INTERNATIONAL HELLENIC UNIVERSITY | University Center of International Programmes of Studies (UCIPS)

Data analysis from the Greek National Catalogue of Services with the use of KNIME

Koutela Kyriaki
(kkoutela@ihu.edu.gr)

SCHOOL OF SCIENCE & TECHNOLOGY

Postgraduate Program of Data Science

Master Thesis

Supervisor: Berberidis Christos
Supervising Committee Members: Peristeras Vasilios, Magnisalis Ioannis

December 2022
Thessaloniki - Greece

Acknowledgements

I would like to express my gratitude to Dr. Christos Berberidis, my supervisor, for providing me with the opportunity to work on this project and for believing in me. I would also like to thank Mr. Ioannis Konstadinidis and Ms. Elena Kapantai for their guidance and assistance. Additionally, I want to thank my dear friend and former team partner, Mr. Alexios Michailidis, for being a constant source of support throughout my master's program.

Abstract

In the digital age, high-quality data is essential for transparent and effective decision-making, as well as for the provision of important public services. However, ensuring data quality is a complex and multifaceted task that depends on various factors, such as the sources of the data, the processes used to collect and manage it, and the integrity of the systems that store and access it. In this study, we focused on the data quality of information available on the mitos.gov website of the Greek government. Using the KNIME platform, we extracted data from the website and evaluated its quality using a set of quality dimensions. To this end, we developed a model for clustering descriptions of the data records and used it to assign topics to each cluster. Our findings provide valuable insights into the data quality of the Greek government website and highlight some issues that need to be addressed in order to improve the reliability and usefulness of the information available on the website. Based on our analysis, we offer a set of recommendations for improving the data quality and data integration of the website.

Table of contents

1. Introduction

In today's digital world, data is an invaluable resource that drives business, government, and society. The proliferation of information from various sources has created a data-driven economy, where data is the currency that powers decision-making, innovation, and growth. However, with this abundance of data comes the challenge of ensuring its quality. Poorly managed data can lead to incorrect or misleading conclusions, as well as a loss of trust and credibility.

Data quality is a complex and multifaceted concept that refers to the completeness, timeliness, accuracy, uniqueness, availability, credibility, processability, and relevance of data. Data quality is essential for the reliability and usefulness of the information, and it depends on various factors, such as the sources of the data, the processes used to collect and manage it, and the integrity of the systems that store and access it. Ensuring data quality is a critical task for any organization, as it affects the quality of the products, services, and decisions that are based on the data [38].

One common issue that can affect data quality is the existence of data silos, or isolated pockets of information within an organization. Data silos can lead to data quality issues such as inconsistent or unreliable data, as it may not be properly maintained or updated. There are a few indicators that may suggest the presence of data silos within an organization, including lack of data integration, inefficient data sharing, limited visibility into data, and difficulty in accessing data [58]. Data silos can arise due to various factors, such as organizational silos, different technologies, or different processes. They can lead to duplication of data, inconsistencies, and a lack of visibility into the data landscape. Data silos can also hinder data integration and interoperability, as well as the ability to derive insights and value from the data [42].

Data standards and controlled vocabularies can help to improve data quality and reduce the issues caused by data silos. Data standards provide guidelines for how data should be collected, formatted, and shared, while controlled vocabularies provide a consistent set of terms and definitions that can be used to classify and describe data [60].

Using these tools helps to ensure that data is consistently defined and understood across different departments and organizations.

In this thesis, we focus on the data quality of information available on the mitos.gov website of the Greek government. The website is a crucial source of information for citizens, businesses, and other stakeholders, as it provides access to a wide range of public services, policies, and programs. We believe that it is essential for the data on this website to be of high quality, as it is used for transparent and effective decision-making, as well as for the provision of important public services. To this end, we have developed a model for clustering descriptions of the data records. These clusters can serve as a baseline for future searches to identify duplicate records within the same cluster and assess the quality of the other columns of data. By analyzing the data quality of the Greek government website and identifying potential data silos, we hope to provide valuable insights and recommendations for improving the reliability and usefulness of the information available on the website.

2. Data types

Business data refers to data that is generated and used by a business or organization in the course of its operations. This data can come from a variety of sources and can be used for a wide range of purposes [59]. It is important for businesses to have clear data specifications, or guidelines for how the data should be collected, structured, and used.

There are many different types of data that a business might collect and use, including financial data, customer data, product data, marketing data, operational data, and human resources data. The specific types of data that a business needs will depend on its industry, size, and goals.

Having clear data specifications can help ensure that the data collected by a business is accurate, consistent, and relevant to the needs of the organization. Data specifications can include details such as the types of data to be collected, the format in which the data should be stored, and the processes for maintaining and updating the data. Having well-defined data specifications can also help facilitate the sharing and integration of data across different departments and systems within a business [1].

2.1 Data specifications

Any information or documentation that gives perspective, shape, and meaning to data is included in data specifications. They offer the knowledge required to manufacture, construct, generate, evaluate, apply, and handle data and information. Data specifications can be metadata, data standards, reference data, data models, and business rules. Without data requirements, it is challenging to generate high-quality data and even more challenging to assess, comprehend, and manage the quality of data. The benchmark against which to evaluate the results of data quality assessments is provided by data specifications. Additionally, they offer guidance on establishing applications, designing data load procedures, and manually entering data [1]. Basic details regarding each data specification field are provided below.

2.1.1 Metadata

Many refer to metadata as data about data [64]. Metadata is data that identifies, describes, and explains other data. We may think of it as a parent-child relationship where the metadata is the parent and the actual data is the child. Since we can see characteristics and attributes of the parent in the kid but not the real parent, there is a relationship between them. Because of this, metadata are linked to their corresponding data, which makes it simpler to comprehend, utilize, and manage them. Both data and metadata must meet the same quality requirements. The basic facts regarding metadata to learn about include their appearance and nature, like databases, matrices, and descriptions, data types, data domains, valid values and the locations of approved lists of values or standards, as well as variables that govern their validity, such as formats, styles, or patterns. Knowing the size, names, and details of the data field is also vital. It is also important to know whether the data is affirmed by the system, whether it has linked reference tables, whether it was created by the system, and whether it is a must, optional, or conditional. It is also crucial to understand the conditions in which information should be added. Finally, durable URIs should be used to identify objects when describing information and data in accordance with the Linked Data principles [1, 10, 11, 12]. One must make sure the following information is in each field's data definition:

- Complete, precise, and clear title, tag, or name, as well as a description
- Required, alternative, and conditional fields are noted (where the conditions are clear and noted)
- Synonyms or identifiers.
- Acceptable formats or types.
- Lists of acceptable values, reference data, and data standards.
- Each table has a title and a brief description.

2.1.2 Data standards

Data standards are defined as agreed-upon specifications for the representation and manipulation of data from various contexts or resources. They are required in order for data to be shareable, portable, and reusable. Standards for data fields and the value sets used to encode the data within those fields are both included in the concept of

standardized data [14]. The information that needs to be collected for this specification includes identification of fields and tables according to protocols, criteria to obey while inputting data, such as allowed prefixes, case (capital, lower, hybrid), and punctuation, as well as external norms that the business adheres to or to which it is obliged. The following information must be contained in each field [1]:

Regarding table and field names:

- Contrast naming acronyms with names of actual tables, views, fields, etc.
- Make sure any acronyms used in names are recognized standard acronyms.
- In the absence of any explicit naming acronyms, search for any agreement within and among the names.

Rules for inputting data:

- Guidelines to abide by when providing data include using reference data, recognized acronyms, casing (higher, lower, hybrid), punctuation, etc.

Lists of acceptable values include:

- Those determined by a reliable source (reference data).

ISO 8601 is an international standard for representing dates and times. It is a popular example of a data standard, and it specifies a standardized format for representing date and time information in a machine-readable format. The ISO 8601 format consists of a string of characters that represent the date and time in a specific order. The basic format for representing a date is "YYYY-MM-DD", where "YYYY" represents the year, "MM" represents the month, and "DD" represents the day. The basic format for representing a time is "hh:mm:ss", where "hh" represents the hour, "mm" represents the minutes, and "ss" represents the seconds. The ISO 8601 format also includes additional options for representing time zones, dates with reduced precision, and other information. It is widely used in computer systems, databases, and other applications where accurate and consistent representation of date and time information is important [50].

2.1.3 Reference data

Briefly put, reference data are collections of values or categorization that other systems refer to. It can be applied to specify an identifier's features and establish hierarchies based on attribute values. It might originate from corporate sources or from outside data sources. In reference data, one should be interested in collecting the set of permitted values and their descriptions, the names of reference tables holding acceptable values, and the domain and format requirements for lists of values. So for each field, certain actions must be taken [1]:

- Examine the list of acceptable values and their meanings.
- Verify if the values contain metadata.
- Verify that lists of values exclusively include true values.
- Check to see if the entries list has all the values needed
- Establish whether there is no ambiguity when choosing a value and no crossover in the meanings of the values.

2.1.4 Data models

A data model is a way of organizing and structuring data in a database. It describes the relationships and dependencies between different data elements and is used to ensure the integrity and accessibility of data. There are several types of data models, each with their own characteristics and suitable for different kinds of data and purposes. Data models provide a framework for storing and querying data. They also define the list of database states, or state changes, as a collection of fundamental integrity rules. These rules may sometimes be expressed as add-update-eliminate rules. Finding primary and foreign keys, as well as data models that can be applied to a specific dataset, is what is of interest here. Also, it's crucial to understand how many objects from one type of class might be associated with items from a different type, if a field is required, optional, or conditional, as well as proposals for data architecture that are relevant to the project's scope. Some basic guidelines for data models include [1, 13, 15]:

- Seek out names and definitions that are simple to comprehend.
- Check the models for compliance in entities and information linkages.

- Determine how the model is used and distributed.
- Check that the data model's acronyms, which include capitalization and punctuation, adhere to naming conventions.

2.1.5 Business rules

Knowing a company's frameworks, rules, requirements, conditions, and limits is crucial since they set the parameters for data integrity and regulate how a company conducts business. The data should be closely monitored and may be updated whenever the business rules are changed. Additionally, it's critical to understand whether a field is required, optional, or conditional, with the conditions being clearly stated so as to prevent errors [1]. Below, some worth-noting actions are stated:

- Check business rules for precise and comprehensive definitions.
- Keep an eye out for instructions on how and when to handle a document or data field, such as when significant state alterations may occur and how the data should respond. A DQ rule can then be established for the related data behavior and then verified for its compliance with the DQ rule.

2.2 Data integrity fundamentals

Data presence, completeness, validity, structure, content, and other fundamental aspects of data are discussed in Data Integrity Fundamentals. Most DQ dimensions build on the information learned in Data Integrity Fundamentals. In Data Integrity Fundamentals, the data profiling technique is used to assess the fundamental characteristics of the data, including completeness/fill rate, validity, collection of entries and occurrence distributions, trends, intervals, maximum and lowest values, accuracy, and consistency. Since they may all be assessed using the same approach to data monitoring, this includes several separate characteristics of data quality (for example, completeness and validity) [1].

Although the existence of a value is the foundation for completeness, more verification is required to ensure that the values are accurate. Knowing whether fields are required or not, as well as the circumstances, is necessary in order to evaluate results. Completeness should be 100% if the field is necessary. Count and percentage of records

with values in each of the mandatory fields should also be calculated. Further investigation should be done if the percentage is less than 100%. The application must be updated if the field is needed by the business but not by it. If that is not possible, list the conditions for data entry and explain to those who do the entries dated if the field is needed by the business but not by it. If that is not possible, list the conditions for data entry and explain to those who do the entries what they should enter and why. In this instance, the quality of the data must be closely reviewed. On the other hand, if the field is required by the application but not by the organization, implying that the data to be entered is not significant or known, meaningless data is frequently submitted to satisfy the technical requirement for a response in the area. [1].

3. Data quality

3.1 Definition of data quality

Data quality (DQ) is a metric that shows how well a data set fits the needs of a company. High quality data is essential for making trustworthy decisions, and the definition of DQ may vary depending on the specific requirements of a business [3]. Some key aspects of DQ to consider include accuracy, completeness, timeliness, etc. Ensuring that data meets these standards is crucial for making reliable decisions. DQ has evolved into the key factor needed to show a return on the most data-oriented initiatives and activities. DQ is the key determinant of success or failure in all data-related endeavors, from data analytics, data science, data governance, metadata management, and interoperability enhancement [1], to problems concerning computer science, quality control, human factors research, and statistics [2]. Due to the fact that the definition of DQ relies on the needs of the business, a specific data collection may be regarded as high-quality by one company but low-quality by another. There are various ways to define DQ, but generally speaking, it refers to the state of the data at any point in its entire lifespan [3].

3.2 Data quality dimensions

It's crucial to comprehend the components of DQ in order to guarantee the reliability of the data. DQ dimensions compose the metrics to assess if the data is suitable for usage or whether revisions and enhancements are required. It will not be possible to evaluate the data quality using just one dimension. A trait, feature, or attribute of data is referred to as a DQ dimension. Information and data effectiveness requirements can be categorized using dimensions, and they are employed to categorize, quantify, enhance, and manage the caliber of data and information [1].

When examining various works of literature, one can see that there are a number of theoretical frameworks, set of metrics, dimensions, and characteristics for evaluating DQ, and that the same dimensions frequently appear under different names and with little variations in their definitions. Since some frameworks are only applicable for certain fields [2], choosing the right metrics for data quality requires understanding of the

business requirements or nature of the data. One must, for instance, be aware of the structure of their data [2], the demands of their organization, the intended use of the data [3], the general needs of the users [2], and the areas of data quality that need to be improved [4]. In accordance with the Data Management Association UK and the official portal for European data, organizations have evaluated the quality of their data based on these main dimensions: Completeness, Timeliness, Accuracy, Uniqueness, Availability, Credibility, Processability and Relevance [3, 11].

3.2.1 Completeness

Completeness can be defined as the degree to which the data are suitable in broad, depth, and complexity of the project for the task being performed [2] or whether the information complies with the requirements or specifications of the database [6]. This dimension should simply provide answers to the questions of whether the data are comprehensive and how detailed they are [3]. Simply said, it is the percentage of entities and attributes that are present in the schema [5], or alternatively, the quantity of null, missing values, errors and omissions in a particular data set [3, 6]. The ability to do data mining and analysis in a variety of applications will be hampered by a data collection that contains too many gaps. For instance, if a field requests that the user provide personal information and they do so in all but one of the optional fields, the data set will be deemed complete. In order to improve the quality of this dimension, one should assess whether all of the requisite information is available, and whether there are any missing elements. Data and metadata must be stated so as to understand what they are about and to avoid changes whenever new entries arrive. One must have knowledge on where the data are coming from, what they contain, what is their theme and their purposes. Moreover, datasets should have a version number which is going to change when updates take place. In that way, one can monitor the changes [16].

Data, model, attribute, value and feature completeness are the different types of completeness that can be distinguished. Data completeness is the quantifiable gap between the database and the requirements, and it is independent of application. When all of the objects listed in the specification are present, even highly generalized databases can be considered complete. Model completeness is defined as the compatibility between

the concrete description and the specification needed for a certain database application. Due to it being application-specific, model completeness is a factor in usability. The degree to which each important property of a feature has been encoded is known as attribute completeness. Value completeness is the aspect to which values exist for all characteristics. According to Brassel, Bucher, Stephan, and Vckovski's publication "Elements of Spatial Data Quality", feature completeness can be defined over space, time, or theme. They use the example of a database that shows the locations of buildings in a particular state, and they claim that the database is considered to be incomplete based on space, time, and theme if it only contains entries about buildings that are located only in a portion of the state, if it lacks building entries from a specific time period, or if it only includes residential buildings and excludes non-residential ones, respectively [6, 7].

The metric of completeness as a percentage is defined as [4]:

$$Completeness = \left(1 - \frac{number\ of\ incomplete\ cells}{number\ of\ cells}\right) * 100$$

Where

$$number\ of\ cells = (number\ of\ rows) * (number\ of\ columns)$$

It shows the percentage of complete cells in a dataset. It refers to the cells that are not empty and that have significant data written into them. The normalized metric of completeness is defined as:

$$Normalized_{Completeness} = \frac{Completeness}{100}$$

Figure 3.2.1 (1): High completeness example [11].

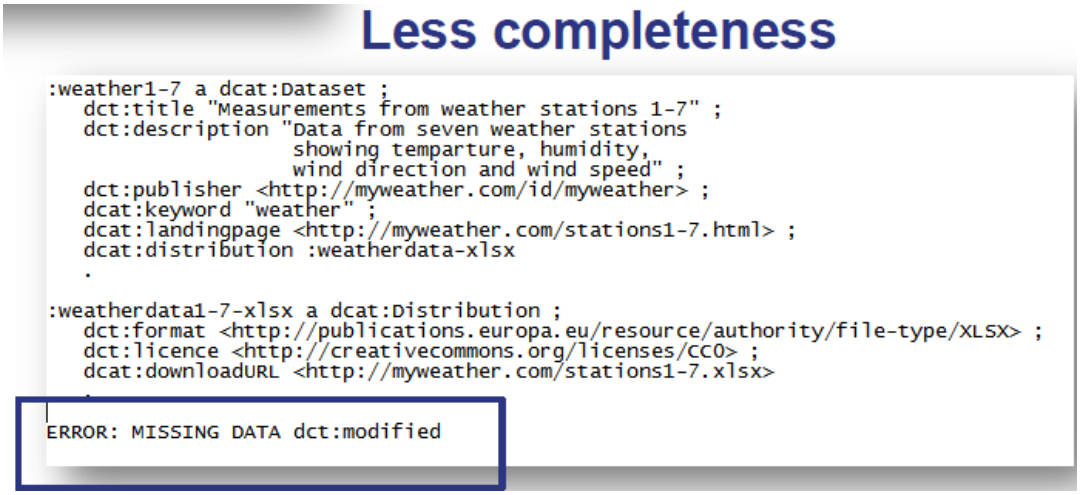

Figure 3.2.1 (2): Less completeness example [11].

