of the temporal attributes and temporal relationships of features. Temporal quality can be described by data quality elements as follows along with their related measures (Şeval Alan 2022:10)

— **accuracy of a time measurement:** closeness of reported time measurements to values accepted as correct

— **temporal consistency:** correctness of chronology

— **temporal validity:** Temporal validity can be considered as timeliness, currentness or actuality (ISO/IEC 25012. For instance a date value such as '30 Feb 2024' indicates a domain inconsistency. Thus It can be said that temporal validity is different from domain consistency of temporal values. To exemplify temporal invalidity, example of a dataset that contains objects that did not exist in the timestamp associated with their record can be given.

Thematic quality: Thematic quality is correctness of quantitative and non-quantitative attributes as well as true classifications of features and their relationships. Standard examines thematic quality with 3 data quality elements:

— **classification correctness:** comparison of the classes assigned to features or their attributes to a universe of discourse

— **non-quantitative attribute correctness:** measure of whether a non-quantitative attribute is correct or incorrect;

— **quantitative attribute accuracy:** closeness of the value of a quantitative attribute to a value accepted as or known to be true. (ISO 19157:2023)

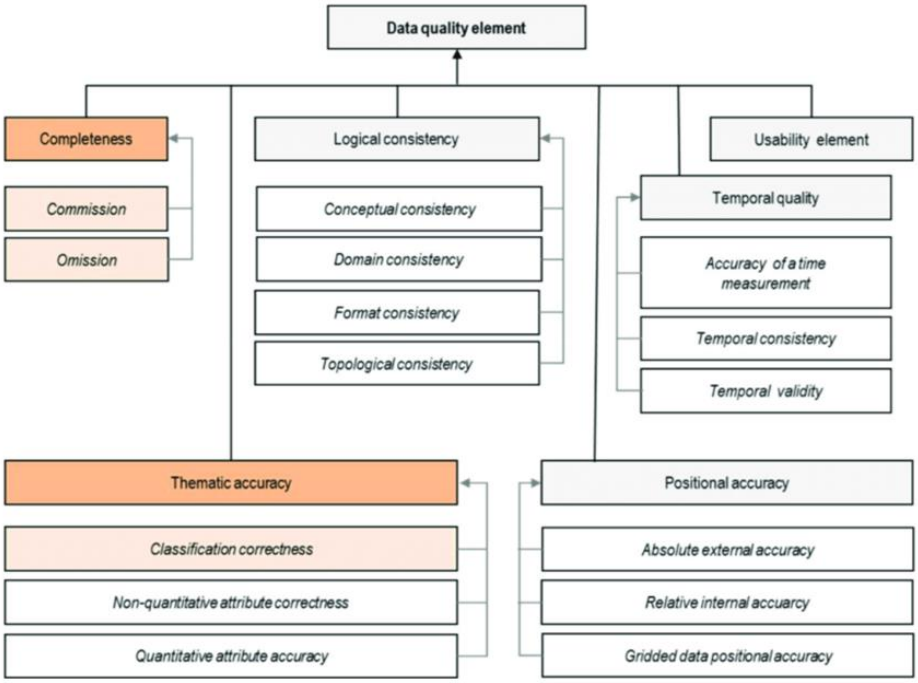


Figure 5: Overview of the ISO 19157:2013 data quality elements (Mõisja et al. 2018)

2.5.2. Metaquality Elements

Meta-quality elements are a set of quantitative and qualitative statements about quality assessment and its outcome. Information about the quality and appropriateness of the assessment method, the measure applied and the result given can be as important as the result itself.

Metaquality can be described through elements as follows:

— **Confidence:** reliability of a data quality process' result

— **Representativity:** A result that is representative of data within the sample data quality used

— **Homogeneity:** The expected or tested uniformity of results obtained for a data evaluation evaluation. (ISO 19157:2023)

2.5.3. Data Quality Evaluation Methods

A data quality evaluation procedure comprises one or more data quality evaluation methods. Data quality evaluation methods can be divided into two main classes: direct and indirect. Direct evaluation methods determine data quality through the comparison of the data with internal and/or external reference information. Indirect evaluation methods infer or estimate data quality using information on the data such as lineage

Direct Evaluation: Direct evaluation is an evaluation method based on direct scrutinizing of items within a given dataset. The direct evaluation methods can be divided into 2 parts as internal evaluation or external evaluation. Internal direct evaluation method assesses quality of data in a given data set without taking reference data into account. Whereas external direct evaluation method uses external reference data for associated data set.

Both internal and external methods uses inspection procedures as follows:

— **full inspection:** data quality evaluation assesses each item that is contained by scope

— **sample-based inspection:** to carry out data quality evaluation subset of data set is determined for given scope. If this method is used, it needs to be stated in the quality evaluation report

Indirect Evaluation: Indirect evaluation is carried out through using external knowledge about target data set thus it may not be objective As In this type of evaluation data quality is subject to be estimated. This external knowledge may include information about dataset's usage, lineage and purpose or other data quality reports on the dataset or data used to produce the dataset. For instance knowing about methods or tools used for capturing associated data set (Ureña-Cámara et al. 2018)

2.5.4. Data Quality Evaluation Methods

Geospatial data quality process starts with specification of data quality units. A data quality consists of scope and quality elements. For instance conceptual consistency, completeness (commision and omision), thematic classification etc. All data quality units are enumerated. Such as QU-1, QU-2 etc.

After qualiy units are specified, data quality measures are specified such as a given conceptual schema compliance, number of missing items, rate of missing items, excess items etc. and measures are enumerated as well.

Then, data quality evaluation procedures are specified. For instance direct external producedure etc. After this process, data quality results are elicited.

