University of West Bohemia
Faculty of Applied Sciences
Department of Computer Science and Engineering

Master's Thesis

Methodology Design for Dataset Quality Assessment

Pilsen, 2021 Marek Lovčí

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2020/2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Marek LOVČÍ**
Osobní číslo:	**A19N0093P**
Studijní program:	**N3902 Inženýrská informatika**
Studijní obor:	**Informační systémy**
Téma práce:	**Návrh metodiky pro vyhodnocení kvality datových sad**
Zadávající katedra:	**Katedra informatiky a výpočetní techniky**

Zásady pro vypracování

1. Seznamte se se současnými metodikami pro vyhodnocení kvality datových sad.
2. Navrhněte metodiku umožňující univerzální strukturovaný postup pro ohodnocení datasetů.
3. Ověřte možnost automatické klasifikace zvolených datových sad z hlediska kvality.
4. Proveďte zhodnocení dosažených výsledků.

Rozsah diplomové práce: **doporuč. 50 s. původního textu**
Rozsah grafických prací: **dle potřeby**
Forma zpracování diplomové práce: **tištěná**

Seznam doporučené literatury:

dodá vedoucí diplomové práce

Vedoucí diplomové práce: **Doc. Dr. Ing. Jana Klečková**
Katedra informatiky a výpočetní techniky

Datum zadání diplomové práce: **11. září 2020**
Termín odevzdání diplomové práce: **20. května 2021**

L.S.

Doc. Dr. Ing. Vlasta Radová
děkanka

Doc. Ing. Přemysl Brada, MSc., Ph.D.
vedoucí katedry

V Plzni dne 24. září 2020

Návrh metodiky pro vyhodnocení kvality datových sad

Zadání

1. Seznamte se se současnými metodikami pro vyhodnocení kvality datových sad.

2. Navrhněte metodiku umožňující univerzální strukturovaný postup pro ohodnocení datasetů.

3. Ověřte možnost automatické klasifikace zvolených datových sad z hlediska kvality.

4. Proveďte zhodnocení dosažených výsledků.

Methodology Design for Dataset Quality Assessment

Assignment

1. Learn about current methodologies for assessing the quality of datasets.

2. Create a methodology that allows for a universal structured procedure for dataset evaluation.

3. Examine the ability to classify selected datasets for quality automatically.

4. Analyze the results obtained.

Poděkování

Tímto bych rád poděkoval Doc. Dr. Ing. Janě KLEČKOVÉ za odborné vedení, za cenné rady a čas, který strávila čtením a konzultací této práce.

Prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

I hereby declare that this master's thesis is completely my own work and that I used only the cited sources.

Pilsen ....

Marek LOVČÍ

Abstrakt

Tato diplomová práce se zabývá evaluací kvality datových sad. Shrnuje metodiky, které jsou současným standardem v oboru a na jejich bázi definuje metodiku novou. V další části práce byla prozkoumána možnost automatické klasifikace kvality datových sad a navržen algoritmus, který požadavek splňuje. V poslední části práce byla metodika i klasifikace předvedena na vyhodnocení kvality katalogu s daty COVID-19.

Klíčová slova Kvalita dat, kvalita informací, hodnocení kvality informací, hodnocení kvality dat, COVID-19

Abstract

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

Keywords Data Quality, Information Quality, Information Quality Assessment, Data Quality Assessment, COVID-19

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Today, we live in what many refer to as the Information Age, in which digital data production is central to all ecosystems. Data is being used to drive growth in businesses of all sizes, large and small. All industries require the use of data to analyze, manage, and control various systems. Decisions based on data are a rapidly growing phenomenon in the business world. And now, with the data, a business owner or manager can make more effective strategic decisions [8]. Making decisions can be risky at times due to the possibility of inaccurate or insufficient data. However, because the data is unstructured and complex, maintaining and governing it is critical for organizations [7]. But how can we be certain that the data is objectively correct?

As Orr (1998) stated,

> data quality is the measure of the agreement between the data views presented by an information system and the same data in the real world. A system's data quality of 100% would indicate, for example, that our data views are in perfect agreement with the real world, whereas a data quality rating of 0% would indicate no agreement at all [28].

Assuring a certain level of data quality is a standard IT project. There are some initiatives for improving Open Data Quality, such as 5 Star Data, but none of them are comprehensive. In short, objectively assessing data quality is difficult.

The main issue with this topic is that data quality is shrouded in misconceptions. Data quality is a business issue, not an IT issue. However, IT enables the business to improve itself by providing tools and processes. Bad data has an impact on every system and every person who interacts with it. As a result, it should be everyone's responsibility to uphold good standards and practices, which will increase trust in data used for reporting and analytics.

There are two major reasons for Data Quality Management implementation failure. The first one is related to a lack of data quality processes, such as a lack of proactive DQ surveillance [31]. The second is a scarcity of data quality measurements [21].