3.2.2 Timeliness[1]

Timeliness measures how updated the data is and whether the age of the data is adequate for the task at hand, as data that is out of date won't be useful to an organization [2, 5]. The frequency of (meta) data updates is another aspect of timeliness. Usage describes the degree to which metadata keys are included to describe the data [8]. Additionally, this dimension considers whether or not the data was available at the appropriate moment or whether it came on schedule. For instance, data is timely if it is

[1] Often referred to as currency.

provided when it is needed [3, 5]. Due to this, many refer to timeliness as a factor that reflects user expectations.

The metric of timeliness determines how old the current attribute value is and is defined as [9]:

$$Timeliness = \frac{1}{(mean\ attribute\ update\ frequency) * (age\ of\ attribute\ value) + 1}$$

Where the mean attribute update frequency indicates how frequently the value of the considered attribute is changed on average over a predetermined time period (f.e., 5 times per month). Regarding the output value provided by this function, we can state that the closer it is to 0, the more out-of-date it is, and the closer it is to 1, the more current it is. This can be easily confirmed by assuming that if the mean attribute update frequency is 0, meaning that it never becomes out of date, then the timeliness is 1, meaning that it is up to date. Additionally, we get the value of 1 for timeliness if the age of the attribute value is 0 (i.e., if the attribute value was newly acquired).

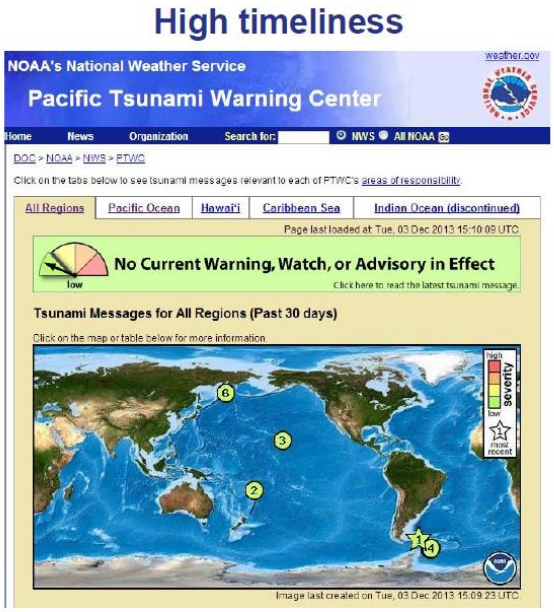

Figure 3.2.2 (1): High timeliness example [11].

Figure 3.2.2 (2): Less timeliness example [11].

3.2.3 Accuracy

Information that doesn't comply with certain syntax standards (such as format, type, or range) or doesn't meet business rules is considered to be invalid, according to the data quality dimension of accuracy [3]. Checking whether all of your information adheres to a particular format or set of business rules is thus an atypical metric for this dimension. The degree to which data adheres to a set of guidelines, criteria, or standards is important. For example, the WGS84 and SDMX formats are commonly used for coordinates and statistics. It is advisable to use regional vocabulary and the most widely accepted standards in the area where the data or metadata is most relevant [11]. Also, default values need evaluation to determine whether they are correct or not [16]. Popular examples include date formats, since many systems ask for dates to be entered in a particular format, and file types, as many programs only accept files in csv format.

Accuracy can also be defined as the extent to which data is presented in the same format, is compatible with previous data, and aligns with commonly accepted definitions. [2, 3]. Many businesses keep the same information in multiple locations, and if the information matches, it is seen as consistent [1]. This dimension can be viewed in a variety of ways, but often refers to the number of records that contravene integrity

constraints [2]. For example, whether or not data held in one location corresponds to pertinent ones stored elsewhere, if there is consistency of identical or duplicate data points across tables (like each entry in one row or column follows the same format and size), or if the system indicates that someone is being paid when they actually no longer work for the organization. Therefore, in order to resolve problems with inaccuracy, it is vital to examine the data sets to see whether they are consistent across all instances, if any records breach referential integrity, and if there are any instances where the information contradicts itself [3, 5].

Another way to measure accuracy is by how well the data accurately represents the event or object it is depicting and how closely it reflects the actual environment it is trying to represent. The information should be reliable and accurate enough for its intended use [16]. It is also referred to as error-freeness in other literature on accuracy, which is another way of saying how precise, trustworthy, and certified the data are [2, 3]. It should go without saying that correct data is crucial for usefulness. The percentages of inaccurate values are an example of those measures [2]. For instance, the system should not report someone's birth year as 1997 if they are 40 years old, hence, it is necessary to determine if the data requires aggregation or disaggregation [16]. Therefore, in order to increase the accuracy metric, both inaccurate data and data that do not accurately reflect the environment must be changed.

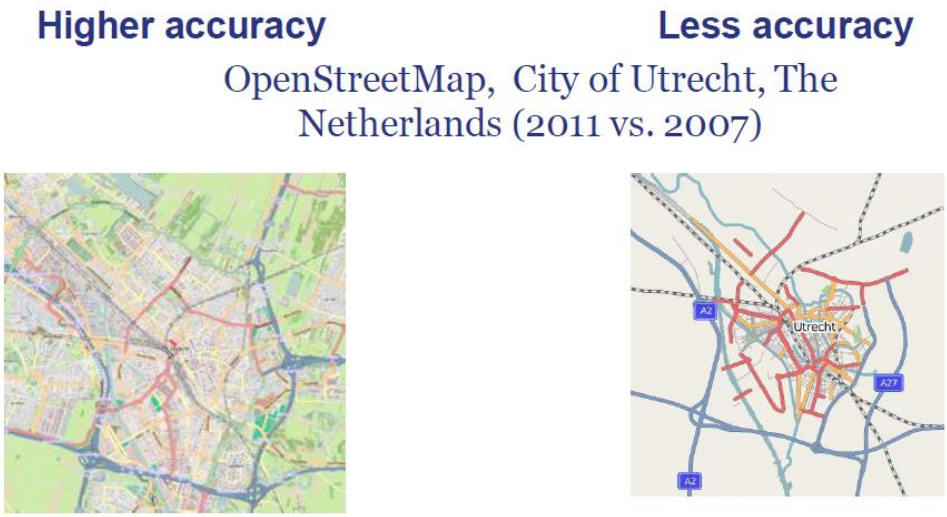

Figure 3.2.3: Accuracy example [11].

3.2.4 Uniqueness

Information is said to be unique if it only ever appears once in a database or system. To meet this data quality requirement, you must check your data to make sure that no duplicates exist. It's crucial to keep in mind that data duplication might occur even when two entries share all other characteristics but differ just in their names [3]. Duplicate records have a lot of unintentional expenses. The credit limit for one firm may unintentionally be crossed when purchases made by that company are linked to duplicate documents, putting the company at an extreme financial risk [1].

To detect duplication, it is necessary to know which data field indicates a record's uniqueness. However, de-duplication tools should be formed as these algorithms might be based on regions that are currently missing data, meaning that the completeness metric would be low. Additionally, the data may be inaccurate or of low quality (e.g., missing values), or it may contain information that shouldn't be there (such as when an identification number is entered in error for a street name field), or the quality of the data is poor (e.g., incorrect values). The de-duplication method's output will be inaccurate if the input is erroneous [1].

3.2.5 Availability

Data availability refers to both how easily data can be retrieved and how persistent it is over time. Instances of this dimension would be a dataset with a consistently defined http: URI, leading to the appropriate resource and doesn't produce an error and the data's summary is located in the data portal's search service. So, for that reason, the design and maintenance of the data's URLs is crucial, as well as the person or organization that is going to undertake that construction [11].

3.2.6 Credibility

The credibility dimension is about how reliable the origin of the data is, like government or official data, and easily verified ones. As a result, in order to comply with this dimension, one should look for datasets that are widely accepted, and if one creates their own data, it must be done in a way that others may trust it [11].

Figure 3.2.6 (1): High credibility example [11].

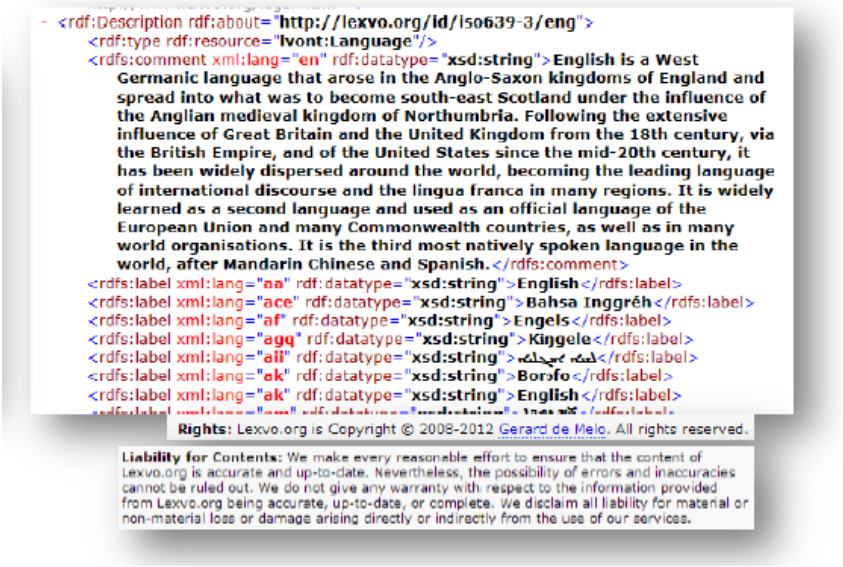

Figure 3.2.6 (2): Less credibility example [11].

3.2.7 Processability

The degree to which data can be comprehended and handled by automated systems is known as its processability. Consider a dataset that contains data such as date and time in a W3C format rather than a text that has been codified using publicly available controlled vocabulary, standards, and code lists.

Figure 3.2.7 (1): High processability example [11].

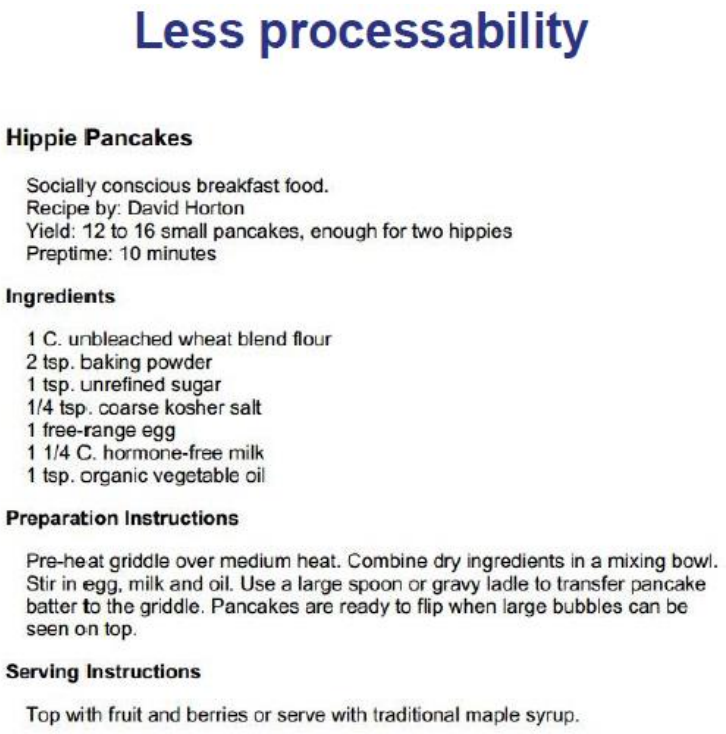

Figure 3.2.7 (2): Less processability example [11].

3.2.8 Relevance

The relevance of data is determined by the existence of necessary information on the datasets so as to enable their use in the application. The main purpose of this dimension is to save time and money by providing the appropriate constraints and details for their current or future use case. For example, in temperature data calculations where rounded quantities for Fahrenheit or Celsius already exist or a dataset containing numbers in acres for measuring the surface of a field. It is important to remember that data features that are useless for their application should be eliminated [11].

In order to determine the relevance of a piece of information, it is useful to consider factors such as the topic of the information, the purpose of the information, and the intended audience of the information. This can help to determine whether the information is directly related to the topic being researched, meets the needs of the person seeking it, and is intended for the same audience. Assessing the relevance of information is an important aspect of information retrieval and search engines, as it is often used to rank and prioritize search results based on how relevant they are to the user's query [54].

High relevance

	Engine (cm3)	Fuel type	CO2 (g/km)	Tax (%)
Car Type 1	900	Gasoline	90	0
Car Type 2	1.100	Gasoline	120	5
Car Type 3	1.300	Gasoline	125	5
Car Type 4	1.400	Gasoline	150	5
Car Type 5	1.800	Diesel	180	10
Car Type 6	2.200	Diesel	190	10
Car Type 7	2.500	Gasoline	210	15

Figure 3.2.8 (1): High relevance example [11].

Less relevance

	Engine (cm3)	Fuel type	CO2 (g/km)	Color	Tax (%)
Car Type 1	900	Gasoline	90	Red	0
Car Type 2	1.100	Gasoline	120	Silver	5
Car Type 3	1.300	Gasoline	125	Black	5
Car Type 4	1.400	Gasoline	150	White	5
Car Type 5	1.800	Diesel	180	Silver	10
Car Type 6	2.200	Diesel	190	Blue	10
Car Type 7	2.500	Gasoline	210	Black	15

Figure 3.2.8 (2): Less relevance example [11].

4. European Data Standards and Controlled Vocabularies

Data standards and vocabularies are crucial for maintaining data quality as they help to ensure consistency and accuracy in data representation and storage. Standardized data definitions, formats, and structures minimize errors and inconsistencies, particularly when data is shared or integrated from multiple sources. They also ensure interoperability and ease of understanding for different systems and individuals. Vocabularies, such as controlled lists of terms and definitions, standardize language and terminology used to describe data, reducing ambiguity and improving accuracy by promoting consistent and correct usage of terms [20]. In summary, data standards and vocabularies are vital for maintaining high quality, reliable, and trustworthy data [60].

4.1 Data Standards

According to the Federal Enterprise Data Resources, the term “standard” refers to a reference that is widely accepted and that enables the value, approval, or sharing of an item, a procedure, or a quantity of information. Depending on the domain, the word has a different meaning. In the context of digital technology, the term "standard" frequently refers to a general set of specific requirements for how data is defined, analyzed, or transferred, as opposed to the government sector, where the word is frequently used to refer to a demand, a conformance measure, or a set of guidelines and requirements that an item must satisfy. To fulfill consistency and interoperability concepts, standards are frequently used for various systems, resources, and individuals.

Data standards are frequently composed of individual data standards, also known as "components," and a number of components may be combined to form a more complete "set" of data standards. A system or software that gathers a group of components to be utilized and grouped together for a variety of applications is known as a data standards framework.

A data standards package details how the majority of the various parts of a complete data standard system should be implemented. Although this is generally alluded to as an application portfolio, several frameworks and groups have alternative

definitions for their own collection of parts. For instance, this is known as Information Exchange Package Documentation in the context of NIEM (IEPD).

An adaptable or complete set of reused data standards components is known as a data standards framework, and it can be put together to create an almost endless number of data standards packages. They are designed in order for elements to be combined and matched to satisfy a variety of applications. The National Information Exchange Model (NIEM), the Dublin Core Metadata Initiative (DCMI), and Schema.org are a few examples of data standards frameworks [17].

Data types, identifiers, vocabulary, schema, format, and API are a few of the components of common data standards [17].

4.1.1 Data type

Data type is a way of describing the kind of information a piece of data is. Data types that are often used include boolean, array, numerical (integer or float), string, and date.

4.1.2 Identifiers

A unit of data or a particular element in a codelist or categorization system may be referenced quickly and easily using identifiers. An identifier is a string of letters that uniquely identifies an item. It might be local, like UIDs (Unique Identifiers), or it can be global, like UUIDs (Universally Unique Identifiers). Despite UUIDs, identifiers may also employ widely recognized code lists and categorization systems. For instance, the ISO 3166-1 alpha-2 code for Greece is "GR," but the North American Industry Classification System (NAICS) code for Motorcycle, Boat, and Other Motor Vehicle Dealers is "44122". Identifiers that may be retrieved online are referred to as "permanent identifiers" in this context. This type of identification is active and durable since it is a Uniform Resource Identifier (URI) for the hypertext transfer protocol secure (http/s), which you may put into the url bar of your web browser to navigate to the destination. A URI pattern and a local identifier (local ID) are the two parts of a URI [17, 18].