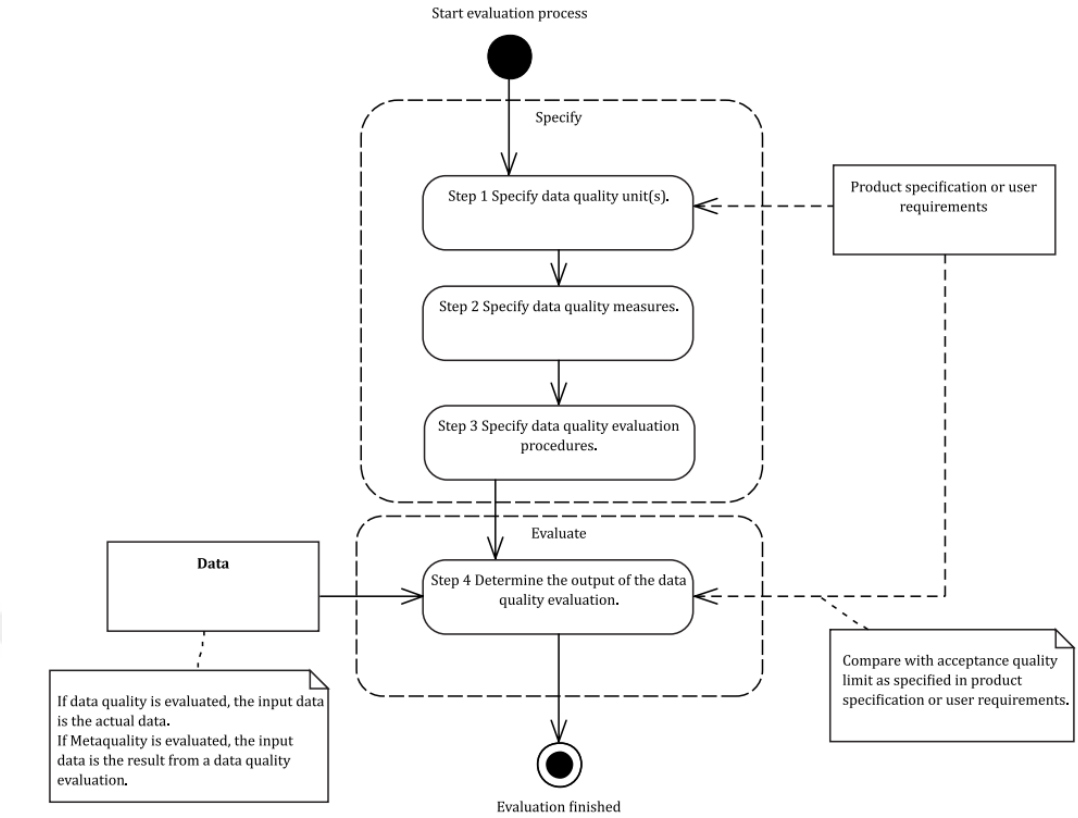


Figure 6: Geospatial Data Quality Process (ISO 19157:2023)

When geospatial data quality is evaluated. It is assessed according to order as follows:

1- Data is evaluated in terms of logical consistency
2- Data is evaluated in terms of format consistency
3- Data is evaluated in terms of completeness
4- Data is evaluated in terms of accuarcy (ISO 19157:2023)

2.5.5. Reporting Geospatial Data Quality

Data quality can be reported as metadata and should include quality evaluation report In order to provide more details. A good data quality report also should include derived results, actual results and aggregated results and needs to include details about data quality evaluation, such as evaluation steps. On the other hand, optionally a quality evaluation report may additionally be issued.

Data quality reports also need to be prepared in a hierarchically way starting from top level to bottom where atomic unit of quality work resides. (ISO 19157:2023)

2.6 ISO 19157 DQ APPROACH

ISO 8000 reaches data quality issue with a broader angle Since, ISO 19157 standard solely focuses on increasing quality of data in terms of completeness, logical consistency, temporal quality, positional accuracy and metaquality without considering assignation of some data related roles such as data ownership, data custodian etc. Also ISO 8000 put forward data quality dimensions such as availability and uniqueness which are not included by ISO 19157.

CHAPTER III
STUDY'S METHOD, MODEL, DATA AND CASE STUDY

3.1 MODEL OF STUDY

In this study, 3 types of geospatial data set (tables of a PostGIS relational gesopatial database called konumsaldb) is used. These datasets are created via .NET Core Web API which is responsible to take user inputs from a front-end World Sphere application to convert these inputs as geography data type to save in a spatial database. In the spatial data sets most of data is subject to data quality work, all data is recorded with EPSG:4326 coordinate system with geography data type.

All data in data sets are generated for test associated Web API testing purpose. And data quality issues emerged due to lack of good programming skills.

Tree data sets consisting of various data types. For instance tblpoint table stores point type locations , tblline table stores line or multilline type data and last one tblpolygon table consists of polygon type data. At first stage, all data sets are profiled and currrent data quality problems are extracted. After then associated data quality problems are fixed through regarding some business requirements.

Both prior and after to case study, data is used within analysis and analysis results are compared in terms of evaluating reliability.

3.2 DATA QUALITY WORK ON SAMPLE DATA SETS

In the geospatial database using in case study which is called "konumsaldb" consists of 3 tables and each table stores different type of vector data. One stores point type, other stores line and multiline data and last one stores polygon data.

Each data set comprises of same table structure along with a column of which data type is geography with EPSG:4326.

3.2.1. Profiling and Data Quality Improvement Study

Data profiling is a study carried out to analyze data, to find duplicates, and to reveal data frequencies and patterns. Within the scope of the thesis, profiling studies

were carried out with the Ataccama DQ Analyzer program and data quality improvement studies were carried out with the DataCleaner program.

3.2.1.1. tblpoints Table Profiling Process

tblpoint table there exists 207 rows. In this table there exists point information for various sights in city of Ankara.

In table stores point data. There exists plenty data with various quality problems. Before commencing data quality work, An analysis taking center as Kizilay Square and brings other points within 500m Radius is carried out. To do this analysis, We simply use the SQL command as follows:

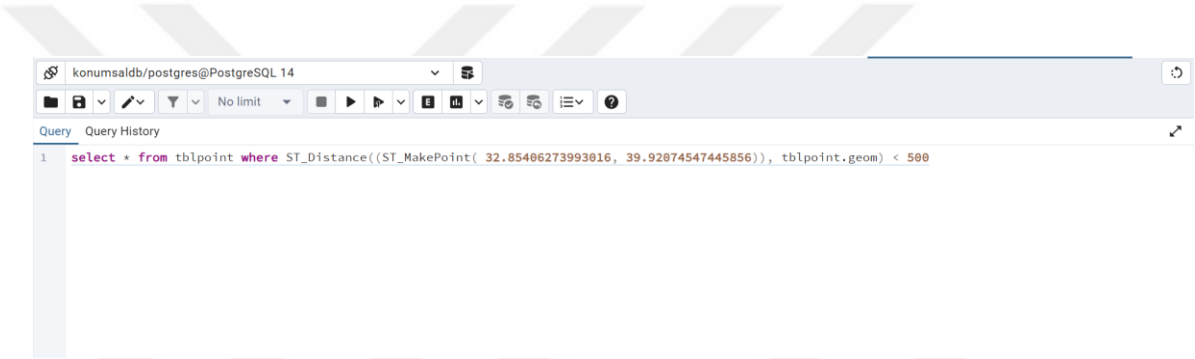
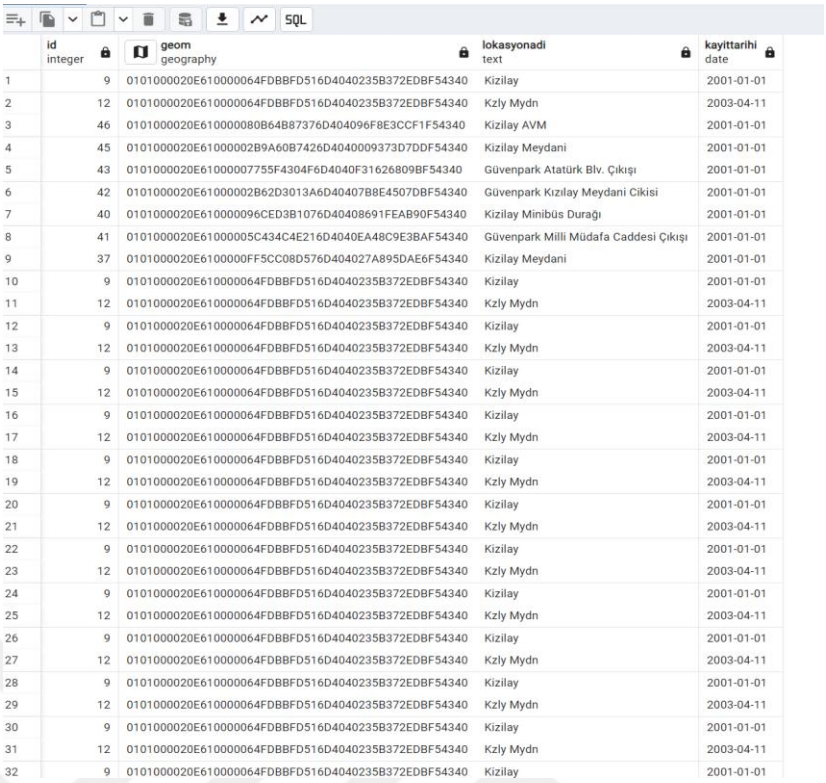