The cost of bad data is defined as *direct* + *indirect* costs. Manual and automatic master data cleaning incurs direct costs. [21]. Indirect cost, on the other hand, is financial loss caused by poor-quality master data which leads to (i) inadequate managerial decision, (ii) process failure and (iii) missed opportunity [21].

Despite the fact that the goal of data quality assessment is to reduce *costs* and *complexity*, the data quality process can still result in low quality master data. The typical perpetrators in such cases are (i) lack of DQ measurements (or faulty definition) and (ii) absence of clear roles in the data life-cycle process [21].

The solutions to the problems mentioned above are as follows (i) a data model definition (metadata) and (ii) proactive data quality surveillance [31].

"One certain way to improve the quality of data: improve its use!" [28]

Work emphasis

There are many things that must be taken into account when implementing the DQ methodology. To name a few [26]:

- stakeholders and participants' involvement,
- metadata management,
- architecture styles,
- functional services,
- data modeling,
- data consolidation and integration,
- management guidance,
- master data identification and
- master data synchronization.

Batini et al. (2009) recognized activities of data quality methodology. In the most general case, the list is composed of four phases listed below [5]. Additional steps are defined in each of the four sections, but we will not go over them in detail here.

1. *State reconstruction*, which is aimed to get information about business processes and services, data collection, quality issues, and corresponding costs.
2. *Measurement*, where the objective is to measure the quality of data collection along relevant quality dimensions.
3. *Assessment*, which refers to the event when *measurements* are compared with certain reference values to determine the state of quality and to assess the causes of poor data.
4. *Improvement* concerns the selection of the steps, strategies, and techniques for reaching new data quality targets.

To narrow the scope, this thesis will concentrate on these components in order to achieve a good and reasonable goal within a limited time and resource: (i) Definition of Data Quality Methodology and (ii) Data Quality Score Measurement & Assesment.

Measurement of Quality (MoQ) is part of the *measurement phase*. The idea is to select the quality dimensions affected by the quality issues identified in the DQ requirements analysis and define corresponding metrics [26]. Measurement can be objective when it is based on quantitative metrics, or subjective, when it is based on qualitative evaluations by data administrators and users [26].

To complete this assignment, we will concentrate on automatic assessment of the final dataset score and define a procedure to objectively evaluate the quality score of the given dataset. As a result, we will be able to assign a score to the dataset, giving us an idea of its current qualitative state. Fully automatic evaluation of Quality Scores will not be possible, as we will see in Chapters 4 and 5. Therefore, we will use a semi-automatic approach, in which we will evaluate the qualitative foundation of the dataset, but subsequent levels will be evaluated automatically using the “drill-up” approach.

Research Purpose

Data quality is a never-ending topic of discussion and is critical in a variety of fields, including telecommunications, healthcare, manufacturing, banking, and insurance, among others. There are numerous characteristics and methodologies that contribute to good data quality, and they vary depending on the domain, with certain data characteristics being more important than others. The goal of this research is to understand the characteristics that contribute to data quality in any domain.

The primary research goal of this study is to define and apply methodology for assessing data quality, as well as to identify, collect, analyze, and evaluate quality metrics for data in order to quantify and improve their value. To suit the specifics of our case study in the Chapter 5, we will select the principal characteristics that contribute to data quality in the field of selected data. The quality metrics chosen should be those that objectively quantify data value.

We will use and assess data from the (currently ongoing) COVID-19 pandemic. The reason for the selection is the general availability, large collection, and open license of selected datasets.

Chapter 2

Related Work

To answer the main thesis question, a review of existing studies will be needed. The topic of DQ and the cost on business is well researched. One of the oldest articles was written by Gerald A. Feltham in 1968 with title “The Value of Information”. Many articles and studies were written on the topic since then, therefore we can recognize some basic structures when talking about DQ methodology.

2.1 Hybrid Approach

In the article *Data quality assessment: The Hybrid Approach* the authors defined data quality as “fit for use”. They reviewed several assessment techniques, including:

- AIMQ (Lee et al., 2002),
- TQDM (English, 1999),
- cost-effect of low data quality (Loshin, 2004) and
- subjective-objective data quality assessment (McGilvray, 2008).

The result of the study is a general framework for creating customized, bussiness unique data quality assessment process. The process consists of seven consecutive activities: (i) select data items, (ii) select a place where data is to be measured, (iii) identify reference data, (iv) identify DQ dimensions, (v) identify DQ metrics, (vi) perform measurement and (vii) conduct analysis of the results.

The methodology can be summarized as follows. The input data (i) are measured (vi) and thus the dimensions (iv) and metrics (v) are obtained. Metrics are applied to the data in the central repository (ii). If necessary, the data may be validated against the reference data (iii).