Figure 4.1.2: Parts of a URI [18].

4.1.3 Vocabulary

A vocabulary defines phrases that are consistent in their semantic meanings and often restricted to a certain namespace or area. The codes are the individually or collectively machine-readable identifiers for the data pieces. In a worksheet with data tables, they may be compared to the column headers, and their descriptions can be found in a data dictionary [18].

4.1.4 Schema

A schema can define the connections between relevant data in a data model or database schema. SQL, UML, XSD, or JSON Schema can all be used to express schemas [18].

4.1.5 Format

The format specifies the document format, encryption, syntax, and application type used to store, display, or transport your data. Online, formats are linked to specific media IDs like json. The basic data formats are XML, CSV, and JSON [18].

4.1.6 API

An application programming interface (API) is utilized as a predetermined framework when reading or writing data from a filing system, data storage, or even a network. REST APIs for the internet and OBDC and SQL for data storage are examples of common data API types [18].

4.2 Controlled Vocabularies

A controlled vocabulary is a set of terms and expressions that have been arranged for the purpose of indexing content and retrieving it via searching. They are essential for enhancing data accessibility. It often has a defined field of use, includes recommended

and modified terminology, and describes a certain domain [19]. Vocabularies are a helpful tool for transmitting information and titling objects with significance at a given level of detail [20]. Since controlled vocabulary enables consistency in desired keywords and attaches similar terms to relevant information, it is crucial because, without it, libraries would not reliably employ the exact term to identify the same individual, location, or object in the indexing stage. Users need to be guided through the search process since they might not be familiar with the right keyword, causing them to search for a specific topic using a variety of similar words or more general terms. The main goals of a regulated vocabulary are to gather alternative terms and synonyms for themes, connect them logically, or organize them into divisions. A controlled vocabulary's linkages and connections make sure that these interconnections are established and kept up to date between categorization and lookup [19]. These connections must have a mutual understanding, and the greater the constraint there is over the terminology in use, the less likely it is that there will be misunderstandings or unintended search results. This is because richer meanings require more control because they add information or allow for differentiation [20]. Thesauri, taxonomies, basic controlled lists, synonym ring lists, and controlled vocabularies are among the most popular types of vocabularies. Current controlled vocabulary resources are accessible in various forms, including textual, spreadsheets, and XML [21].

4.2.1 Controlled Lists

A controlled list is a straightforward dictionary often used to restrict words and their meaning. Every entry is distinct (meanings do not conflict), belongs to the exact same category or classification scheme, has an equivalent degree of specificity, and is organized alphabetically or logically. When utilizing controlled lists, one must take caution since they might have to be generated from bigger vocabularies or they may have been created for a particular database or circumstance and might not be useful in other contexts. To guarantee uniformity between libraries, descriptions of the terminology should be made accessible [19].

4.2.2 Synonym Ring Lists

A basic group of concepts that are deemed similar for recovery purposes is referred to as a synonym ring. In the majority of controlled vocabularies, only objects with authentic synonymy or equivalent semantics must be assumed equal. Synonym rings, on the other hand, are differentiated and, despite being categorized as controlled vocabularies, are nearly always utilized for retrieval instead of classification or indexing purposes. For that reason, they are employed expressly to query expansion (broaden retrieval) and they may really encompass near-synonyms which have alike or related interpretations instead of limiting themselves to a few words having genuine synonymy. Synonym rings are usually employed in the background of a digital data network as collections of flat lists. They are especially helpful at giving information that is expressed in words as well as other situations of unrestricted language. Since they are knowledgeable about the unique information content, the requirements of the users, and potential queries or search results, experts manually create the most effective synonym rings [19].

4.2.3 Taxonomies

An organized categorization for a specific field is called taxonomy. It consists of hierarchically structured controlled vocabulary phrases where each word in taxonomy is related to other words in the taxonomy with one or even more parent-child (wider or narrower) connections. There are several kinds of family connections that can exist, including complete or partially, genus or kinds, as well as instance ties. In actuality, though, every child of a specific parent has the same kind of bond. Taxonomy and a thesaurus could be similar, although taxonomy often contains shorter hierarchies and a simpler structure. As an instance, it frequently lacks equality (synonyms or altered phrases) or relevant terminologies (dependent relationships). Taxonomies are commonly used to describe how organisms and plants are classified scientifically. The term "taxonomy" can also, in general usage, apply to any categorization or grouping of keywords or headers, especially a controlled vocabulary that serves as a navigational framework for a webpage [19].

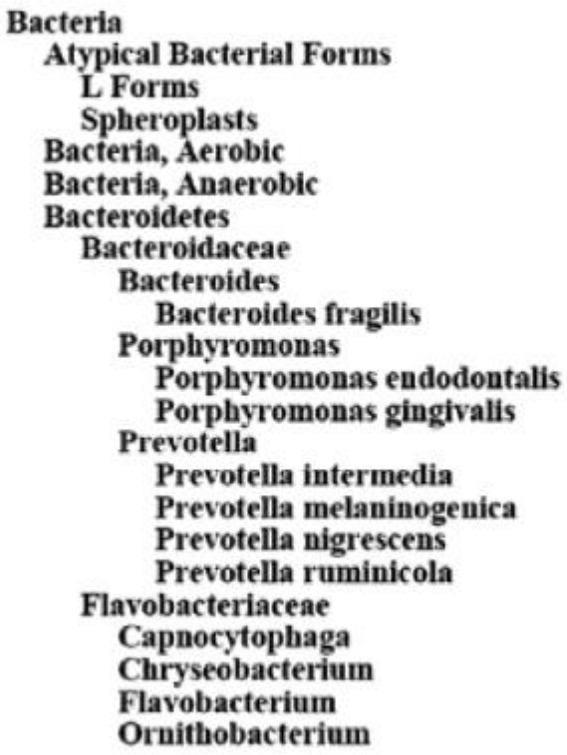

Figure 4.2.3: An example of the taxonomy of the Flavobacteriaceae family with siblings and wider/narrower contexts from the U.S. National Center for Biotechnology Informations [19].

4.2.4 Thesauri

A thesaurus integrates extra elements with the qualities of synonym ring lists and taxonomies. A thesaurus is a logical system with semantic characteristics of distinct concepts that includes connections between synonyms, scenarios that are more general or more specific (family relationships of parent and child), as well as other linked ideas. It can contain one or more languages. A thesaurus may include relationships such as equality (synonym), hierarchy (entire/partly, genus/kinds, or incident), and synergistic. Additionally, definitions, citations, and other ancillary or descriptive material regarding a topic may be included in the thesaurus. A thesaurus is more complicated in comparison to a basic list, a synonym ring list, or a typical taxonomy and they are often regarded as the most complicated controlled vocabularies [21]. Thesauri use the robust and adaptable vocabulary management that is often advised for use as authority in databases referring to art works and cultural history. Every controlled vocabulary shown with predefined link

characteristics, organized in a recognized sequence, and frequently utilized for exploring in after coordinated data storing and retrieving systems can also be referred to as a thesaurus [19].

4.2.4.1 Eurovoc

Eurovoc is a multilingual, multithematic thesaurus with an emphasis on European Union (EU) institutions, law, and legislation. It provides over 8000 entities throughout all 23 of the EU's primary languages and three additional ones that correspond to EU applicant nations. Many information organizations in the EU, libraries and research departments of legislative bodies, as well as other governmental and commercial institutions in related (and non-related) countries of the EU [21, 22]. It is presented in accordance with the W3C Company's approved semantic web standards and incorporates the most recent developments in thesaurus standards. Such tools not only provide modeling adaptability and reasoning skills, but they also make it easier to link EuroVoc with some other controlled vocabularies and thesaurus, like national and field-specific vocabularies. When it comes to enhancing accessibility to research topics in European and national legal documents and several other data resources, EuroVoc is perhaps the most practical controlled vocabulary. It includes every area of work for the institutions of the EU. Based on EU legislation and areas of activity, the themes are divided into 21 domains and 127 smaller concepts. EuroVoc is referred to as a language independence thesaurus since it enables users to search for a topic in their language and return results in another language [22].

EuroVoc can be downloaded from the EU Open Data Portal in the form of XML and SKOS/RDF. When combined with certain other legal materials provided by the Publications Office for reutilization and data linkage, EuroVoc is accessible within the SPARQL endpoint of the Publications Office [22].

To connect topics in a way that is obvious and understood, Eurovoc uses thesaurus relations or abbreviations. In the framework of EuroVoc, concepts are arranged in both hierarchical and non-hierarchical relationships. When the domains of two concepts overlap completely, hierarchical linkages between them should be developed in order for them to be more clear and understandable. In hierarchical relationships, a class

is represented by a bigger concept, while its children or components are described by a smaller concept. "BT" and "NT" are the acronyms used as indicators of a hierarchical relationship. "BT" (sometimes known as a general or greater term) is used to show the connection between a certain notion and a more general one, as well as the degree of hierarchical layers between every broader term and the specific one. The abbreviation "NT" (narrower term) is applied to denote the link between a broad concept and a more expert one, and shows the sequence of hierarchical layers that separate each narrower term from the larger term [23].

Other abbreviations that are used to link concepts together are the following: "RT", "UF", "USE", "SN", and "HN". The connection of two ideas that are thematically or logically related is denoted by the acronym "RT", or related term. Since the user is presented with topics that have related or comparable meanings and could be utilized for indexing, classifying, or searching, the link within concepts must be made clear. The prefix "RT" is inserted before related topics. Often, a subject can be expressed by multiple words pertaining to the same concept. When that happens, one phrase should be identified as the desired label and the rest as the least preferred ones. The favored phrase is preceded by "USE," while the unfavorable term is preceded by "UF." A scope note “SN” defines the thesaurus term's purpose. When another idea is mentioned in a scope note, there needs to be a separate scope note for every one of those additional concepts. In other words, it establishes or explains a concept's meaningful bounds as they relate to its usage in the organized vocabulary. The retrievability of a topic can be impacted by shift changes whenever a thesaurus is modified. A history note, or "HN," must be made in certain circumstances. A brief explanation of the concept's history and evolution includes information such as whether it has altered through periods or has replaced any other terms [23].

5. NLP methods for improving data quality

Unstructured data is difficult to process and analyze since it is created from a variety of sources and formats. There is a rising demand for new automated methods for document analysis as the quantity and volume of textual information rise dramatically. An important problem in many Natural Language Processing (NLP) applications, including document categorization, machine translation, plagiarism detection, information retrieval, and text summarization, is determining how similar two texts are to one another. It is therefore frequently used to compare papers, chapters, or sentences to find topical agreement or resemblance.

To ensure data quality in NLP, it is important to carefully pre-process and clean the data before using it as input to NLP algorithms. This can involve tasks such as correcting errors, standardizing formatting, and removing duplicates. It may also be necessary to use controlled vocabularies or other data standards to ensure that the data is consistently defined and understood [61].

5.1 Natural Language Processing tools

To effectively analyze and process unstructured data, natural language processing (NLP) relies on various tools such as programming languages, software, and libraries. Tensorflow and Python are open-source programming languages commonly used for analyzing large amounts of data from computing devices. Python offers several libraries that support NLP tasks, such as the Natural Language Toolkit (NLTK) and Pandas, which assist with importing, preprocessing, and cleaning unstructured data. The Numpy library is also useful for scientific computing with large multidimensional arrays and matrices [24].

For high-performance parallel computation on the Graphics Processing Unit, NVIDIA created the CUDA toolkit (GPU). It is tailored for GPU-accelerated projects and comes with GPU-accelerated packages, compilers, development tools, and a runtime environment. The framework is well-known for its capacity to carry out effective

computational primitives and is employed in variety of domains, including deep learning, image, and graph analysis [24].

5.2 Text similarity

Text similarity can be calculated using a variety of methods, for instance the Jaccard index, cosine similarity, and Levenshtein distance, along with a variety of programs and libraries, like NLTK, Gensim, and scikit-learn [24]. Text similarity is used by search engines to rank search results based on their relevance to a query. It is an important aspect of information retrieval and is often implemented using vector models. There are two types of text similarity: lexical and semantic. Lexical similarity measures the similarity of strings based on their surface features or sequence, while semantic similarity measures the similarity of words based on their context [24].

The similarity between two documents can be determined by comparing the term vectors for each document and examining the correlation between them. The cosine angle among these vectors indicates how similar they are, with a higher text similarity when the angles between the vectors are smaller. The definition of an orthogonal angle is one greater than 90 degrees, leading to unconnected sentences [24]. Cosine similarity is widely used to evaluate the similarity between text documents in various contexts, including information retrieval and clustering. One key feature of cosine similarity is that it does not depend on the length of the documents being compared. This means that documents with the same content but different lengths will be treated the same when using cosine similarity to measure their similarity [25].

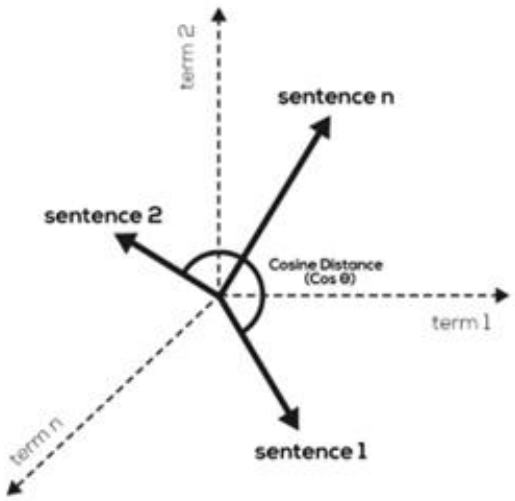

Figure 5.2 (1): Unrelation between sentences 2 and n [24].

The following equation shows the Cosine similarity between two sentences:

$$\cos(S1, S2) = \frac{S1S2}{\|S1\|\|S2\|} = \frac{\sum_{i=1}^{n} S1_i S2_i}{\sqrt{\sum_{i=1}^{n}(S1_i)^2}\sqrt{\sum_{i=1}^{n}(S2_i)^2}}$$

where S1S2 is the combination of two phrases' dot products, multiplied by their combined product magnitude $\|S1\|\|S2\|$ [24].

5.3 Text preprocessing

Text pre-processing involves the actions taken to prepare text for analysis in order to reduce the amount of data and improve the efficiency of subsequent analysis. Usually it starts by removing punctuation, as it does not carry meaning. Next, the text is transformed to lowercase, as words in different cases have the same meaning. The text is tokenized, where it is divided into individual tokens, and common words that do not carry meaning, or stop words, are removed. Then stemming is applied to the remaining words by reducing them to their root form to eliminate variations in word forms [26].

5.4 Sentence embeddings

Sentence embeddings are numerical representations of sentences or phrases that capture their meaning in a compact, low-dimensional space. They can be created by training a neural network on a large dataset of labeled sentences and their corresponding tasks, such as sentiment analysis or question answering. These embeddings capture the

meaning of the sentences in a way that can be used as input for machine learning models and are useful for NLP tasks because they allow for the efficient processing of large amounts of unstructured text data [27].

Sentence embeddings can be very effective in performing various tasks, even with a small amount of task-specific training data (Yang et.,al 2018). In the "Universal Sentence Encoder" research paper, several models were introduced for encoding sentences into embedding vectors that can be transferred to other NLP tasks. These models were efficient and produced accurate results on a range of transfer tasks. The research found that sentence embeddings can be easily used to compute semantic similarity scores with excellent performance on the STS (semantic textual similarity) task. Additionally, these sentence encoding models can be fine-tuned for specific tasks by adjusting the model's parameters using gradient-based updates when incorporated into larger models [27].

5.5 Sentence transformers

Using BERT-based models, the SentenceTransformers Python module offers sentence and text embeddings. Through techniques like cosine similarity, it may be used to locate sentences with comparable meanings in more than 100 different languages. Tasks like semantic textual similarity, semantic search, and paraphrase mining can all benefit from this. SentenceTransformers utilize Sentence-BERT (SBERT), a top performing method for generating sentence embeddings. SBERT is built on transformer models such as BERT and involves taking the mean of the output [28, 29].

The multilingual version of the model was developed by taking a monolingual model and adding support for additional languages by expanding the embeddings. The training process involves mapping translated sentences to the same location in the vector space as the original sentences, using the monolingual model to generate sentence embeddings for the source language and then training a new system on the translated sentences to mimic the behavior of the original model. This method is simple to use to add support for new languages and requires little data. It also helps keep the desired characteristics of the vector space, and it requires fewer resources for training. It has been

shown to work well with over 50 languages from various language families and can be easily expanded to cover more than 400 languages with the available code [28].

5.6 Clustering

There is a high volume of text documents that is constantly growing and includes those found online, in digital libraries and repositories, and in personal digital information. This large amount of text creates problems in terms of effectively organizing and managing it. Clustering is a method that can be utilized to automatically organize a large number of data objects into a smaller number of cohesive groups. This approach has been successful in organizing text documents and continues to be a topic of research [25].