Figure 7: Query using in analysis [1]

The query brings 33 rows result as follows. But as it can be figured out in Figure 6, there exists plenty dublicate data. Dublicate data violates uniqueness data quality dimension of ISO 8000 Data Quality framework as well as in a spatial analysis it causes wrong results to return. Thus decision makers make wrong decision for strategic planning. For instance let's consider a scenario where end user search for number of squares with 5 km diameter from Anitkabir for a public transportation planning. Expected result for his kind of analysis should only be 3. And it would be Kizilay, Sihhiye and Ulus Square. But according to data on the table. Likely result would be 33 instead.

[1] In PostGIS DBMS via associated query, all points within 500m radius are called and Unique records to be retrieved is expected.



	id integer	geom geography	lokasyonadi text	kayittarihi date
1	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
2	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
3	46	0101000020E6100000080B64B87376D404096F8E3CCF1F54340	Kizilay AVM	2001-01-01
4	45	0101000020E61000002B9A60B7426D4040009373D7DDF54340	Kizilay Meydani	2001-01-01
5	43	0101000020E61000007755F4304F6D4040F31626809BF54340	Güvenpark Atatürk Blv. Çıkışı	2001-01-01
6	42	0101000020E61000002B62D3013A6D40407B8E4507DBF54340	Güvenpark Kızılay Meydani Cikisi	2001-01-01
7	40	0101000020E610000096CED3B1076D40408691FEAB90F54340	Kizilay Minibüs Durağı	2001-01-01
8	41	0101000020E61000005C434C4E216D4040EA48C9E3BAF54340	Güvenpark Milli Müdafa Caddesi Çıkışı	2001-01-01
9	37	0101000020E6100000FF5CC08D576D404027A895DAE6F54340	Kizilay Meydani	2001-01-01
10	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
11	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
12	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
13	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
14	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
15	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
16	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
17	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
18	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
19	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
20	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
21	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
22	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
23	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
24	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
25	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
26	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
27	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
28	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
29	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
30	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01
31	12	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kzly Mydn	2003-04-11
32	9	0101000020E610000064FDBBFD516D4040235B372EDBF54340	Kizilay	2001-01-01

Figure 8: Result of query using in anaylsis [2]

This circumstance may cause time and financal loss. Since wrong places may be determined as stop point. And this error may be noticed late or even may not and for this reason public resources may be wasted.

[2] As result of query, numerous dublicate point values are retrieved. This circumstance violates uniqueness data quality dimension of ISO 8000. In data analysis dublicate values may lead miscalculations and this situation also leads wrong decisions to be made.

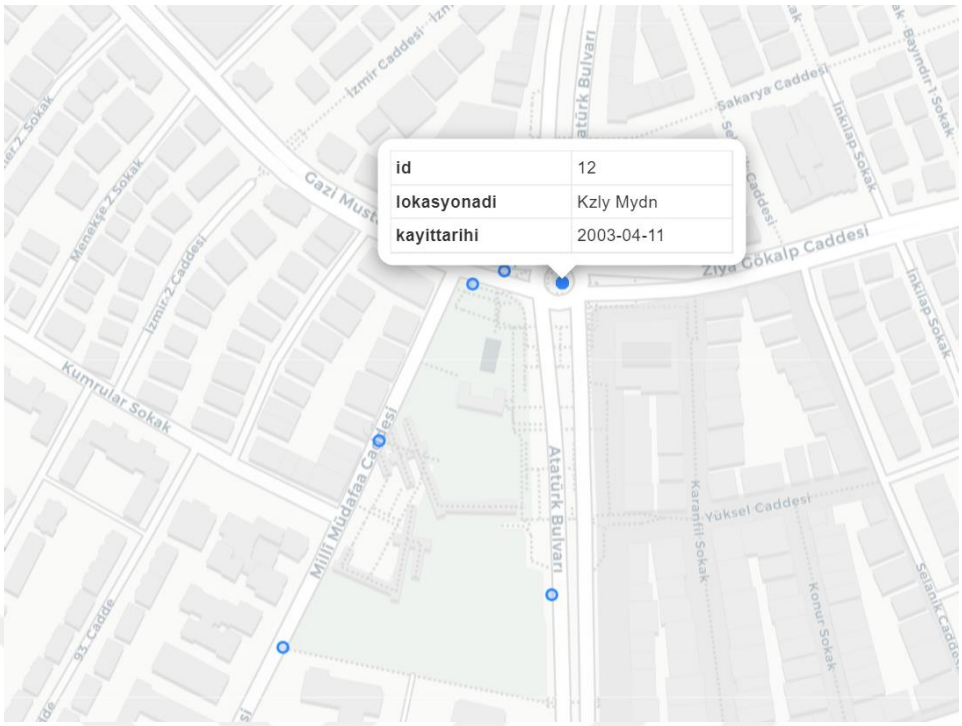


Figure 9: Analysis results on map display [3]

According to cleaning process, There exists 174 rows having same value for geom (data type:geography) column. There exists 4 rows indicating same value currently exist in table. And 29 rows with unique values.

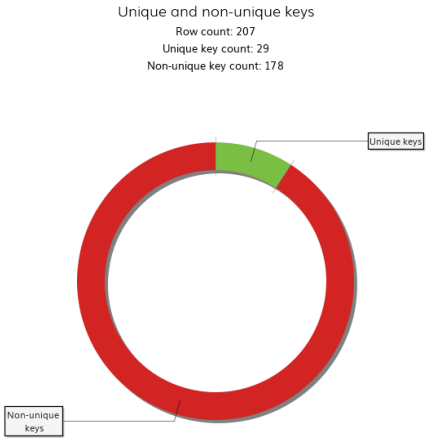


Figure 10: Profiling tblpoint table according to all columns[4]

[3] In Figure 7, dublicate point data representing actually same location is presented along with other values by using a base map from OpenStreetMap project. Biases between point locations stemmed from CRS difference.