In order to facilitate effective browsing and navigation, the clustering technique divides up a massive collection of unstructured text documents into a smaller number of coherent and relevant clusters. There are various methods for determining the distance or similarity between documents, such as squared Euclidean distance, cosine similarity, and relative entropy. In addition to aiding in navigation, clustering can also assist in further analysis and utilization of the dataset, such as for information retrieval and extraction, by grouping together similar types of information sources. Clustering is often used to organize search engine results in a way that helps users find and focus on the most relevant information. This involves grouping similar documents together and separating different documents into different clusters. Nevertheless, determining what makes two documents similar or different can be challenging and may depend on the context. For instance, when clustering research papers, two documents may be considered similar if they have similar themes. Determining a measure of similarity or distance between objects is a crucial step in the clustering process. This measure reflects the degree of proximity or separation between the objects and should be based on the characteristics that are believed to differentiate the clusters within the data. However, the specific characteristics that are relevant can vary depending on the data or context of the problem, and there is no universally ideal measure for all clustering situations [25].

One way to cluster and perform semantic search is to use SBERT to transform each sentence into a vector in a vector space, such that semantically similar sentences are located close to one another. With SBERT, the process of finding the most similar pair of

sentences within a set of 10,000 sentences becomes significantly more efficient, taking only the time needed to compute 10,000 sentence embeddings instead of the 65 hours required with BERT [29].

In this research, we are going to use agglomerative clustering. It is a method of clustering that involves creating clusters by merging individual data points into progressively larger groups. It is a bottom-up approach to clustering, in which each data point starts out in its own cluster, and pairs of clusters are merged together as the algorithm progresses [49]. The main advantage of agglomerative clustering is that it is relatively simple to implement and can be applied to a wide range of data types. However, it can be sensitive to the order in which clusters are merged, and it can be computationally expensive for large datasets.

There are several different methods for defining the distance between clusters and deciding which pairs of clusters should be merged at each step of the algorithm. Some common methods include single linkage, complete linkage, and average linkage. Single linkage clustering involves merging the two closest clusters, where the distance between clusters is defined as the minimum distance between any two points in the two clusters. Complete linkage clustering involves merging the two clusters that are furthest apart, where the distance between clusters is defined as the maximum distance between any two points in the two clusters. Average linkage clustering involves merging the two clusters with the smallest average distance between all pairs of points in the two clusters [48].

Agglomerative clustering can be useful for exploring the structure of a dataset and identifying patterns and relationships within the data. It is often used in combination with other clustering methods or as a preprocessing step for other machine learning tasks.

5.7 PCA

Principal component analysis (PCA) is a statistical technique that is used to reduce the dimensionality of a dataset. It is a common method for data preprocessing and feature extraction in machine learning and data analysis. PCA is based on the idea that the most important information in a dataset is often contained in the patterns and trends that are present across multiple variables. By identifying and projecting these patterns

onto a new set of coordinates, PCA can reduce the complexity of the data and highlight the underlying structure [51].

To perform PCA, the data is first centered and scaled, and then a set of orthogonal "principal components" are computed. These principal components are linear combinations of the original variables that capture the maximum variance in the data. The first principal component is the one that captures the most variance, followed by the second principal component, and so on [52].

PCA can be used to visualize high-dimensional data, identify patterns and trends in the data, and reduce the complexity of the data before applying machine learning algorithms [51]. It is a useful tool for exploring and understanding complex datasets, but it is important to note that it can also discard important information that is not captured by the principal components [53].

5.8 Topic modeling

Topic modeling is a technique in natural language processing that is used to identify the main themes or topics in a large collection of documents. It does this by analyzing the words and phrases used in the documents and grouping them into clusters based on their co-occurrence patterns [55]. There are several algorithms for performing topic modeling, including Latent Dirichlet Allocation (LDA), Structural Topic Modeling (STM), and the BERTopic model.

The BERTopic model is a variation of the LDA model that uses a combination of word embeddings and a bidirectional encoder representation from transformers (BERT) to model the co-occurrence patterns of words in documents. BERTopic has been shown to outperform other topic modeling algorithms in certain cases, particularly when working with short texts such as social media posts or news articles [55].

Topic modeling can be useful for various tasks, such as information retrieval, document classification, and content recommendation. It can help to organize and summarize large amounts of text data, and can be used to discover hidden patterns and trends in the data [57].

6. Related work

Improving data quality in government organizations is a critical issue that has received attention from various fields, including computer science, information systems, and public administration. In this section, we review several relevant studies that have addressed this issue and identified key challenges and strategies for improving data quality in government organizations.

Holzer, König, and Weinhardt (2003) examined the challenges and solutions for improving data quality in e-government systems in their paper "Data Quality in E-Government: Challenges and Solutions" [43]. They found that data quality problems in e-government systems can be attributed to a variety of factors, such as complexity, lack of standardization, and difficulty in integrating data from multiple sources. To address these challenges, they proposed several strategies, including the implementation of data governance frameworks and the use of data quality tools.

"Using Natural Language Processing to Assess Data Quality in Clinical Text" by A.J. O'Connor et al. (published in the "Journal of the American Medical Informatics Association" in 2013) [63] describes a study in which the authors employed natural language processing (NLP) techniques to evaluate the quality of clinical text data. The study involved using text classification and information extraction to identify errors and inconsistencies in the data, such as incorrect or missing information. The authors also used NLP to identify patterns that may signal data quality issues, such as a high frequency of certain types of errors. The study found that NLP can be effective in identifying and addressing data quality issues in clinical text.

In the paper "Assessing Data Quality of Open Government Data Using Natural Language Processing" by C. Chen et al. (published in the "Government Information Quarterly" in 2016) [62], the authors explore the use of natural language processing (NLP) techniques for evaluating the quality of open government data. The study involves using NLP techniques such as text classification and information extraction to identify errors and inconsistencies in the data and extract relevant information for quality

assessment. The results of the study suggest that NLP can be effective in improving the quality of open government data.

Wang and Liu (2004) conducted an exploratory study of data quality management in government organizations in their paper "Data Quality Management in Government: An Exploratory Study" [44]. They identified several factors that impact data quality in government organizations, including organizational culture, data management processes, and the use of data quality tools. They also proposed several strategies for improving data quality, such as establishing data governance frameworks, implementing data quality management processes, and training staff on data quality best practices.

7. Experimental design

7.1 General methodology

The goal of this research is to use KNIME to gather specific data from the Greek National Catalogue of Services, evaluate its data quality using data quality metrics, and then apply clustering and topic modeling techniques. The outcomes could be used for future tasks to check for duplicate data that has been written in different ways.

KNIME (Konstanz Information Miner) is an open source data analytics platform that allows users to create and share data processing workflows. It provides a visual interface for creating data pipelines, integrating various data sources and processing tools, and analyzing data using machine learning and statistical techniques. KNIME is designed to be user-friendly and can be used by data scientists, analysts, and business users to explore, analyze, and visualize data. It supports a wide range of data types and formats, including structured and unstructured data, and can be used for tasks such as data cleansing, transformation, and integration, as well as predictive modeling and data visualization [47].

7.2 Data Used

The "National Registry of Administrative Proceedings" dataset is a collection of data related to processes, including the names of the processes, alternative titles, types of applications and their owners, and descriptions of associated evidence.

The GET Request node was used to retrieve the services dataset from the mitos.gov website using an API key. The data was returned in a JSON format in the node's output. The JSON Path and Ungroup nodes were used to filter the data and extract the IDs and names. The String Manipulation node combined the URL and IDs, which were then used in the GET Request to extract the data from each page, and with the Rule Engine node, data that had a null value as an id was filtered out. The output was used to obtain the processes and processes evidence datasets. The get processes metanode includes a series of nodes that manipulates and filter the JSON body to extract the processes dataset, which has 78 columns. The data quality analysis focuses on columns

where more than 60% of the rows have no missing values. The get processes evidences metanode is similar to the get processes metanode and generates a table with 18 columns. The process evidence description from the get processes evidences was used for clustering and topic modeling tasks.

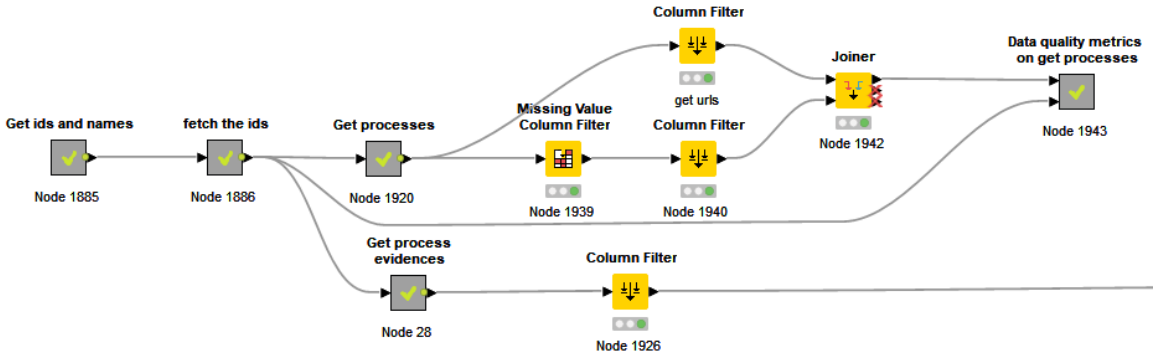

Figure 6.2 (1): Data extraction workflow in KNIME.

8. Experimental results

I have chosen to examine the data quality metrics mentioned by the "official portal of European data," which include completeness, timeliness, accuracy, uniqueness, availability, credibility, processability, and relevance. By examining these metrics, I aim to understand how the data from the "official portal of European data" compares to these standards and identify any areas that may need improvement.

8.1 Completeness

The statistics node gives us information about the number of missing values in each column of the dataset. This helps us understand how complete or incomplete the data is. The analysis showed that there were more than 1,000 missing values in 7 of the columns.

Column	Number of missing values
process_application_related_url	1.933
process_sdg_notes	1.386
process_validity_duration_type	1.338
process_application_note	1.284
process_alternative_titles	1.113
process_remarks	1.102
process_sdg_resource	1.040

We calculated the completeness of the data both for the whole table and for each individual column. The process we used for calculating the completeness of the entire table is shown in the figure 8.1 (1). This workflow includes the use of a Math Formula node, which applies the completeness formula described in the section 3.4.1.

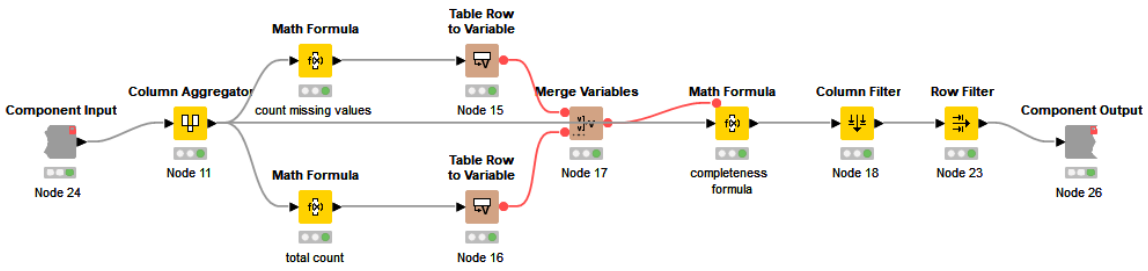

Figure 8.1 (1): Completeness workflow.

The results indicated that the overall completeness of the table is 81.72%. To calculate the completeness of each column, we used the same workflow for each column and then combined the results in a single table. This resulting table was then used to create a histogram.

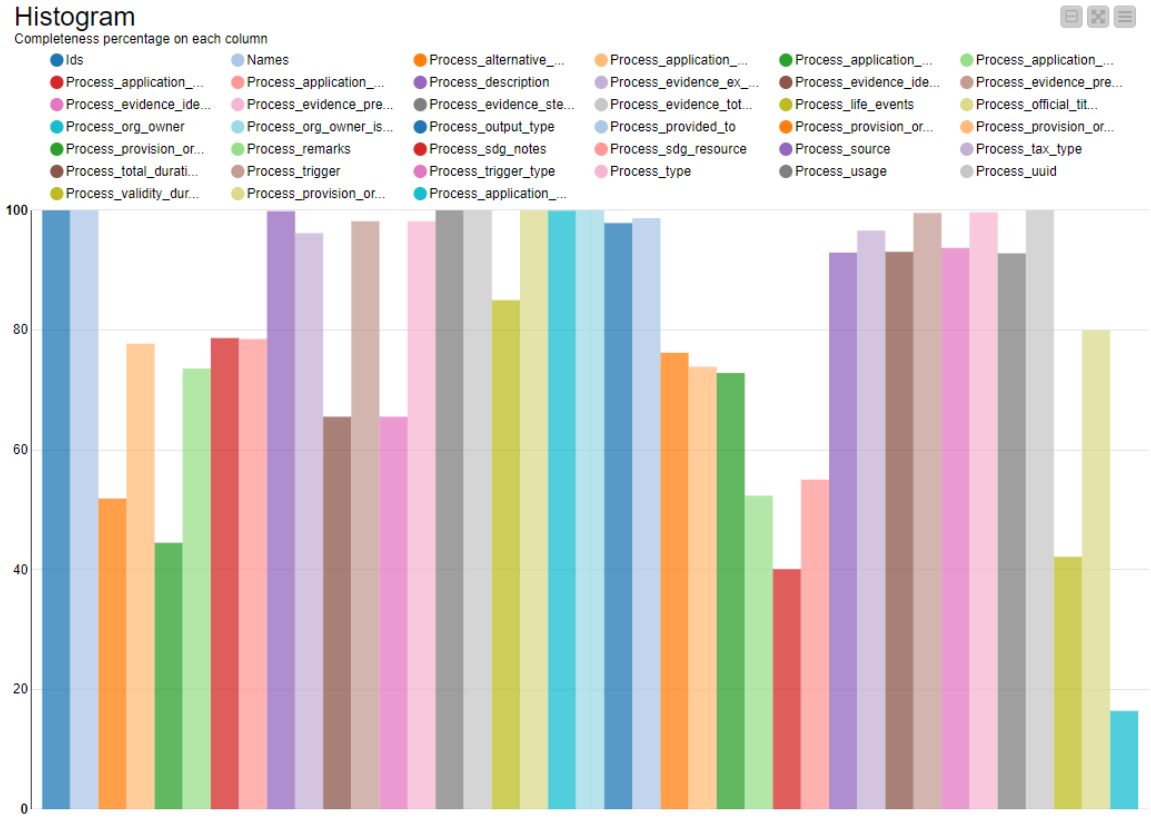

Figure 8.1 (2): Histogram on completeness percentage per column.

The histogram confirms our expectation that the columns with the lowest completeness (below 60%) are the same 7 columns that the statistics node identified as having the most missing values.

There are a large number of missing values in some columns of our dataset, which could indicate low completeness in those columns. However, the overall completeness of the dataset is relatively high at 81.72%. The missing values may have a larger impact on future analyses because they are concentrated in specific columns or rows. To address this issue, it might be best to either remove those columns or supplement them with additional metadata.

8.2 Timeliness

To assess the timeliness of this data quality metric, it was necessary to track updates to the data. We evaluated timeliness in two ways: the frequency of updates relative to previous updates, and the frequency of updates relative to the date that the data was extracted. The first measurement required access to previous updates, so we needed a place to store the updates. Therefore, we used the Microsoft Authentication and SharePoint Online Connector to save this information on SharePoint. In addition to saving the two frequencies, we also made sure to include the date of extraction, computation, and saving in the name of each saved file to avoid confusion.

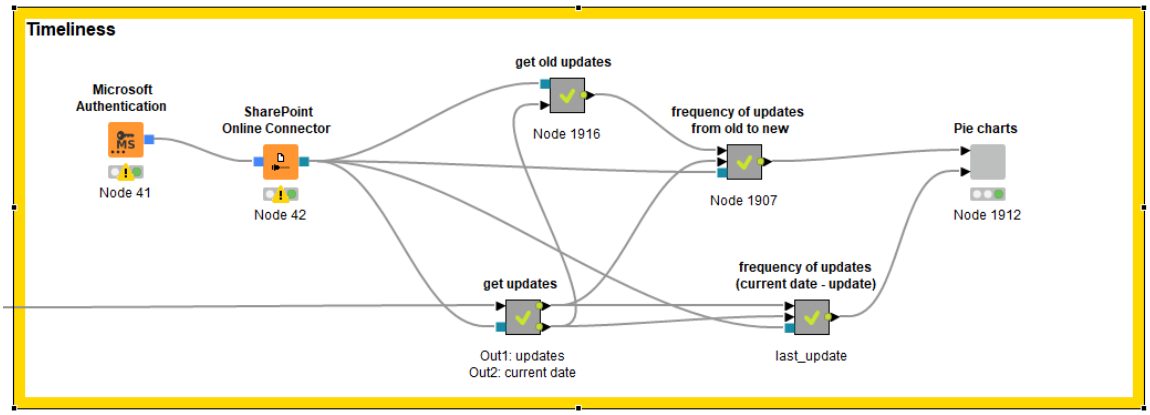

Figure 8.2 (1): Timeliness workflow.