[4] In Figure 8, tblpoint is profiled in terms of uniqueness data quality dimension through considering all columns

On the other hand, there exists 2 rows of which record date (kayittarihi) field is uncomplete. Also data possesses unique values have suitable format in terms of geography, syntaxial and date. In other words, data doesn't have validity problems.

Namely, approximately 14% of data is unique and usable for data analysis in order to obtain good results and perform better decision making activities.

Table 2: tblpoint table's data statistics [5]

Dublicate Geometry	174
Same Place with Different Geometry	4
Rows with Missing Column	2
Unique and Usable Rows	27

According to profiling process tblpoint table violates uniqueness and completeness data quality dimensions. On the other hand, according to geospatial data quality elements the profiling result is as follows

Geospatial Data Quality Score Board	
Table: tblpoint	**Geometry Type:** point
Data Quality Element	**Percentage (%)**
Completeness	99,04
Omission	85,9
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	100
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

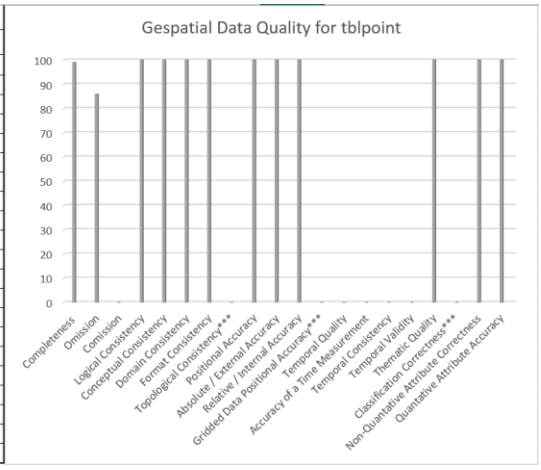


Figure 11: tblpoint table's data profile prior to data processing

Geospatial data quality is measured via full inspection method and through data cleaning process all redundant data is eliminated. After data manipulation

[5] In Table 2, tblpoint table's data quality statistics are given considering completeness data quality dimension along with uniqueness.

processes are carried out, geospatial data quality profile of new data set is as follows:

Geospatial Data Quality Score Board	
Table: tblpoint	**Geometry Type:** point
Data Quality Element	**Percentage (%)**
Completeness	100
Omission	0
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	100
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

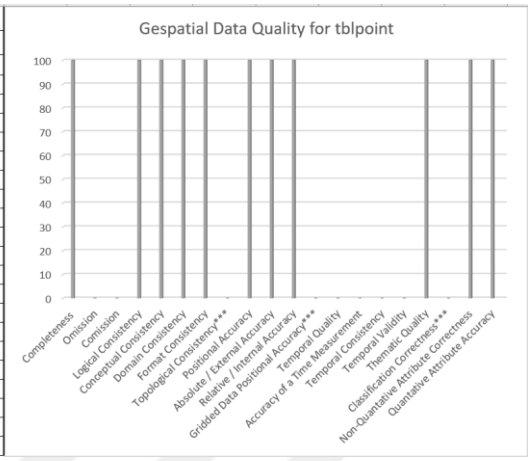


Figure 12: tblpoint table's data profile after data processing

3.2.1.2. tblline Table Profiling Process

In tblline table, there exists 7 multiline data. And none of them is dublicate nor represent same location. However in this type of data there exists another issue called biasing.

Table 3: tblline table's data statistics

Dublicate Geometry	0
Same Place with Different Geometry	0
Missing Column	0
Unique and Usable Rows	7

As seen in the image below. On of 7 Lines is missing from its actual path in some points. This can be a problematic situation in case a spatial data analysis. Since if it existed, adjacent paths would overlap or cross each other. That's why it needs to be fixed.

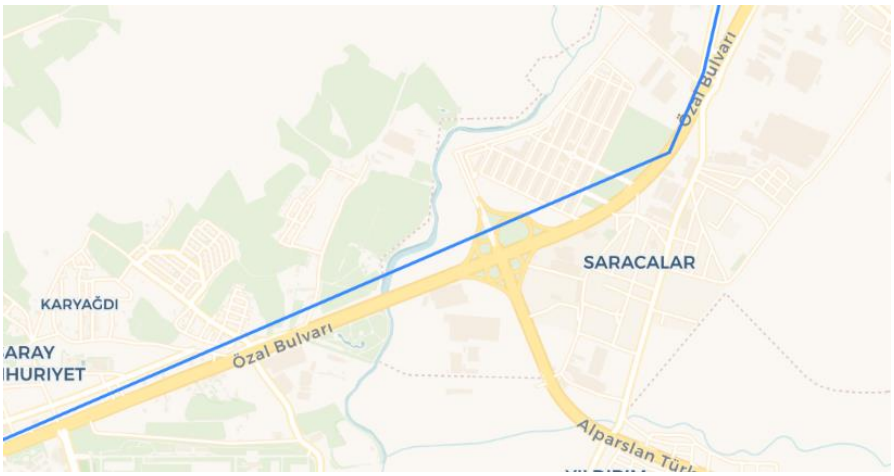


Figure 13: tblline's low quality data on map display [6]

For one record that biased from line that it is supposed to follow, the profiling score card of the associated table is as follows with 85% of Absolute / External Accuracy value Since 1 of 7 value misrepresent real world location.

Geospatial Data Quality Score Board	
Table: tblline	**Geometry Type:** line / multiline
Data Quality Element	**Percentage (%)**
Completeness	100
Omission	0
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	85
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

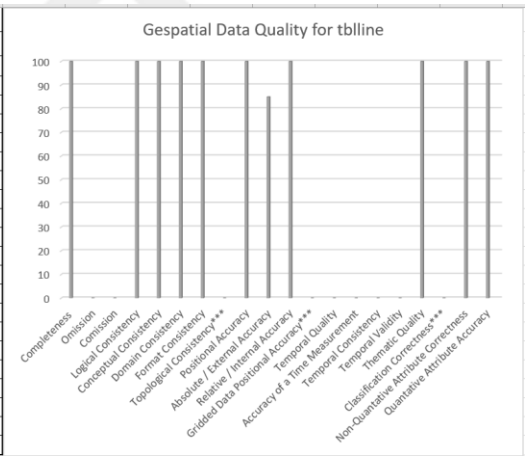


Figure 14: tblline table's data profile prior to data processing

In order to fix data, We need to remove the entry and re-insert data by adding extra vertex points to stabilize the line.

[6] In Figure 9, a multiline intended to follow Ozal Bulvari with minor bias is displayed. This data quality issue occurred because Web API Service responsibe to send geographical coordinates to DBMS sent some missing coordinates then vertex points are miscalculated. This error refers positional accuracy of ISO 19157's data quality elements. This kind of errors may lead wrong results in data analysis. For instance in a shortest route calculation problem, wrong route can be returned to end user.