The get updates metanode has two outputs. The first output contains the updates per id, and the second output has the date that the updates were extracted. It includes a series of nodes that manipulate JSON data to extract the updates, as well as nodes like Cell Splitter, Column Filter, and Column Rename, which format the updates in a consistent manner using the YYYY-MM-DD format. To retrieve the directory path, the List Files/Folders node is connected to the SharePoint Online Connector. To create an XLSX file with a name based on the current date, the Create Date&Time Range node is used, and its output is converted to a variable. This variable is then used with the Java Edit Variable (sample) node to create the file name in the format "Updates_" + $${Stime_now}$$ + ".xlsx". The directory path and file name are combined and converted to a variable that is connected to an Excel Writer node. This process enables the updates to be saved to SharePoint under a unique file name. The get old updates

metanode retrieves all the previously saved updates files from SharePoint and filters out the file that was already saved by the get updates metanode.

The frequency of updates for a metanode is calculated by determining the number of days between the current date and the date the update was made. This is done by inputting the dates of data extraction and updates into the metanode and using a Date&Time Difference node to find the difference between the two dates. The Rule Engine node is then used to categorize the updates based on the amount of time that has passed, as shown in the following table.

Time difference (td) or Date	Label
0	Updated today
0 <td <=31	1 month
31 <td <=60	2 months
60 <td <=90	3 months
90 <td <=120	4 months
120 <td <=150	5 months
150 <td <=180	6 months
180 <td <=210	7 months
210 <td <=240	8 months
240 <td <=270	9 months
270 <td <=300	10 months
300 <td <=330	11 months
330 <td <=361	1 year
td > 361	More than 1 year
Date of update is missing	N/A

The same method that was used in updates is used here to save the frequencies on SharePoint under the name "last_update_" + $${Stime_now}$$ + ".xlsx".

The frequency of updates from old to new metanode outputs is calculated by finding the time difference between the current extracted date and the previously extracted date of updates. The process used to do this is the same as before, with the difference being in the categorization values on the Rule Engine node. The results are saved under the name "freq_of_update_" + $${Stime_now}$$ + ".xlsx" and the categorization values are shown on the following table.

Time difference (td) or Date	Label
0	Not updated
0 <td <=31	1 month
31 <td <=60	2 months
60 <td <=90	3 months
90 <td <=120	4 months
120 <td <=150	5 months
150 <td <=180	6 months
180 <td <=210	7 months
210 <td <=240	8 months
240 <td <=270	9 months
270 <td <=300	10 months
300 <td <=330	11 months
330 <td <=361	1 year
td > 361	More than 1 year
Old + new updates are missing	N/A
Old updates are missing	New entry
New updates are missing	Deleted

The pie charts metanode displays the visualization of the results. This is done by using the Statistics node to count the number of occurrences of each categorization.

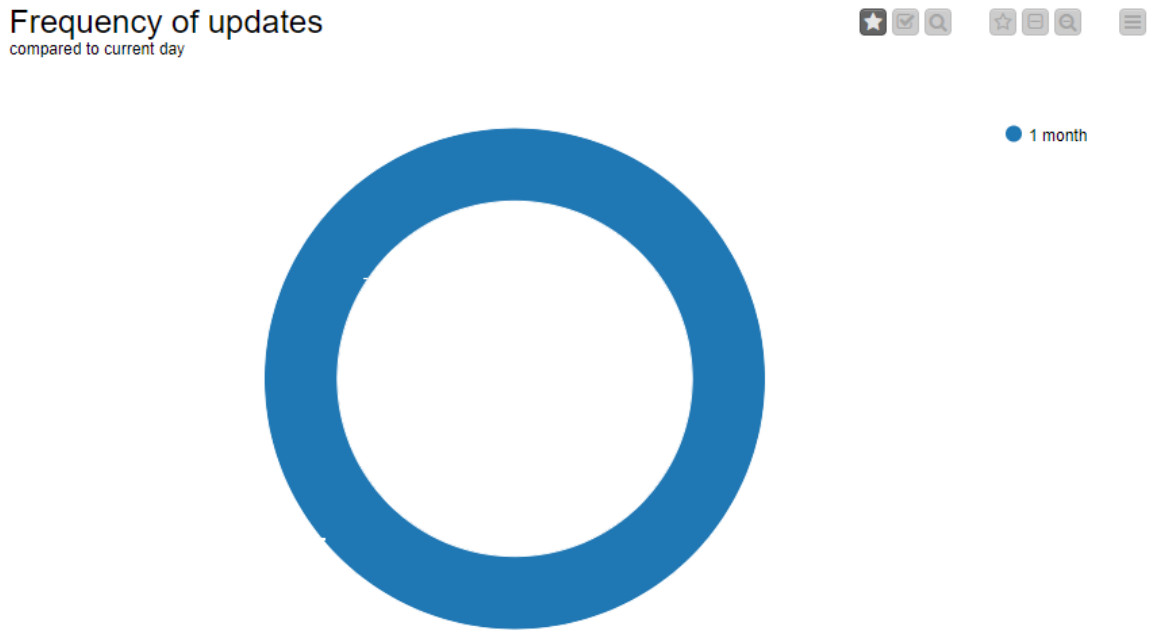

Figure 8.2 (2): Pie chart with the frequency of updates based on the date that they were extracted.

It appears that all of the data in our dataset has been updated within the past month. This suggests that the timeliness of the data is relatively high, as all of the data is relatively current.

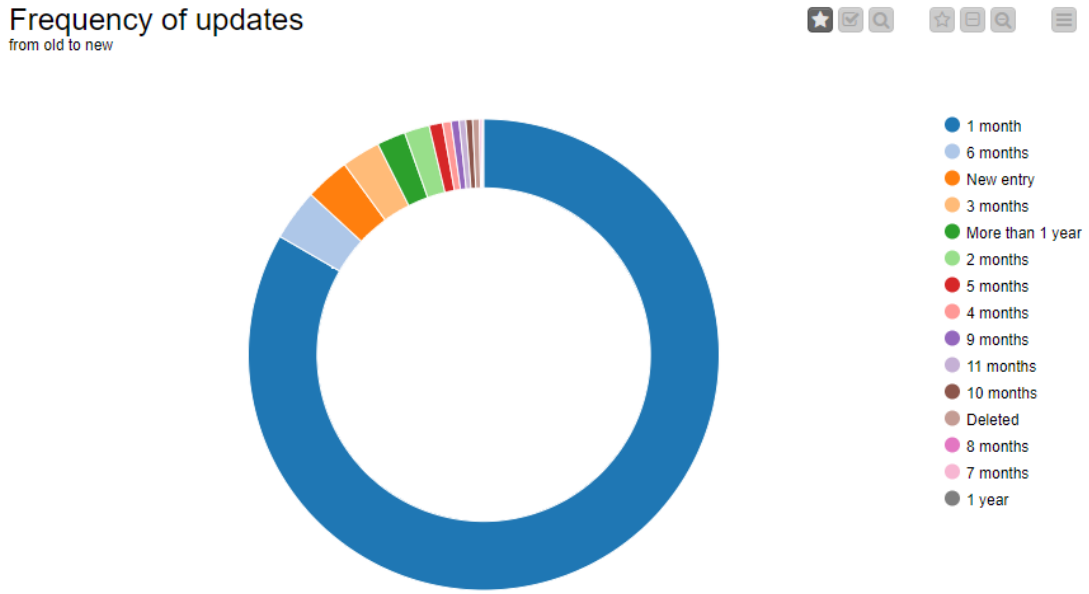

Figure 8.2 (3): Pie chart with the frequency of updates from old to new.

It appears that the majority of our data (83%) is being updated regularly, on a monthly basis. This is a positive indication of the timeliness of our data. However, a small portion of our data (2%) has not been updated in over a year, which could potentially impact the usefulness and relevance of that data. Additionally, another 8% of the data has not been updated in over 6 months, which could also affect the timeliness of that data.

Considering previous updates in addition to current ones is important for evaluating the timeliness of our data. Even though all of the data in our dataset has been updated recently, it is still important to consider data that has not been updated as frequently to ensure its relevance and accuracy. Regularly reviewing and updating our data helps to ensure that it is current and accurate, allowing us to make informed decisions based on the most current information available.

8.3 Accuracy

To determine the accuracy of each column in the dataset, the number of occurrences of each term in the column was calculated using the Statistics node. We selected the columns that seemed to have a sufficient number of occurrences for a specific value. There were a total of 6 chosen columns. Using the Rule Engine node, we classified values according to the following rules: if the value is fixed, it is labeled as "Accurate," if it is missing it is classified as "Null," and if it does not fit into either of these categories it is labeled as "Not accurate or Other."

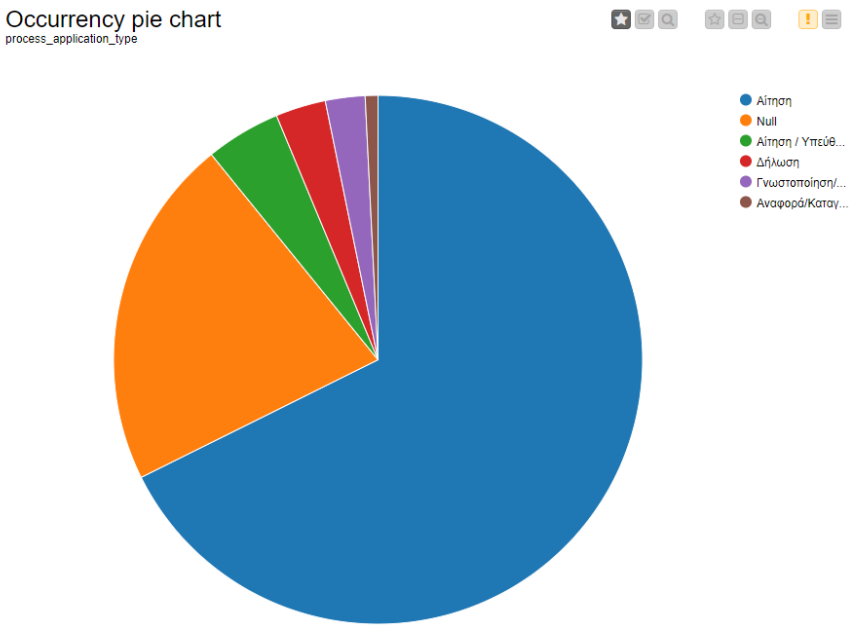

Figure 8.2 (4): Occurrency pie cart on process_application_type.

For example, in the process_application_type column, the values that occurred most frequently and were the only values present were: "Αίτηση," "Δήλωση," "Γνωστοποίηση/Αναγγελία/Ανακοίνωση," "Αναφορά/Καταγγελία," "Αίτηση/Υπεύθυνη Δήλωση.", and "Null". Therefore, if the value is Null, it is categorized as "Null". If the value appears in the list of occurrences and is not Null, it is categorized as "Accurate". All other values are categorized as "Not accurate or Other". If a value is not accurate, it may be a typo error. On the other hand, if a value is classified as "Other," it indicates that it is either a new entry or a value that is not commonly seen.

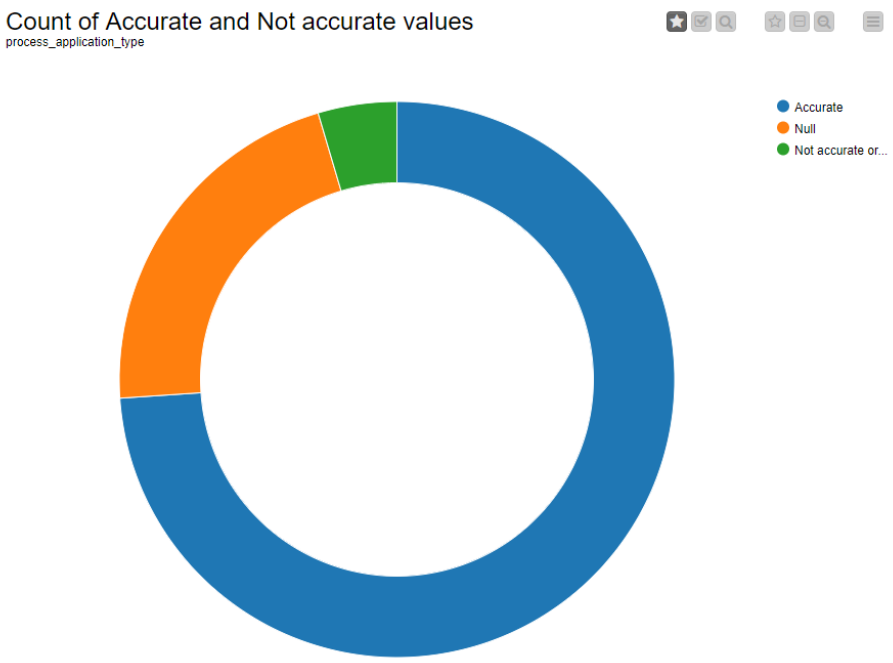

Figure 8.2 (5): Count pie chart of Accurate and Not accurate values.

In the "process_application_type" column, 73.8% of the values are classified as "Accurate," but there are also a significant number of "Null" (21.2%) and "Not accurate or Other" (5%) values present. These null and "not accurate or other" values could potentially impact the accuracy of the data in that column. The "process_org_owner_is_private" column has 100% "Accurate" values, indicating a high level of accuracy for that column. The "process_sdg_resource" column has a relatively high level of accuracy with 55.1% "Accurate" values, but also has a significant number of "Null" values (44.9%). The "process_total_duration_steps_type" column has a high level of accuracy with 93.3% "Accurate" values, but also has a relatively small number of "Null" values (6.7%). The "process_type" column has a very high level of accuracy with 99.7% "Accurate" values and only a small number of "Null" values (0.3%). The "process_validity_duration_type" column has a relatively low level of accuracy with 56.8% "Null" values and 43.1% "Accurate" values, and a small number of "Not accurate or other" values (0.1%). Overall, it appears that the accuracy of the 6 columns varies, with some columns having relatively high levels of accuracy and others having lower levels. It may be worthwhile to consider why there are null and "Not accurate or other"

values present in some of the columns and whether there are steps someone can take to improve the accuracy of the data in those columns.

8.4 Uniqueness

The Uniqueness workflow consists of a Statistics node that outputs the occurrences of each value. With the Missing Value node, the null values are given a fixed value ("Null"). After inspecting the results, we noticed that for the columns ids, names, and process_uuid, the lower and upper bounds of their count are both 1, meaning that they consist of unique values.

Columns: 6	Column Type	Column Index	Color Handler	Size Handler	Shape Han…	Filter Handler	Lower Bound	Upper Bound
ids	String	0					?	?
Count (ids)	Number (int…	1					1	1
names	String	2					?	?
Count (names)	Number (int…	3					1	1
process_uuid	String	4					?	?
Count (process_uuid)	Number (int…	5					1	1

Figure 8.4 (1): Lower and upper bounds of columns that consist of unique values.

The results for the columns "process_alternative_titles" and "process_official_title" are notable because, aside from the first 10 rows, all of the other values are unique. Additionally, with the exception of the Null values count in "process_alternative_titles", the count for the first 10 rows ranges between 2 and 4. This suggests that further analysis may be necessary in the future to determine if these values are linked to duplicate records. If they are, then these two columns will also contain unique values.

Table "default" - Rows: 2313 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	process_alternative_titles	Count (process_alternative_titles)	process_official_title	Count (process_official_title)
Row0	Null	1113	Έκδοση άδειας λειτουργί…	2
Row1	Μετατροπή άδειας οδήγησης, …	4	Ανανέωση Άδειας Λειτου…	2
Row2	Ανάκληση άδειας υδατοδρομίου	3	Έλεγχος και έγκριση τω…	2
Row3	Ηλεκτρονικό Μητρώο τουριστ…	3	Ειδικές Περιπτώσεις κτήσ…	2
Row4	myPhoto	2	Μεταβολή Γνωστοποίησ…	2
Row5	Άδεια λειτουργίας Κ.Δ.Β.Μ., …	2	Υποβολή τροποποιητικής…	2
Row6	Μεταβολή όρων και προϋποθ…	2	ERASMUS+ στους τομείς…	1
Row7	Έντυπο Ε9	2	myPhoto για πολίτες	1
Row8	Τροποποιητική δήλωση Ε9	2	myPhoto για φωτογράφ…	1
Row9	Διαχείριση και παρακολούθησ…	1	Έκδοση νέας άδειας υδα…	1
Row10	Αλλαγή νομικής μορφής της …	1	Έκδοση νέας άδειας υδ…	1

Figure 8.4 (2): Occurrences count for the columns process_alternative_titles and process_official_title.

8.5 Availability

For the availability dimension, we decided to verify the functionality of the links in the process_application_related_url column. To do this, the links were removed of duplicates and saved in a CSV file called mylinks.txt in the KNIME workspace environment. They were then manipulated using Python in the Python Source node. The links were labeled as 0 if they were not functioning and as 1 if they were functioning. These labels were then stored in a csv file called "working_links.csv" and outputted from the node.