To do this first Geography data type needs to be expressed as human readable text via built-in PostGIS ST_AsText command. Then after We follow the edge points and find the bias point and add extra points from this point to place line in correct position.

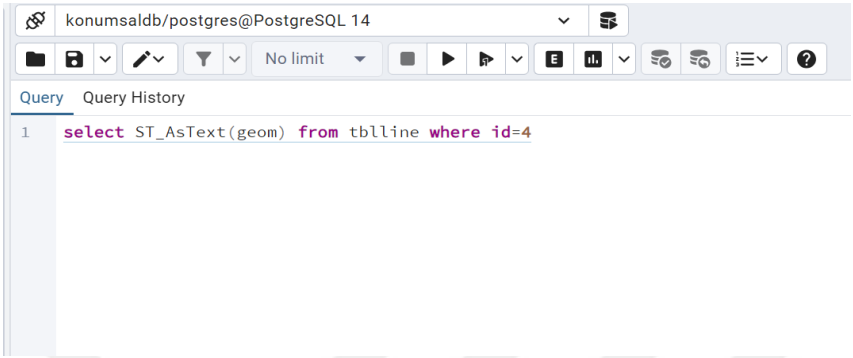


Figure 15: Usage of ST_AsText command

The ST_AsText query gives us the vertex points' coordinates so Coordinates can be fixed and along with new coordinates data can be re-generated.

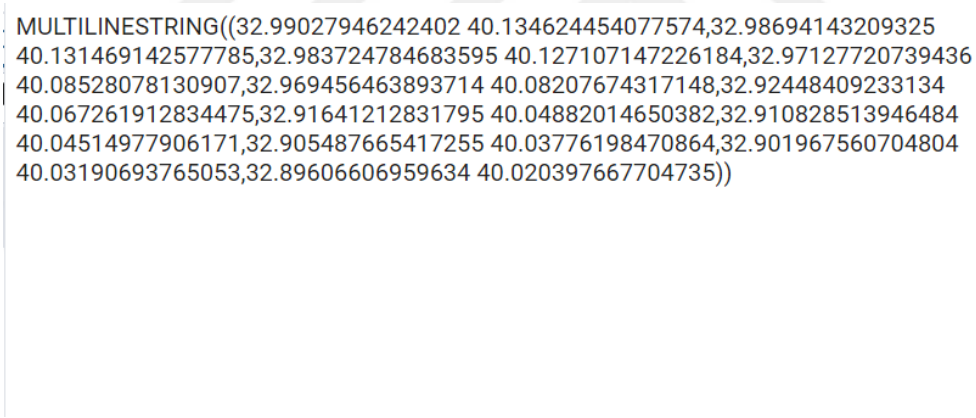


Figure 16: Result of ST_AsText command

This type of data quality issue may lead miscalculations in shortest path calculations or in spatial data analysis that aims to find out cross roads. That's why they need to be found out and fixed.

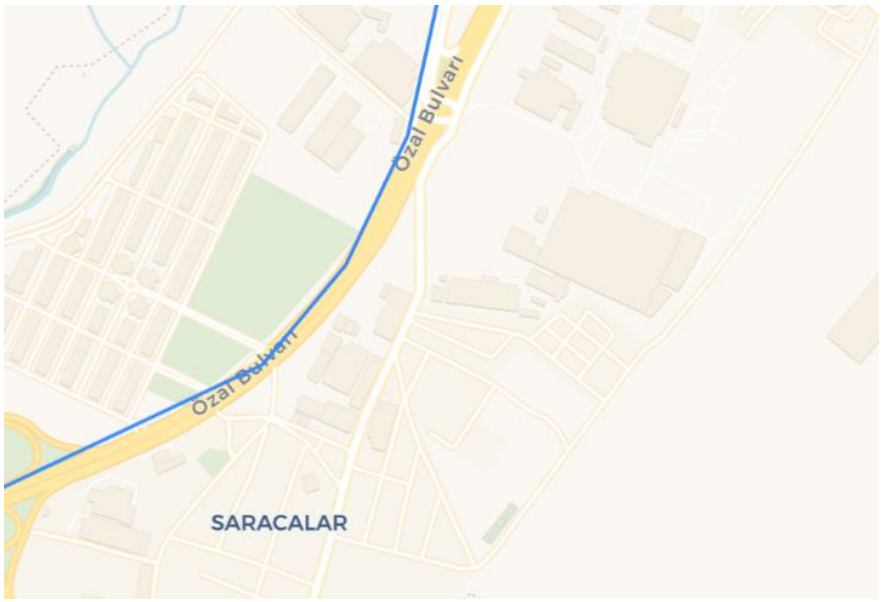


Figure 17: tblline's corrected data on map display

Data of tblline table violates absolute/external accuracy. After fixing process analysis making through data of tblline gives 100% success for absolute / external accuracy.

Geospatial Data Quality Score Board	
Table: tblline	**Geometry Type:** line / multiline
Data Quality Element	**Percentage (%)**
Completeness	100
Omission	0
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	100
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

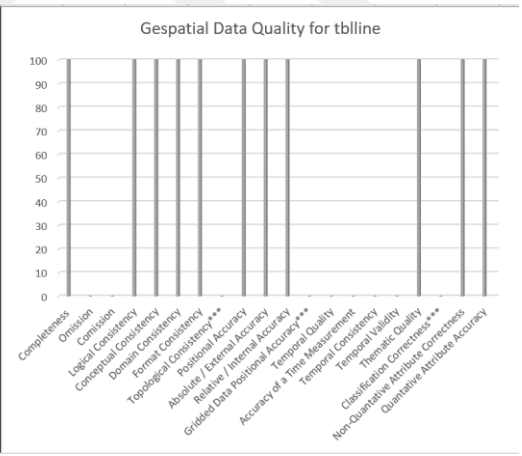


Figure 18: tblline table's data profile after data processing

3.2.1.2. tblpolygon Table Profiling Process

In polygon type vector data which is included by tblpolygon same rules with tblline are valid. There exists 10 records within the table and it includes some data quality issues.