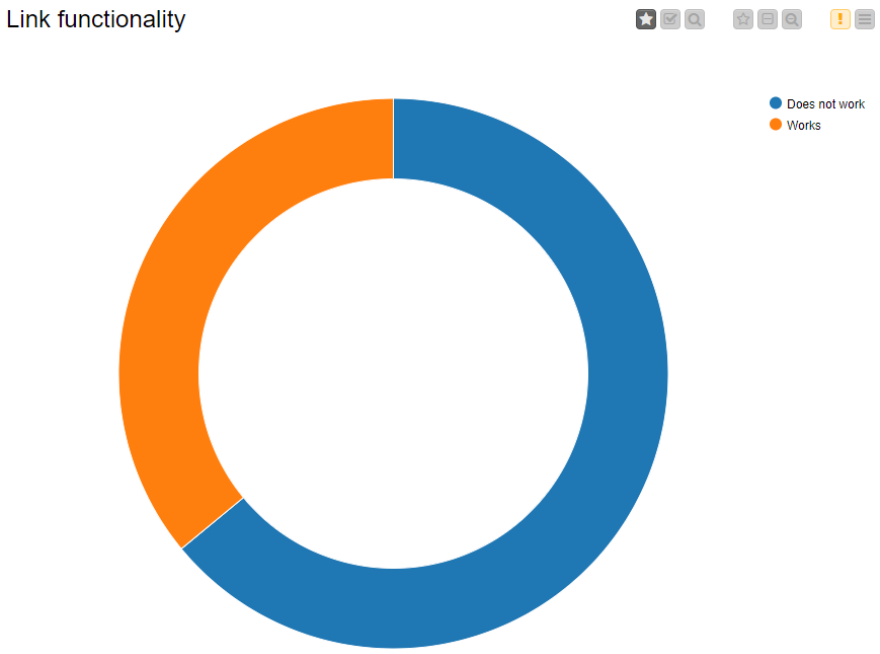

Figure 8.5 (1): Link functionality pie chart.

However, upon further examination, we discovered that the code was not entirely accurate as it incorrectly classified some working links as non-working. This was often due to the links not having the proper format, such as "http://link.gr", even though they were actually functioning. An example of this case of misclassification is the link "gov.gr". Nevertheless, there were also instances where the code correctly identified non-working links, which suggests that it has the potential to be improved upon. Overall, it seems that the availability of our data may be somewhat limited due to the inaccuracies

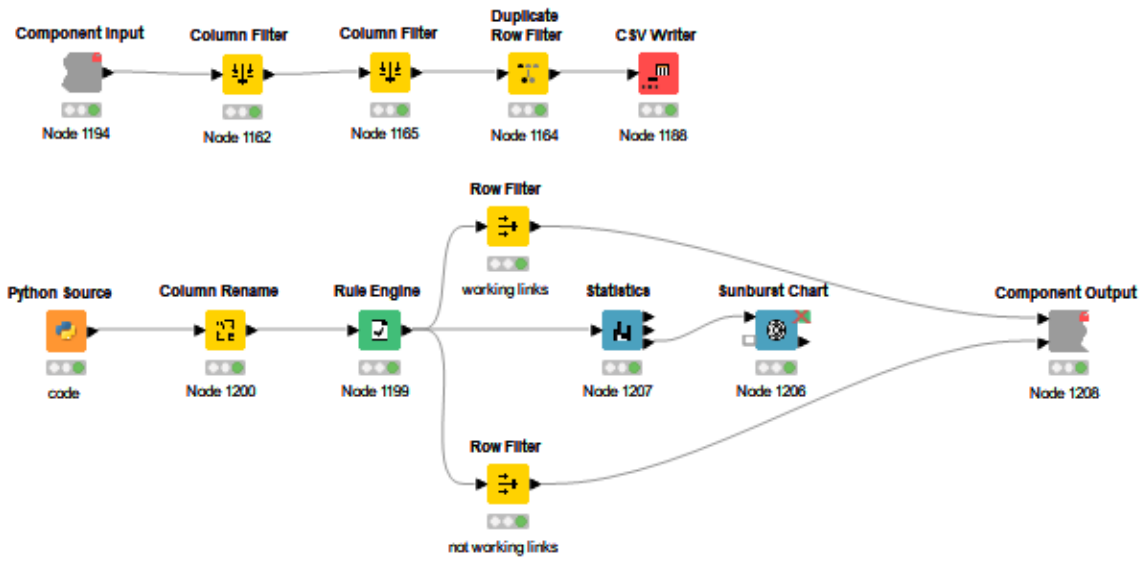

Figure 8.5 (2): Workflow for the availability dimension.

8.6 Credibility

As stated in section 3.4.6, data from government sources is generally considered to be highly credible. Therefore, since the data in this case is sourced from an official government site, we can conclude that it is entirely reliable.

8.7 Processability

Since the data is presented in a JSON format that includes text and updates in ISO 8601 format, it can be easily read and processed by machines. Overall, it seems that our data are in a structured and organized format, which should make them relatively easy to process and analyze. The inclusion of text and numerical values, as well as dates in ISO 8601 format, suggests that our data may be suitable for a wide range of analyses, including statistical and time-series analyses.

8.8 Relevance

The importance of the relevance quality dimension is often overlooked when it comes to government data [30]. As mentioned previously in section 3.4.8, the relevance of data can vary depending on the intended application, and our dataset includes a large number of different dimensions. To increase the relevance of the data and provide more context and meaning to the content, we need to add metadata to the dataset. This will also

allow us to fill in any missing values. Overall, enhancing the dataset with metadata will help to improve the relevance dimension and make the data more useful.

8.9 Text similarity, clustering and topic modeling

The main objective of this task is to utilize data from the "process_evidence_description" column that has been saved previously to form clusters of data with their own centroids, representatives, and topics. After that, the new data being analyzed will be compared to the centroids of each cluster using cosine similarity scores, and will be placed into the cluster with which it has the highest similarity. Finally, the embeddings of the data will be plotted in a scatterplot to display the top 20 clusters. The complete workflow for this task is illustrated in Figure 8.9 (1).

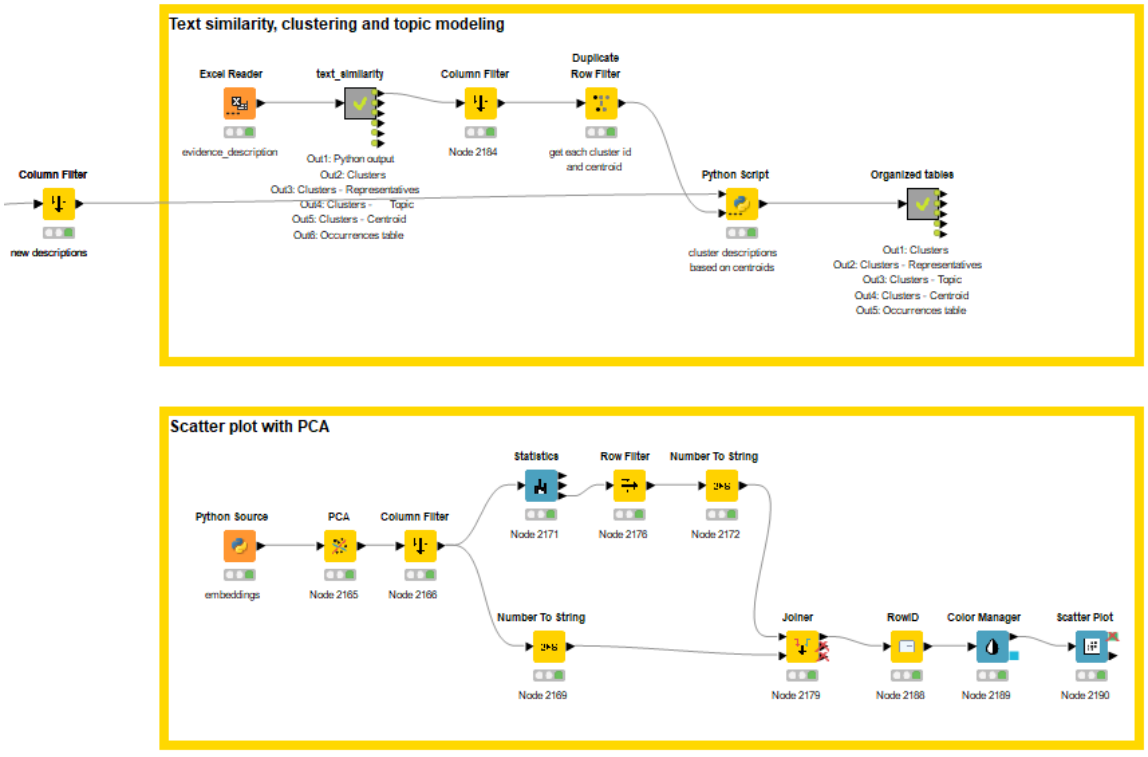

Figure 8.9 (1): Workflow of text similarity, clustering, and topic modeling.

The "text_similarity" metanode uses data from the "process_evidence_description" table as its input. This data is inputted into a Python

Script node and divided into smaller chunks, with each sentence being no longer than 512 characters. As the data will be transformed using document embeddings, there is no need to perform any preprocessing on the data. This is because all parts of a document are important in understanding the general topic of the document, and the document embeddings take this into account. As a result, there is no need to modify the data in any way before it is transformed. The sentence transformer model "paraphrase-multilingual-MiniLM-L12-v2" is then used to convert each row of data into an embedding. The "paraphrase-multilingual-MiniLM-L12-v2" is a sentence transformer model that converts sentences and paragraphs into a dense vector space with 384 dimensions. The embeddings were normalized. This model can be used for tasks like clustering or semantic search [29]. The embeddings were then used to determine the similarity between each sentence by calculating the cosine similarity. The pairs of sentences with the highest similarity scores (at least 94%) were then combined into a new dataframe.

Row ID	S desc1	S desc2	D cosine_score
Row0	Συγκεντρωτική κατάσταση με τα ονόματα των καθηγητώ...	Συγκεντρωτική κατάσταση με τα ονόματα των καθηγη...	0.992
Row1	Αντίγραφο της δήλωσης έναρξης εργασιών ή της δήλωση...	Αντίγραφο της δήλωσης έναρξης εργασιών ή της δήλω...	0.989
Row2	Κατάλογος εγγράφων νομιμοποίησης νομικών προσώπων	Κατάλογος εγγράφων νομιμοποίησης νομικών προσώπ...	0.988
Row3	Βεβαίωση του οικείου Εμπορικού και Βιομηχανικού Επιμελη...	Βεβαίωση του οικείου Εμπορικού και Βιομηχανικού Επιμ...	0.985
Row4	Υπεύθυνη δήλωση 2 Πολιτικών Μηχανικών για τη στατική ...	Υπεύθυνη δήλωση 2 Πολιτικών Μηχανικών για τη στατι...	0.982
Row5	Βεβαίωση Πυροσβεστικής Υπηρεσίας, συνοδευόμενη από σ...	Βεβαίωση Πυροσβεστικής Υπηρεσίας, συνοδευόμενη απ...	0.982
Row6	Βεβαίωση Πυροσβεστικής Υπηρεσίας, συνοδευόμενη από σ...	Βεβαίωση Πυροσβεστικής Υπηρεσίας, συνοδευόμενη απ...	0.982
Row7	Αντίγραφο του καταστατικού (ή οργανισμού λειτουργίας, ...	Αντίγραφο του καταστατικού (ή οργανισμού λειτουργία...	0.981
Row8	Ορισμός και στοιχεία αντικλήτου, εφόσον απαιτείται	Ορισμός και στοιχεία αντικλήτου, εφόσον απαιτείται.	0.968
Row9	Ορισμός και στοιχεία αντικλήτου, εφόσον απαιτείται	Ορισμός και στοιχεία αντικλήτου, εφόσον απαιτείται.	0.968
Row10	Βεβαίωση από Πολεοδομία ότι επιτρέπεται στο συγκεκριμέ...	Βεβαίωση από Πολεοδομία ότι επιτρέπεται στο συγκεκρι...	0.964

Figure 8.9 (2): Results of sentences that have a cosine similarity score of equal to or over 94%.

In order to speed up the computation, we decided to perform the clustering on the data that had high cosine similarity scores and was included in the previous dataframe. For this reason, the desc1 and desc2 columns (shown in Figure 8.9 (2)) were combined into a list. The previously mentioned list was used as input for agglomerative clustering, a clustering method that does not require the specification of the number of clusters. In this case, the number of clusters was not defined. In addition to calculating the full tree, a linkage distance threshold was used to determine when clusters should not be merged for distances of 0.7 or greater. The final result was 1.122 clusters. Figure 8.9 (3) shows a

table of these clusters, where each column represents a single cluster and the column's name is the cluster's identification number.

Table "default" - Rows: 116 | Spec - Columns: 1122 | Properties | Flow Variables

Row ID	S 0	S 1	S 2	S 3
Row0	Τεχνική Έκθεση	Άριθμος Φορολογικού Μητρώου (ΑΦΜ)	Πλήρης αρχιτεκτονική μελέτη.	Πιστοποιητικά ADR οδηγών.
Row1	Τεχνική έκθεση χωροθέτησης	Αριθμός Φορολογικού Μητρώου	Σειρά αρχιτεκτονικών σχεδίων.	Πιστοποιητικό ADR οδηγών.
Row2	Τεχνική έκθεση	Άριθμος Φορολογικού Μητρώου (ΑΦΜ)	Πλήρης αρχιτεκτονική μελέτη	Πιστοποιητικά ADR οδηγών.
Row3	Τεχνική έκθεση χωροθέτησης.	Αριθμός Φορολογικού Μητρώου (ΑΦΜ) αιτ...	Αρχιτεκτονικά σχέδια οικοδομής	Πιστοποιητικά ADR οδηγών.
Row4	?	Αριθμός Φορολογικού Μητρώου (ΑΦΜ) αιτ...	?	Πιστοποιητικά ADR των οδηγ...

Figure 8.9 (3): Table of clusters.

After careful examination of the clusters, we found that some of them were very similar and should be merged into a single cluster. To do this, we need to calculate the centroid for each cluster and then determine the cosine similarity between the centroids. This will allow us to identify which clusters are similar enough to be combined.

To calculate the centroids for each cluster, we first needed to convert the descriptions in each cluster into normalized embeddings using the same transformers model that was used earlier in the process. These embeddings are numerical representations of the words in each description, and they are represented as tensors with 384 values. We then took the average of all of the embeddings in a cluster to form the cluster's centroid. The final centroid for each cluster is a tensor with 384 values, and it is given an identification number (the centroid id) that matches the identification number of the cluster it originated from (the cluster id).

Table "default" - Rows: 384 | Spec - Columns: 1122 | Properties | Flow Variables

Row ID	D 0	D 1	D 2	D 3
Row0	-0.024	-0.014	0.053	0.041
Row1	0.043	0.044	0.051	0.039
Row2	-0.014	-0.082	-0.005	0.011
Row3	-0.038	0.01	0.03	0.018
Row4	0.024	-0.021	-0.035	0.018
Row5	-0.003	0.036	-0.045	-0.011
Row6	-0.028	0.034	-0.127	0.064
Row7	0.025	0.088	0.007	0.037
Row8	0.012	0.015	-0.041	-0.022
Row9	0.031	0.004	0.022	-0.088
Row10	0.017	0.051	-0.062	0.043

Figure 8.9 (4): Table of centroids.

To identify similar centroids, the cosine similarity was calculated for each pair of centroids. Only those pairs that had a cosine similarity score of 90% or higher were retained, while all other pairs were discarded. The IDs for the remaining pairs of centroids were then grouped together. For instance, if the pairs [4, 102] and [27, 4] were found, this would indicate that 4 has a high similarity with both 102 and 27, so these IDs would be grouped together in a list [4, 102, 27], while the other pairs would be discarded. After grouping the IDs into lists, the IDs within each list were sorted in ascending order. Since the cluster and centroid IDs were the same, for each list of similar IDs, every ID except for the first one was selected for its cluster to be merged with the first ID's cluster. These IDs were then removed from the clusters and centroids tables. For example, if we have the list [4, 27, 104], the clusters with IDs 27 and 104 will be merged into the cluster with ID 4, and the centroids and clusters with IDs 27 and 104 will be removed. Finally, the column names for the clusters and centroids tables (which are currently the IDs) will be renamed with numbers starting from 0 and increasing by a step of 1. This renaming is necessary to avoid confusion and make it easier to manipulate the data in the future. By following these steps, we were able to group together similar clusters and ultimately end up with 524 clusters.

Table "default" - Rows: 823 | Spec - Columns: 524 | Properties | Flow Variables

Row ID	S 0	S 1	S 2	S 3
Split Value 1	Τεχνική Έκθεση	Τα δικαιολογητικά δεν αναφέρονται στην παρούσα κ...	Υπεύθυνη δήλωση του ν. 1599/1986	Για τον Διευθυντή Σπουδών: Φωτοαντίγρ...
Split Value 2	Τεχνική έκθεση χωροθέτησης	Τα δικαιολογητικά δεν αναφέρονται στην παρούσα κ...	Υπεύθυνη δήλωση του ν. 1599/1986	Διευθυντής σπουδών: Φωτοαντίγραφο τη...
Split Value 3	Τεχνική έκθεση	Τα δικαιολογητικά δεν αναφέρονται στην παρούσα κ...	Υπεύθυνη δήλωση του ν. 1599/1986	Διευθυντής σπουδών: Φωτοαντίγραφο τη...
Split Value 4	Τεχνική έκθεση χωροθέτησης.	Τα δικαιολογητικά δεν αναφέρονται στην παρούσα κ...	Υπεύθυνη δήλωση του ν. 1599/1986	Για τον Νόμιμο εκπρόσωπο και τον διευθυ...
Split Value 5	?	Τα δικαιολογητικά δεν αναφέρονται στην παρούσα κ...	Υπεύθυνη Δήλωση ν. 1599/1986.	Φωτοαντίγραφο της αστυνομικής ταυτότ...