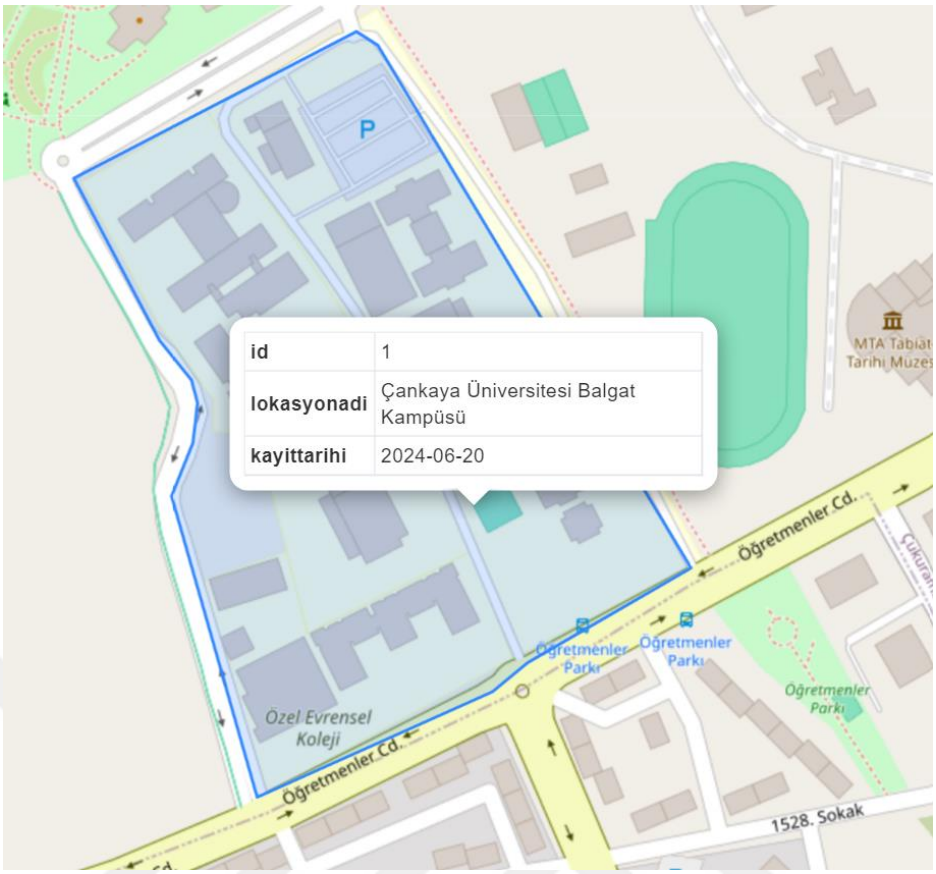


Figure 19: Sample polygon data from tblpolygon[7]

It is important to describe adjacent polygons. Since in case they crosses, anaylsis can be made may varies. Also it is important that adjacent polygons shouldn't overlap.

If there exists any kind of overlaping or interwined polygons, They may be found via SQL statements as follow.

```
SELECT a.lokasyonadi AS ad1, b.lokasyonadi AS ad2
FROM tblpolygon a, tblpolygon b
WHERE a.id < b.id AND ST_Intersects(a.geom, b.geom);
```

[7] Figure 13, a polygon encompasses Cankaya University's Balgat Campus is displayed. It is a perfect polygon and does not include any quality issue. Overflows to road stems from CRS since base map used belongs OpenStreetMap which uses EPSG:3857

As it could be figured as follows, In tblpolygon table, there exists 2 rows that interwined each other.

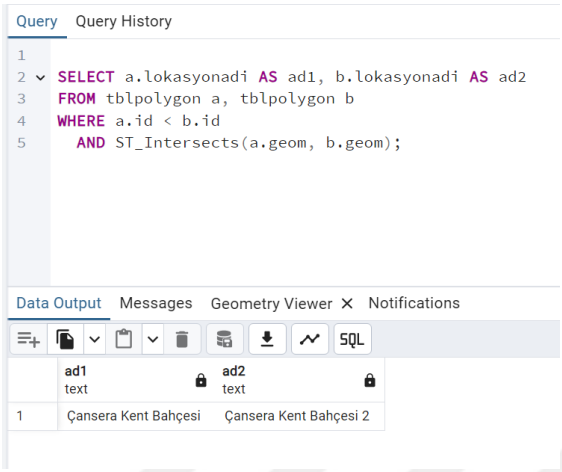


Figure 20: Result of SQL query designed to find overlapping polygons[8]

Geospatial Data Quality Score Board	
Table: tblpolygon	**Geometry Type:** polygon
Data Quality Element	**Percentage (%)**
Completeness	100
Omission	10
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	100
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

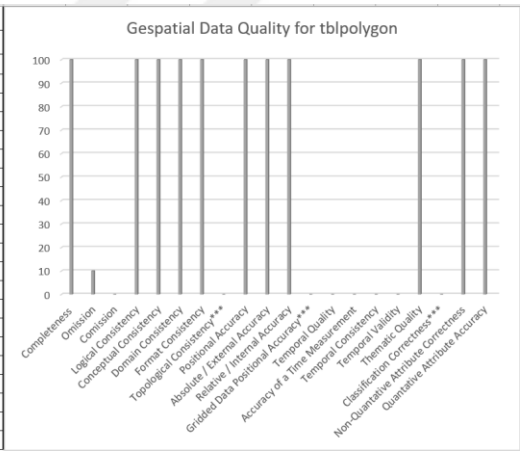


Figure 21: tblpolygon table's data profile prior to data processing

According to profiling results, there exists omission problem in which one row is recorded as dublicate. Since there exists 10 records totally, 10% of all records include omission.

As resolution, data cleaning process can be carried out and as result one of dublicate values can be removed. In addition to this polygons in table scrutinized in

[8] In Figure 14 via SQL query on it, Overlapping polygons within a table are intended to find out. The Query can find whether they are dublicate values or not. And query returns that there exists 2 dublicate values which are also overlapping as result.

terms of temporal, logical or positional accuracy and there couldn't be found any data quality but this.

Geospatial Data Quality Score Board	
Table: tblpolygon	**Geometry Type:** polygon
Data Quality Element	**Percentage (%)**
Completeness	100
Omission	0
Comission	0
Logical Consistency	100
Conceptual Consistency	100
Domain Consistency	100
Format Consistency	100
Topological Consistency***	N/A
Positional Accuracy	100
Absolute / External Accuracy	100
Relative / Internal Accuracy	100
Gridded Data Positional Accuracy***	N/A
Temporal Quality	N/A
Accuracy of a Time Measurement	N/A
Temporal Consistency	N/A
Temporal Validity	N/A
Thematic Quality	100
Classification Correctness***	N/A
Non-Quantative Attribute Correctness	100
Quantative Attribute Accuracy	100

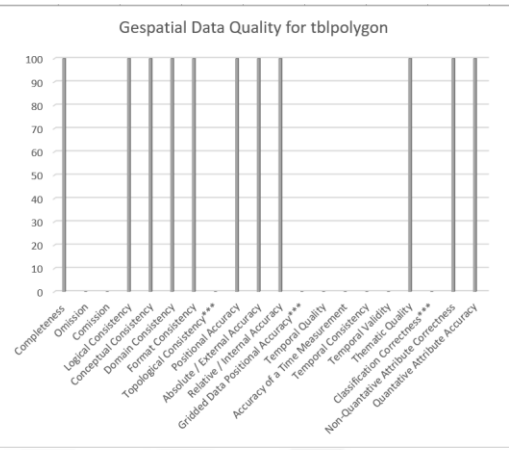


Figure 22: tblpolygon table's data profile after data processing

CHAPTER IV
RESULTS

Unfortunately, there is no single data quality approach that can suit all governmental institutions' needs in data quality improvement efforts. Although data quality activities are implemented on analytical data, this is an issue that not only data analytics or business intelligence teams but also institutions should think about as a whole.

Although public institutions serving in our country today incur serious costs in terms of data storage, their rate of utilizing data is significantly low. Although artificial intelligence and big data technologies have frequently occupied the agenda of the IT world lately, the number of public institutions using these technologies is very low in our country for this reason, These technologies need to be adopted by public institutions haven't yet met with them as quick as possible.

The use of these technologies and data analytics in the context of geographical data is a foreign phenomenon, with the exception of a few public institutions. However, geographic data analysis is a phenomenon that can be used in a wide range of public areas, from public security to urban planning, from urban planning to education planning, and should be at the backbone of decision-making processes.So populating harness of spatial data and GIS softwares in public institutions that haven't met with these technologies is a must.

Public institutions those harnessing spatial data and using GIS softwares need to consider their data's quality in other words They should purchase ISO standards and apply the practices of them then They need to call for Institute of Turkish Standards for audit in order to obtain ISO 19157 certification.

It is important to apply practices of ISO 8000 along with ISO 19157 since, ISO 8000 mandates determination of data related roles such as data ownership, chief data officer, data custodian etc. Establishment of data team consisting of employees having these roles may lighten burden of administrative staff

It is highly advised that data quality management process should be carried out via an automated tool rather than doing manually. So public institutions intended to carry out data quality work are expected to search for data quality tools.

On the other hand conducting data quality studies in the public sector requires more than applying these standards. In this context, the psychological side of the study, as much as its technical aspect must be understood well. More clearly, it is important for the staff in the managerial role in the institutions to start working by knowing the attitudes of the employee towards working, ethical characteristics of the staff. In other words, since data quality management is a phenomenon that can change the business culture in the institution, it is important for the institutional management to be able to successfully carry out change management and be prepared for it. In this context, situations such as reorganization of the institution, changing the technologies used or introducing new technologies should be planned and implemented.

From this perspective , it is important to create a coding culture in the institutions for developer teams, to prevent code complexity and to create a feedback mechanism by controlling this.

On the other hand, it is also important for software developers to develop and implement unit tests and function tests of the relevant software on test servers after writing their codes. Although the development of these tests may seem to increase development costs at first glance, it actually provides the opportunity to prevent many problems that may occur in the future. At this point, if necessary, separate dedicated teams should be formed for test development.

On the other hand, data quality studies are not short-term and one time works, but are long-term and require a comprehensive action plan and good coordination between technical units and business units.

In addition to the software development process, standardization of data coming from different sources to data storage systems is also an important issue and is one of the most important factors affecting geographical data quality. Since lineage is a spatial data quality element and has a direct effect on data quality. In this context, data origin verification and standardization checks should be carried out.

Another factor that causes poor quality in the data is the excess of human intervention. In this context, it is also important to apply DataOPS practices, which is the trend of the 2020s in data governance, to the institution, to establish pipelines by

contacting the relevant institutions for the data to be transferred, and to create data quality commitments and have them signed by the relevant managers.

REFERENCES

AÇİLER Sezer (2020), *Veri Nedir Veri Tabanı Nedir ?*, https://www.Iienstitu.com/Blog/Veri-Nedir , DoA. 21.06.2024.

ALAN Şeval (2022), *Coğrafi Verilerin Veri Kalitesi Parametreleri Açısından İrdelenmesi* (Master's Thesis), Karadeniz Technical University Graduate School of Science Department of Geomatics Engineering, Trabzon.

AMAZON AWS Tk (2024), *Veri Modelleme Nedir?*, https://aws.amazon.com/Tr/What- Is/Data-Modeling/, DOA. 21.01.2024.

ARAZ Aytaç (2013), *Metaveri Standartlarının İrdelenmesi ve Harita Genel Komutanlığı için Bir Metaveri Profili Oluşturulması* (Master's Thesis), Aksaray University Graduate School of Science, Aksaray.

BUZELLI Brian (2023), *Data Quality Engineering in Financial Services*, 1'st Press, O'Relly Sebastopol CA, USA.

CAMERON F. Kerry and MORRIS B. John (2019), *Why data ownership is the wrong approach to protecting privacy?*, https://www.brookings.edu/blog/techtank/2019/06/26/why-data-ownership-is-the-wrong-approach-to-protecting-privacy/, DoA. 15.03.2024.

CHASINOV Nick (2018), "What Is The General Data Protection Regulation And Should You Care?",*Forbes Magazine*, https://www.forbes.com/sites/theyec/2018/08/10/what-is-the-general-data-protection-regulation-and-should-you-care/ , DoA. 21.06.2024.

CHISLOM Malcolm (2011), *The Data Land Scape*, https://www.dataversity.net/the-Data-Landscape/#, DoA. 21.06.2024.

CICHY Corinna and RASS Stephen (2023), "An overview of data quality frameworks", *IEEE Access,* Vol. 7, pp. 24634–24648.

COETZEE Serena, IVANOVA Ivana, MITASOVA Helena and BROVELLI A Maria (2020), "Open geospatial software and data: A review of the current state and a perspective into the future", *ISPRS International Journal of Geo-Information*, Vol. 9, pp. 90.

DATAWORKS (2024), *Stages in the Data Management*, https://www.dataworks.ie/5-stages-in-the-data-management-lifecycle-process/, DoA. 21.02.2024.

DEVILLERS Rodolphe, BéDARD Yvan and JEANSOULIN Robert (2005), "Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS", *Photogrammetric Engineering & Remote Sensing*, Vol. 11, pp. 205-215.

DOGER Şehat and KURGUN Avşar (2021). "Şarap Üretiminde Veri Kalitesine İlişkin Eksik Veri Sorunlarının Derin Öğrenme İle Çözlülmesi: Üretici Çekişmeci Ağlarla Bir Uygulama", *International Journal of Contemporary Tourism Research*, Vol.5, pp. 99-111.

DÜLGE Senem (2009), *Bilgi Yönetimi Çözümleri ve İş Zekası Projelerinde Veri Kalitesi Yönetimi Uygulamaları* (Master's Thesis), Marmara University Graduate School of Social Sciences, Istanbul.

EARLY Stephens (2021),"The Intersection Of Data Quality And Compliance",*Forbes Magazine*,https://www.forbes.com/councils/forbestechcouncil/2021/03/12/the-intersection-of-data-quality-and-compliance/, DoA. 21.02.2024.

ERYUREK Evren, GILAD Uri, LAKSHMANAN Valliappa, KIBUNGUCHY-GRANT Anita and ASHDOWN Jessi (2021), *Data Governance: The Definitive Guide*, O'Relly, Sebastopol CA – ABD.

FEDERAL CDO COUNCIL (2023), *What is a Chief Data Officer*, https://www.cdo.gov/., DoA. 02.02.2024.