Figure 8.9 (5): Table of grouped clusters.

In order to select a description to represent each cluster, we calculated the embeddings of each sentence in the cluster using the same sentence transformer model as before. We then calculated the cosine similarity between each of these embeddings and the centroid of the cluster. From all of the embeddings within the cluster, the description associated with the embedding that had the highest score was chosen to be the representative for the cluster. The table shown in Figure 8.9 (5) presents the selected representative descriptions for each cluster, organized by their respective cluster or centroid ID.

Table "default" - Rows: 1 | Spec - Columns: 524 | Properties | Flow Variables

Row ID	S 0	S 1	S 2	S 3
Split Value 1	Τεχνική έκθεση	Τα δικαιολογητικά δεν αναφέρονται στην παρού...	Υπεύθυνη δήλωση του ν. 1599/1986	Για τον Διευθυντή Σπουδών: Φωτοαντίγρα...

Figure 8.9 (6): Table of representatives.

In order to represent each cluster with a specific topic, we applied topic modeling using the BERTopic multilingual model to the clusters of sentences. The model requires a list of at least 10 values as input, so if a cluster contained fewer than 10 descriptions, we

multiplied the list to generate additional self-similar texts. The chosen topic for each cluster was determined by the topic (labeled as "topic 0") that the majority of sentences in the cluster belonged to. If a cluster had no identifiable topics, it was assigned the topic labeled as "-1", which contained all sentences that could not be assigned to any other topic. The table displayed in Figure 8.9 (6) provides a summary of the topics that have been identified for each cluster, with the cluster or centroid ID serving as the reference for each topic. This allows the reader to see which topics have been associated with each cluster and how these topics are related to the overall organization of the data.

Table "default" - Rows: 1 | Spec - Columns: 524 | Properties | Flow Variables

Row ID	S 0	S 1	S 2	S 3
Split Value 1	0_τεχνική_έκθεση__	-1_και_στη_προσδιορίζονται_περίπτωση	0_του_άρθρου_δήλωση_υπεύθυνη	-1_σπουδών_και_διευθυντή_τις

Figure 8.9 (7): Table of topics.

Finally, a dataframe was created and outputted from the Python Script node. This dataframe includes the following columns: 'Description', 'Cluster', 'Centroid', 'Representative', and 'Topic'. This dataframe provides a summary of the descriptions and their associations with various clusters, centroids, representatives, and topics. Moreover, all the tables shown in the above figures have been saved in Excel format.

Table "default" - Rows: 5470 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Description	S Cluster	[...] Centroid	[...] Represen...	[...] Topic
Row0	Τεχνική Έκθεση	0	[-0.02418128797216923,0.0426...	[Τεχνική έκθε...	[0_τεχνική_έκθεση...
Row1	Τεχνική έκθεσ...	0	[-0.02418128797216923,0.0426...	[Τεχνική έκθε...	[0_τεχνική_έκθεση...
Row2	Τεχνική έκθεση	0	[-0.02418128797216923,0.0426...	[Τεχνική έκθε...	[0_τεχνική_έκθεση...
Row3	Τεχνική έκθεσ...	0	[-0.02418128797216923,0.0426...	[Τεχνική έκθε...	[0_τεχνική_έκθεση...
Row4	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row5	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row6	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row7	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row8	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row9	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...
Row10	Τα δικαιολογη...	1	[0.05806858119155679,8.6388...	[Τα δικαιολογ...	[-1_και_στη_προσδι...

Figure 8.9 (8): Final table.

Now that we have all of the essential data (clusters, centroids, representatives, and topics), the goal is to assign each new description to a specific cluster. This process involves calculating the embeddings of the new descriptions using the same sentence

transformer model as before, and then comparing the cosine similarity between these embeddings and the centroids for each cluster. The cluster with the highest score will be chosen as the cluster for the new description. This allows us to organize the new descriptions in a meaningful way, by grouping them with other descriptions that are similar in content and meaning. After extracting the "centroid" and "cluster" columns from the final table with the Column Filter, we can use the Duplicate Row Filter to remove any duplicate rows, resulting in a table with unique cluster and centroid combinations.

Table "default" - Rows: 524 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	S Cluster	[...] Centroid
Row0	0	[-0.02418128797216923,0.042621846310794353,-0.014024036703631282,...
Row4	1	[0.05806858119155679,8.638874028942414E-4,0.030057901915694987,...]
Row26	2	[-0.07624841887842525,0.11467689885334535,-0.03382737553593787,...]
Row105	3	[0.026533739970001087,0.10893842490280375,-0.026303779552964604,...]
Row122	4	[-0.011373624870819705,0.06667715470705714,-0.049945121364934106,...
Row136	5	[-0.03982252143323421,0.046335508860647676,-0.024881330877542497,...
Row156	6	[-0.033558797906152904,0.05313612846657634,-0.0793405519798398,...]
Row270	7	[-0.005492503181673014,0.0943411519894233,-0.02724180255944912,...]
Row346	8	[-0.013044431785795283,0.07324304019517087,-0.06112890912497297,...]
Row1169	9	[-0.04928220622241497,0.09529111534357071,-0.06889781169593334,...]
Row1173	10	[0.07885527921219666,0.013946341273064414,-0.04373891403277715,...]
Row1185	11	[0.023884073458611967,0.06362291350960732,-0.019997746776789428,...]
Row1195	12	[0.041394757639084546,0.038988548064870496,0.010935844121766942,...]

Figure 8.9 (9): Table of unique centroids and cluster IDs.

The table of unique centroids and cluster IDs, along with the new descriptions, are inserted into a Python Script node. Within the node, the Excel files containing grouped clusters, representatives, and topics are also uploaded. The new descriptions can then be processed by slicing them and limiting them to 512 characters in length in order to be used with the sentence transformer model. After formatting the new descriptions and the unique centroids and cluster IDs, the centroids can be placed into a list and converted into a tensor, which is the required format for the cosine similarity model. The new descriptions can then be transformed into embeddings using the transformers model, and the cosine similarity score between the embeddings and the centroids can be calculated. The maximum score and the corresponding centroid can be chosen, and the centroid ID can be used to determine the appropriate cluster, representative, and topic for the new

description. This process allows us to accurately place the new descriptions into the appropriate clusters and associate them with the relevant centroids, representatives, and topics. The resulting table that outputs the Python Script node is shown in the figure below.

Table "default" - Rows: 10416 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Description	C...	Representative	Centroid	Topic
Row0	Έντυπο νομικής ον...	258	[Κατάθεση γενικού σχεδιαγρά...	[0.0262405816465616...	[-1_της_μονάδας_κάτοψη_σχεδιαγράμματ...
Row1	Δελτίο Τραπεζικών ...	63	[Έντυπο υπαγωγής στις Πρότ...	[-0.0035869384417310...	[0_2036_περιβαλλοντικές_με_κοινή]
Row2	Επίσημα νομιμοποιη...	510	[Αντίγραφο πτυχίου, διπλώμα...	[-0.0210192622616887...	[0_αντίγραφο_από_το_νόμιμα]
Row3	Βεβαίωση έναρξης ή...	184	[Υπεύθυνη Δήλωση του φορέ...	[-0.0539864450693130...	[-1_του_της_και_λειτουργικές]
Row4	Νόμιμα συνταγμένο...	489	[Πρόσθετα Δικαιολογητικά Υπ...	[-0.0276302006095647...	[0_όπως_υποψηφίων_πρόσθετα_κατηγορι...
Row5	Κωδικοί TAXISnet	103	[Φορολογική Ενημερότητα γι...	[0.0166960331526669...	[-1_κεντρικής_πλην_διοίκησης_για]
Row6	Δεύτερος παράγοντ...	94	[Φωτοαντίγραφο τίτλου σπου...	[-0.0016667331353976...	[-1_υπεύθυνου_του_φωτοαντίγραφο_επισ...
Row7	Υπεύθυνη δήλωση τ...	2	[Υπεύθυνη δήλωση ανάληψη...	[-0.0762484188784252...	[-1_λειτουργίας_υπεύθυνη_καθηκόντων_...
Row8	Υπεύθυνη δήλωση τ...	2	[Υπεύθυνη δήλωση ανάληψη...	[-0.0762484188784252...	[-1_λειτουργίας_υπεύθυνη_καθηκόντων_...
Row9	Κατάλογος εγγράφ...	323	[Δύο φωτογραφίες.]	[-0.0206584899375836...	[0_πρόσφατες_δυο_φωτογραφίες_δύο]
Row10	Αντίγραφο της δήλ...	111	[Ασφαλιστική Ενημερότητα]	[-0.0212715042755007...	[-1_ενημερότητα_ασφαλιστική_πιστοποιητι...

Figure 8.9 (10): Final table of new descriptions.

To create a scatter plot of the top 20 clusters, we need to have the sentence embeddings in a table with each row representing an embedding and the columns representing the values of each embedding. The first column will contain the cluster id so that we know which cluster the embedding belongs to. This process was carried out using a Python Source node in which the grouped clusters were uploaded and the sentence transformers model was used to convert each sentence into an embedding. The final result of the process described above is shown in the figure that follows.

Table "default" - Rows: 10442 | Spec - Columns: 385 | Properties | Flow Variables

Row ID	cluster_id	0	1	2
Row0	0	-0.147	0.157	-0.121
Row1	0	-0.04	0.097	-0.021
Row2	0	-0.153	0.19	-0.042
Row3	0	-0	0.101	-0.01
Row4	1	-0.053	0.155	-0.289
Row5	1	-0.057	0.207	-0.331

Figure 8.9 (11): Table of embeddings.

The table of embeddings described above is input into the PCA node, which uses all of the columns except the cluster_id column to perform dimensionality reduction and reduce the data to 2 dimensions.

Table "default" - Rows: 10442 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	I cluster_id	D PCA dimension 0	D PCA dimension 1
Row0	0	0.853	1.718
Row1	0	0.896	1.341
Row2	0	0.796	1.805
Row3	0	0.922	1.288
Row4	1	0.639	0.479
Row5	1	0.637	0.548
Row6	1	0.639	0.479
Row7	1	0.768	0.222
Row8	1	0.768	0.222
Row9	1	0.639	0.479
Row10	1	0.649	0.452

Figure 8.9 (12): Dimensionality reduction table.

Using the Statistics node, we can see how many elements are in each cluster, which allows us to choose the top 20 clusters. The cluster_id column is converted to a string so that it can be used in the Color Manager node to assign a unique color to each cluster. This helps to clearly distinguish the different clusters in the scatter plot.

Table "default" - Rows: 20 | Spec - Columns: 3 | Properties

Row ID	S cluster_id	I Count (cluster_id)
Row0	14	2314
Row1	65	434
Row2	44	420
Row3	5	330
Row4	12	269
Row5	6	204
Row6	47	190
Row7	59	182
Row8	256	162
Row9	62	145
Row10	10	142
Row11	25	130
Row12	43	127
Row13	53	114
Row14	114	99
Row15	88	92
Row16	32	71
Row17	77	70
Row18	104	70
Row19	87	67

Figure 8.9 (13): Table of top 20 clusters.

The table below lists the top 20 clusters and includes the cluster ID and its representative for each cluster.

Representative	**Cluster_ID**
Πίνακας απασχολούμενου προσωπικού	14
Επιπρόσθετα δικαιολογητικά κατά περίπτωση.	65
Πιστοποιητικό ή βεβαίωση μικροβιολογικού ελέγχου του θαλασσινού νερού του οποίου γίνεται χρήση, το οποίο εκδίδεται από την Υγειονομική Υπηρεσία της οικείας Περιφερειακής Ενότητας ή από άλλα διαπιστευμένα εργαστήρια σύμφωνα με την προβλεπόμενη διαδικασία μετά τη διενέργεια ελέγχων και δειγματοληψιών σύμφωνα με τις ισχύουσες υγειονομικές διατάξεις για τις κολυμβητικές δεξαμενές.	44
Καταστατικό της εταιρείας και τυχόν τροποποιήσεις αυτού, με τα αντίστοιχα ΦΕΚ δημοσίευσής τους ή πιστοποιητικά καταχώρισης στα βιβλία εταιρειών του Πρωτοδικείου της έδρας της εταιρείας.	5
Αποδεικτικό συνταξιοδότησης	12
Για την άδεια λειτουργίας: Πιστοποιητικά αντιεκρηκτικότητας ΑΤΕΧ των αντλιών, των διανομέων, των βυθομετρικών βεργών καθώς και του λοιπού εξοπλισμού, όπου απαιτείται από τη μελέτη.	6
1 . Σε περίπτωση που ο ενδιαφερόμενος εργαζόταν ως μισθωτός: Βεβαιώσεις Προϋπηρεσίας από την επιχείρηση που απασχολείται ο ενδιαφερόμενος με συνυπογραφή και από τον εποπτεύοντα (κάτοχο άδειας ή βεβαίωσης αναγγελίας), στην οποία πρέπει να αναγράφονται τα στοιχεία του εργοδότη και του εποπτεύοντος αδειούχου ή του κατόχου της βεβαίωσης αναγγελίας – αριθμός άδειας και εκδούσα	47

αρχή - τα στοιχεία του ενδιαφερομένου προσώπου, η διάρκεια της απασχόλησης - ακριβής χρόνος απασχόλησης σε ημερομίσθια -, η ειδικότητα κα	
Τυχόν γνωστοποιήσεις που έχουν προηγηθεί της μεταβολής.	59
Αντίγραφο ισχύουσας άδειας παραμονής εφόσον ο υποψήφιος είναι αλλοδαπός	256
Πιστοποιητικό (ενεργητικής) πυροπροστασίας από την αρμόδια πυροσβεστική Υπηρεσία.	62
Υπεύθυνη δήλωση Ν.1599/86 του ενδιαφερομένου στην οποία δηλώνει ότι: α. δεν είναι ιδιοκτήτης ή υπάλληλος ατομικής Ιδιωτικής Επιχείρησης Παροχής Υπηρεσιών Ασφάλειας SECURITY ή Γραφείου Ιδιωτικών Ερευνών, β. δεν είναι εταίρος ή διαχειριστής σε Ιδιωτική Επιχείρηση Παροχής Υπηρεσιών Ασφάλειας, που λειτουργεί ως εταιρεία περιορισμένης ευθύνης, ομόρρυθμη εταιρεία ή ετερόρρυθμη ή ιδιωτική κεφαλαιουχική εταιρεία Ε.Π.Ε., Ο.Ε., Ε.Ε., Ι.Κ.Ε., γ. δεν είναι μέλος του Δ.Σ. ή εκπρόσωπος εταιρείας οποιασδήποτε μορφής που σ	10
Άδεια λειτουργίας του αεροδρομίου	25
Τυχόν ληξιπρόθεσμες οφειλές ελέγχονται αυτεπάγγελτα από τον Δήμο κατά την υποβολή της γνωστοποίησης.	43
Τέσσερις (4) έγχρωμες φωτογραφίες πρόσφατης έκδοσης.	53
Φωτοτυπία Ταυτότητας	114
Βεβαιώσεις ή άλλα αποδεικτικά στοιχεία από τα οποία προκύπτει η εμπειρία εκτέλεσης	88

δοκιμών του προσωπικού του εργαστηρίου	
Αντίγραφο Πρωτοκόλλου Γενικής Επιθεώρησης ή Πιστοποιητικού Ασφαλείας για κάθε πλοίο	32
Υπεύθυνη Δήλωση ν. 1599/1986 ότι τα παραπάνω υποβληθέντα φωτοαντίγραφα αποτελούν γνήσια φωτοαντίγραφα εκ των πρωτοτύπων.	77
Για την έγκριση λειτουργίας: Έγκριση ίδρυσης εγκατάστασης εκτροφής και αναπαραγωγής ζώων συντροφιάς	104
Φωτοαντίγραφο του διαβατηρίου.	87

Using the Joiner node, we added the two dimensions obtained from the PCA node to the table of the top 20 clusters. Then, using the Color Manager node, we assigned a unique color to each cluster, resulting in the final table with the cluster information and the corresponding colors. This table can be used to create a scatter plot of the top 20 clusters, with each cluster being plotted in a different color.

Table "default" - Rows: 5632 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	S cluster_id	I Count (cluster_id)	D PCA di...	D PCA di...
Row0	5	330	0.382	0.634
Row1	5	330	0.365	0.301
Row2	5	330	0.383	0.668
Row3	5	330	0.335	0.495
Row4	5	330	0.081	0.243
Row5	5	330	0.081	0.243
Row6	5	330	0.063	0.314
Row7	5	330	0.081	0.243
Row8	5	330	0.081	0.243
Row9	5	330	0.081	0.243
Row10	5	330	0.081	0.243

Figure 8.9 (14): Dimensionality reduction table with colors.