FISCHER Julia, EGLI Lukas, GROTH Juliane, BARRASSO Catherina, EHRMANN Steffen, FIGGERMEIR Heiko, HENZEN Cristian, MEYER Carsten, MULLER-PFEFFERKON Ralph, RÜMMLER Arne, WAGNER Michael, BERNARD Lars, and SEPPELT Ralph (2023), "Approaches and tools for user-driven provenance and data quality information in spatial data infrastructures", *International Journal of Digital Earth*, Vol. 16, pp. 1510–1529.

FOOTE D. Keith (2023), *Data Quality Dimensions*, https://www.dataversity.net/Data-Quality-Dimensions/#, DoA. 21.06.2024.

GALDIES Peter (2014), "The insider threat to data assets", *Journal of Direct, Data and Digital Marketing Practice*, Vol.15, pp.185-186.

GIS-GEOGRAPHY (2017), *World Geodetic System (WGS84)*, https://gisgeography.com/wgs84-world-geodetic-system/, DoA. 18.7.2024.

GOVERNMENT DATA QUALITY HUB (2021), *What is Data Quality ?*, https://www.gov.uk/Government/News/What-Is-Data-Quality, DoA.21.06.2024.

GREGORY Adrian (2011), "Data governance Protecting and unleashing the value of your customer data assets: Stage 1: Understanding data governance and your current data management capability", *Journal of Direct, Data and Digital Marketing Practice*, Vol. 12, pp. 230–248.

GUNES Evrim and SEKERDIL Resat (2022), “A Qualitative Research On The Relationship Between Entrepreneurial Behaviour and Innovative Work Behaviour”, *Journal of Business Economics and Finance*, Vol.11, pp. 88-101.

GUESS A.R. (2013), *Addressing Issues of Data Ownership*, https://www.dataversity.net/Addressing-Issues-of-Data-Ownership/, DoA.11.01.2024.

GUPTA Ankur (2023), *Data Quality: Why, How, Who and When ?*, https://www.linkedin.com/pulse/Data-Quality-Why-How-Who-When-Ankur-Gupta/ , DoA. 21.12.2023.

HENDERSON Deborah, EARLEY Susan and DATA MANAGEMENT ASSOCIATION (2017), *DAMA-DMBOK: Data Management Body of Knowledge*, 2’nd Press, Technic Publications, Basking Ride New Jersey–ABD.

JUGULUM Rajesh (2014), *Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality*, Wiley, New York ABD.

KALE Panoho (2019), “The Age Of Analytics And The Importance Of Data Quality” ,*Forbes Magazine*, https://www.forbes.com/councils/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/, DoA, 21.06.2024.

KING Tim and SCHWARZENBACH Julian (2020), *Managing Data Quality A Pratical Guide*, BCS Learning Institute, England.

KIRAN Bhageshpur (2019), “Data Is The New Oil -- And That’s A Good Thing”, *Forbes Magaine*, https://www.forbes.com/Sites/Forbestechcouncil/2019/11/15/Data-Is-the-New-Oil-and-Thats-a-Good-Thing/?Sh=60af40f07304, DoA. 21.06.2024.

KOCAK Ahmet and ERGUN Mehmet Ali (2023). “Sağlıkta veri kalitesi ve veri madenciliği uygulamaları”, *Disiplinlerarası Yenilik Araştırmaları Dergisi*, Vol. 3, pp. 23-30.

KOMTAS (2024), *DataOPS Nedir?* , https://www.komtas.com/post/dataops-nedir, DoA. 21.06.2024.

KROGSTIE John (2013), A Semiotic Approach to Data Quality, *Lecture Notes in Business Information Processing*, Vol. 147, pp. 395-410.

KUMAR G. Tayi (1998), “Examining Data Quality”, *Communications of the ACM*, Vol. 41, pp. 54-57.

LOSHIN David (2011), *The Practitioner’s Guide to Data Quality*, Morgan Kaufmann OMG Press, Burlington MA – USA.

LINZ (2024), *WGS 84 / Web Mercator tile scale set definition*, https://www.linz.govt.nz/guidance/data-service/linz-data-service-guide/map-tile-services/wgs-84-web-mercator-tile-scale-set-definition, DoA. 18.7.2024.

López Francisco Javier Ariza, González Pablo Barreira, Pau Joan Masó, Torres Alaitz Zabala, Pascualp Antonio Federico Rodríguez, Vergara Gonzalo Moreno and Balboa José Luis García (2020), “Geospatial data quality (ISO 19157-1): Evolve or perish”, *Revista Cartográfica,* pp.129-154, DOI: 10.35424/rcarto.i100.692.

McGILVRAY Danette (2021), *Executing Data Quality Projects*, 2'nd Press, Academic Press, London.

MONSALVES Diego, CORNIDE-REYES Héctor and RIQUELME Fabián (2023), "Relationships Between Social Interactions and Belbin Role Types in Collaborative Agile Teams", *IEEE Access*, Vol. 11, pp. 17002-17020.

MÕISJA Kiira, UUEMAA Evelyn and TÕNU Oja (2018), "The Implications of Field Worker Characteristics and Landscape Heterogeneity for Classification Correctness and the Completeness of Topographical Mapping", *International Journal of Geo-Information*,Vol.7, pp.10.

PORFIRIO Barbara, ADANIYA A. Nicole, JOSKO Borovina Marcelo João and OIKAWA Marcio (2020), "Classification of Errors in Geographic Data Using ISO 19157", *International Geoscience and Remote Sensing Symposium (IGARSS)*, pp.3231–3234, Waikoloa, USA.

RIZKY Syah (2024), "Organizational Behavior Management in Creating Competent Human Resources", *Jurnal Ilmiah Manajemen Kesatuan*, Vol.12, no 1, pp. 012005.

SYAFIK Muhammad and SUHAIBAH Azri (2023), "A review on the GIS usage in spatio- temporal risk assessment in asset management", *IOP Conference Serires: Earth Environment Science*, Vol. 1274, pp. 01-2005

TALEND TEAM (2024), *GDPR's Perspective on Data Quality*, https://www.talend.com/Resources/Gdpr-Improve-Data-Quality/, DoA. 21.06.2024.

Talend Team (2024), *What is data value ?* , https://www.Talend.Com/Resources/Data-Value/, DoA. 21.06.2024.

TOONDERS Joris (2014), *Data Is the New Oil of the Digital Economy*, https://www.wired.Com/Insights/2014/07/Data-New-Oil-Digital-Economy/, DoA. 21.06.2024.

TORSTEN Reimer, TILLMAN Russell and CRISTOPHER Roland (2017), “Groups and teams in organizations”, *The International Encyclopedia of Organizational Communication*, pp. 1-23, DOI:10.1002/9781118955567.wbieoc092.

TUTORIALSPOINT(2024),*Data Quality Dimensions* [Figure],https://www.tutorialspoint.com/computer_concepts/computer_concepts_data_processing_stages.htm, DoA. 21.02.2024.

YUXIANG Luan, ZHAO Kai, WANG Zheyuan and FENG Hu (2023), “Exploring the Antecedents of Unethical Pro-organizational Behavior (UPB): A Meta-Analysis.”, *J Bus Ethics*, Vol. 187, pp. 119–136.