Using the Scatter Plot node and the information in the final table, we can generate a scatter plot that visualizes the top 20 clusters and their corresponding colors. The resulting plot should show the different clusters as distinct groups of points, with each group having a unique color. The plot may also include axes to show the two dimensions of the data, as well as any other relevant details such as labels or legends.

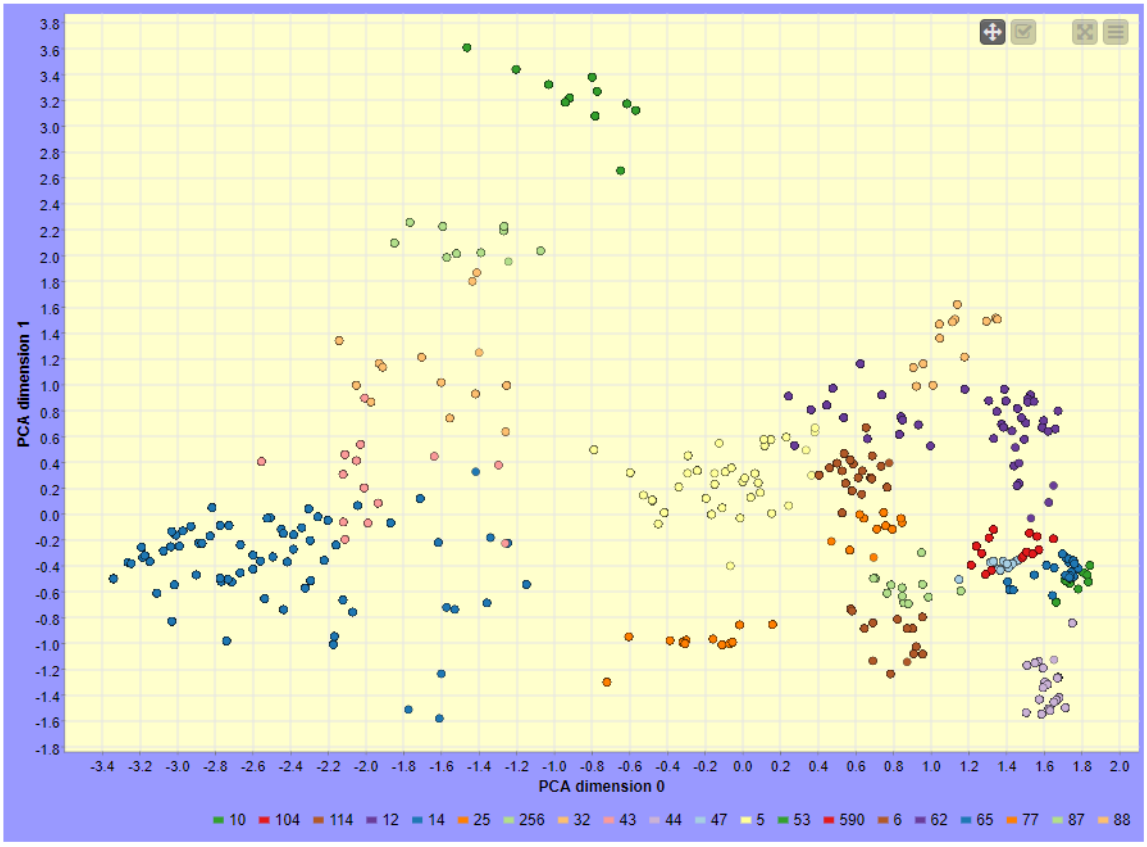

Figure 8.9 (15): Scatter plot of the top 20 clusters

9. Conclusions

Overall, it appears that the quality of the data in our dataset is mixed, with some strengths and some areas for improvement.

The results have shown that the dataset has relatively high overall completeness, with 81.72% of the rows having no missing values. This statement indicates that while the dataset includes a considerable amount of valuable information, it is presented in a format with low data quality. Although an 81.72% completeness rate may appear to be high, it is important to consider that the appropriateness of this percentage can vary depending on the intended use of the data. However, it's also important to note that there are some columns with a large number of missing values, which could potentially impact the usefulness of the data in those columns.

The data in this dataset appears to be timely and up-to-date, with 83% being updated on a monthly basis. However, it is important to consider that the appropriateness of this level of timeliness may vary depending on the intended use of the data. However, there is a small portion of the data (2%) that has not been updated in over a year, which could potentially impact the usefulness and relevance of that data. Additionally, another 8% of the data has not been updated in over 6 months. It's important to consider how frequently the data needs to be updated for an intended use case and whether the current update schedule is sufficient.

In terms of accuracy, some columns have a relatively high level of accuracy, with over 90% of the values being classified as "Accurate". However, there are also some columns with a relatively low level of accuracy, with less than 60% of the values being classified as "Accurate". Additionally, there are some columns with a significant number of missing values, which could potentially impact the accuracy of the data in those columns. It's important to consider the accuracy of the data in the context of an intended use case and take steps to improve the accuracy of any columns with low levels of accuracy or a significant number of missing values. Enhancing the data with metadata could potentially improve the accuracy and relevance of the data.

For the uniqueness dimension, it's good to see that some of the columns have unique values, which could be useful for certain, types of analysis. Two columns are notable because, aside from the first 10 rows, all of the other values are unique. Additionally, with the exception of the Null values count in one of the columns, the count for the first 10 rows ranges between 2 and 4. This suggests that further analysis may be necessary in the future to determine if these values are linked to duplicate records. If they are, then these two columns will also contain unique values.

In terms of availability, we have attempted to verify the functionality of the links in the "process_application_related_url" column. However, it appears that the code used to classify the links as working or non-working was not entirely accurate, with some working links being misclassified as non-working due to formatting issues. It might be worth revisiting this aspect of the analysis to ensure that the links are correctly classified and to take steps to improve the availability of any non-working links. Moreover it would be more useful if the links were already in the correct format to prevent any inconsistencies in analysis in the future.

For the credibility dimension, the data is sourced from an official government site, which suggests that it is reliable.

For the processability dimension, the data (excluding the links from the "process_application_related_url" column) is formatted in a way that is easily readable and processable by machines.

Overall, it's important to regularly review the quality of the data and take any necessary steps to ensure that it is suitable for any intended use case. This may involve updating or improving the data, or identifying and addressing any issues or weaknesses in the data.

On the overall text similarity, clustering, and topic modeling section, it seems that the sentence transformer model and the clustering algorithm were successful at grouping descriptions into clusters that make sense. The use of the centroids and representative topics for each cluster can be useful for classifying new data and understanding the main themes within each cluster. However, it is worth noting that there are still many clusters

with low counts, which may indicate that the clustering is not as effective as it could be. Finally, it seems that the results of the analysis are promising and could be useful for understanding the relationships between descriptions and evidence.

Knime is a powerful analytical tool, but it has its limitations. While some tasks are easier to accomplish in Knime than through coding, others are not. As a result, I had to use Python in conjunction with Knime's nodes to complete certain tasks. However, the Knime community is continually adding new features and functionality, so it is likely that the analytical tool will eventually be able to perform all necessary tasks without the need for any additional coding.

10. Limitations and suggestions for future research

The current research has provided valuable insights into the quality of data for processes. Additionally, we have been able to successfully organize descriptions into clusters and assign a representative, a centroid, and a topic to each cluster, which serve as descriptive summaries of the cluster's content. However, there is still room for further exploration and expansion in future studies.

Based on the fact that certain columns had to be excluded from the analysis due to a high number of missing values, and considering that the relevance dimension could potentially be improved, we have determined that there is a need to enhance the data with additional metadata. As mentioned in section 4, this could be achieved through the incorporation of open government data related to the public sector, adherence to European data standards (section 4.1), or the use of controlled vocabularies (section 4.2). Enhancing data with metadata is an essential step in improving the completeness and relevance dimensions discussed in sections 3.2.1 and 3.2.8. This can ultimately lead to better decision-making and more effective use of the data within the government.

Furthermore, the analysis carried out in section 8 could serve as a foundation for future research to examine the other columns within the same cluster and compare them to identify any duplicate records in the database that are described using different wording. This would be extremely valuable for any organization, particularly for governments, as it could significantly improve the accuracy and uniqueness dimensions discussed in sections 3.2.3 and 3.2.4. Improving these dimensions could lead to more efficient and effective data management within an organization.

Finally, it may be possible to further refine the coding for the availability dimension (section 3.2.5) in order to produce more accurate results. As we have observed, certain links do not adhere to the standard HTTP format, which can negatively impact the accuracy and processability dimensions discussed in sections 3.2.3 and 3.2.7. It would be beneficial for the government to be made aware of these issues and to take steps to address them, such as by updating the links on the mitos.gov website or

improving the formatting of the links. Overall, this would help ensure the availability and usability of the data within the organization.

References

[1] McGilvray, D. (2021). *Executing data quality projects: Ten steps to quality data and trusted information (TM)*. Academic Press.

[2] Cichy, C., & Rass, S. (2019). *An Overview of Data Quality Frameworks. IEEE Access, 1–1*.

[3] Moses B, Gavish L & Vorweck M, (2022), Data Quality Fundamentals. A Practitioner's Guide to Building More Trustworthy Data Pipelines

[4] Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325-337.

[5] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). *Data quality assessment. Communications of the ACM, 45(4), 211*.

[6] Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.

[7] Brassel K, Bucher F, Stephan E-M, Vckovski A (1995) Completeness. In Guptill S C, Morrison J L (eds) Elements of spatial data quality. Oxford, Elsevier Science: 81–108

[8] Šlibar, B., Oreški, D., & Begičević Ređep, N. (2021). Importance of the open data assessment: An insight into the (meta) data quality dimensions. *SAGE Open*, *11*(2), 21582440211023178.

[9] Heinrich, B., Klier, M., & Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data and Information Quality (JDIQ)*, *1*(1), 1-28.

[10] Gordon, K. (2013, November). Principles of data management: facilitating information sharing second edition. BCS.

[11] European Commission, (2014). Open Data & Metadata Quality

[12] Gartner, R., & Gartner, R. (2016). *Metadata*. Springer.

[13] Peckham, J., & Maryanski, F. (1988). Semantic data models. *ACM Computing Surveys (CSUR)*, *20*(3), 153-189.

[14] Richesson, R. L., & Krischer, J. (2007). Data standards in clinical research: gaps, overlaps, challenges and future directions. *Journal of the American Medical Informatics Association*, *14*(6), 687-696.

[15] Hoberman, S. (2015). *Data Model Scorecard: Applying the Industry Standard on Data Model Quality*. Technics Publications.

[16] European Data Portal, (2018). Open Data Goldbook for Data Managers and Data Holders, practical guidebook for organisations wanting to publish open data.

[17] Federal Enterprise Data Resources, Data Standards

[18] McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., ... & Parkinson, H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS biology*, *15*(6), e2001414.

[19] Harpring, P. (2013). *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works*. Getty Publications.

[20] Lima, C., Zarli, A., & Storer, G. (2007). Controlled vocabularies in the European construction sector: evolution, current developments, and future trends. In *Complex Systems Concurrent Engineering* (pp. 565-574). Springer, London.

[21] Paredes, L. P., Rodrıguez, J. M., & Azcona, E. R. (2008). Promoting government controlled vocabularies for the Semantic Web: the EUROVOC thesaurus and the CPV product classification system. *Semantic Interoperability in the European Digital Library*, *111*.

[22] Strohmeier, R. W., Gemert, W. V., Laaboudi, C., & Waniart, A. (2017). Optimizing subject access to legal resources: EuroVoc and VocBench. Benefits of using multilingual controlled vocabularies and an open source collaborative tool for their maintenance.

[23] Publications Office of the EU, EuroVoc Handbook.

[24] Qurashi, A.W., Holmes, V. and Johnson, A.P., 2020, August. Document processing: Methods for semantic text similarity analysis. In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-6). IEEE.

[25] Huang, A., 2008, April. Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (Vol. 4, pp. 9-56).

[26] Gunawan, D., Sembiring, C.A. and Budiman, M.A., 2018, March. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series* (Vol. 978, No. 1, p. 012120). IOP Publishing.

[27] Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. and Sung, Y.H., 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

[28] Reimers, N. and Gurevych, I., 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

[29] Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

[30] Ridge, T., 1990. Accuracy and Relevance and the Quality of Data. *Data quality control: Theory and pragmatics*, *112*, p.105.

[31] Chen, Q., Peng, Y. and Lu, Z., 2019, June. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1-5). IEEE.

[32] Ferrari, A., Witschel, H.F., Spagnolo, G.O. and Gnesi, S., 2018. Improving the quality of business process descriptions of public administrations: Resources and research challenges. *Business Process Management Journal*.

[33] Safarov, I., Meijer, A. and Grimmelikhuijsen, S., 2017. Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, *22*(1), pp.1-24.

[34] Codd, E.F., 1980, June. Data models in database management. In *Proceedings of the 1980 workshop on Data abstraction, databases and conceptual modeling* (pp. 112-114).

[35] Yoon, V.Y., Aiken, P. and Guimaraes, T., 2005. Applying a Metadata Framework to Improve Data Quality. In *Encyclopedia of Information Science and Technology, First Edition* (pp. 146-151). IGI Global.

[36] Moraga, C., Moraga, M.Á., Caro, A. and Calero, C., 2009, August. SPDQM: SQuaRE-aligned portal data quality model. In *Ninth International Conference on Quality Software, QSIC* (pp. 24-25).

[37] Kučera, J., Chlapek, D. and Nečaský, M., 2013, August. Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective* (pp. 152-166). Springer, Berlin, Heidelberg.

[38] Jariwala, A., Chaudhari, A., Bhatt, C. and Le, D.N., 2022. Data Quality for AI Tool: Exploratory Data Analysis on IBM API. *International Journal of Intelligent Systems & Applications*, *14*(1).

[39] Richard Y. Wang & Diane M. Strong (1996) Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 12:4, 5-33

[40] Gartner, R., 2016. Breaking the Silos. In *Metadata* (pp. 87-96). Springer, Cham.

[41] Wibowo, M., Sulaiman, S. and Shamsuddin, S.M., 2017, July. Machine learning in data lake for combining data silos. In *International Conference on Data Mining and Big Data* (pp. 294-306). Springer, Cham.

[42] Patel, J., 2019. Bridging data silos using Big Data integration. *International Journal of Database Management Systems*, *11*(3), pp.01-06.

[43] Holzer, A., König, W., & Weinhardt, C. (2003). Data Quality in E-Government: Challenges and Solutions. ResearchGate.

[44] Wang, H.-F., & Liu, Y.-H. (2004). Data Quality Management in Government: An Exploratory Study. ResearchGate.

[45] Wixom, B., & Schwester, D. (2003). Improving Data Quality in Public Sector Organizations. Journal of Management Information Systems, 20(3), 199-222.

[46] Wixom, B., Schwester, D., & Loshin, D. (2004). Data Quality in Public Sector Organizations: A Review and Research Agenda. Journal of Management Information Systems, 21(1), 199-236.

[47] Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K. and Wiswedel, B., 2009. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, *11*(1), pp.26-31.

[48] Tan, P., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

[49] Jain, A. K., Dubes, R. C., & Narasimha, M. (1988). Algorithms for Clustering Data. Prentice Hall.

[50] Wolf, M. and Wicksteed, C., 1998. Date and time formats. *W3C NOTE-datetime-19980827, August*.

[51] Abdi, H. and Williams, L.J., 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), pp.433-459.

[52] Shlens, J., 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

[53] Bro, R. and Smilde, A.K., 2014. Principal component analysis. *Analytical methods*, *6*(9), pp.2812-2831.

[54] Levitin, A. and Redman, T., 1995. Quality dimensions of a conceptual view. *Information Processing & Management*, *31*(1), pp.81-88.

[55] Silveira, R., Fernandes, C.G., Neto, J.A.M., Furtado, V. and Pimentel Filho, J.E., 2021. Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, *1613*, p.0073.

[56] Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W., 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), pp.1-22.

[57] Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B. and Rand, D.G., 2014. Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), pp.1064-1082.

[58] Speiser, S. and Harth, A., 2010, September. Taking the lids off data silos. In *Proceedings of the 6th International Conference on Semantic Systems* (pp. 1-4).

[59] Dayal, U., Castellanos, M., Simitsis, A. and Wilkinson, K., 2009, March. Data integration flows for business intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 1-11).

[60] Field, D. and Sansone, S.A., 2006. A special issue on data standards. *OMICS: a Journal of integrative Biology*, *10*(2), pp.84-93.

[61] Mishra, S., Arunkumar, A., Sachdeva, B., Bryan, C. and Baral, C., 2020. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*.

[62] Chen, C., et al. (2016). Assessing Data Quality of Open Government Data Using Natural Language Processing. Government Information Quarterly, 33(2), 257-266.

[63] O'Connor, A.J., et al. (2013). Using Natural Language Processing to Assess Data Quality in Clinical Text. Journal of the American Medical Informatics Association, 20(6), 981-987.

[64] Gill, T., 2008. Metadata and the Web. *Introduction to metadata*, *3*, pp.20-38.