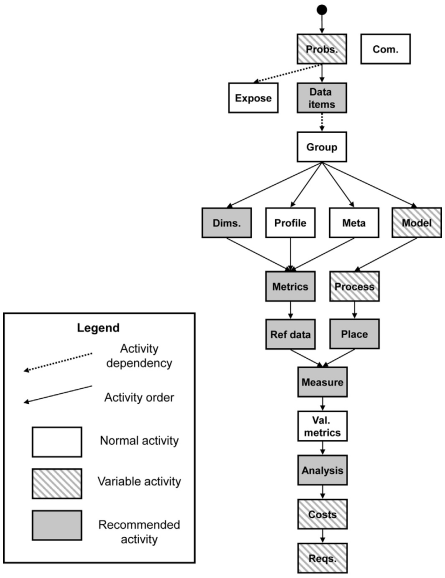
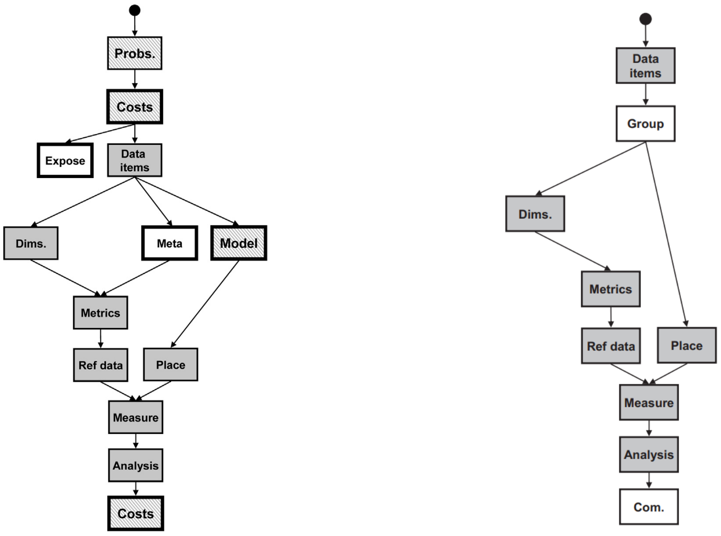


Figure 2.1: A generic AT according to the Hybrid Approach methodology [40]

The methodology is tested by the authors on two practical cases. The first use case is to adapt the framework for an MRO (Maintenance Repair and Operations) company. The second use case is the adaptation of the methodology for the London Underground.

A very important result of the study is a configurable process model. It is possible to design an alternative configuration of the process model to suit the case study or specific domain.



(a) An AT for the MRO organisation (b) An AT for London Underground

Figure 2.2: ATs for the case studies [40]

In the *Hybrid Approach*, the ATs developed between 1998 and 2008 were incorporated, and they all suggest very similar ideas to evaluating DQ [40]. The methodology, thanks to the fact that it takes over the best practices of other methodologies, will be up-to-date for a long time. The only problem arises when multiple stakeholders demand conflicting requirements. If one party requires some activity and the other does not, the activity cannot simply be incorporated due to time and resource costs. A thorough analysis is needed in this regard.

2.2 AIM Quality

The AIM Quality is a information quality assessment and benchmarking methodology for Management Information Systems (MIS). The methodology consists of three main components, a model, a questionnaire to measure information quality, and analysis techniques for information quality interpretation. The methodology has been built on the foundations of other academic studies as well as professional white-papers (e.g., Department of Defense, HSBC, and AT&T), and has been validated on health organizations use cases.

The important components in AIMQ are the IQ dimensions, critical for the information consumers. The authors grouped IQ dimensions into four categories, *intrinsic* IQ (the information itself contains a certain level of quality), *contextual* IQ (quality must be considered within the business context), *representational* IQ (expressing whether the information is comprehensible in the information system) and *accessibility* IQ (expressing whether the information is accessible in the information system, but at the same time securely stored).

The information quality model in AIMQ, Product Service Provider (PSP/IQ) model, has four quadrants relevant to the IQ improvement decision process. The model is shown in the Table 2.1. This model can be used to evaluate how well a company develops *sound* and *useful* information products and delivers *dependable* and *usable* information services to the consumers.

	Conforms to specifications	Meets or exceeds consumer expectations
Product Quality	Sound information	Useful information
Service Quality	Dependable information	Usable information

Table 2.1: The PSP/IQ model [24]

2.2.1 Four PSP/IQ model quadrants

The next four paragraphs contain examples of DQ dimensions contained in each of the quadrants.

Sound Information Dimensions Free-of-error, Concise representation, Completeness, Consistent representation.

Useful Information Dimensions Appropriate amount, Relevancy, Understandability, Iterpretability, Objectivity.

Dependable Information Dimensions Timeliness, Security.

Usable Information Dimensions Believability, Accessibility, Ease of operation, Reputation.

2.3 Comprehensive Data Quality Methodology

A comprehensive data quality methodology (CDQM) for web and structured data is a methodology developed by Batini et al. (2008). The methodology consists of three main phases: (i) state reconsruction (modeling of organizational context), (ii) assessment (problem identification and DQ measurement) and (iii) choice of the optimal improvement process. From the last phase there is feedback to the previous phase.

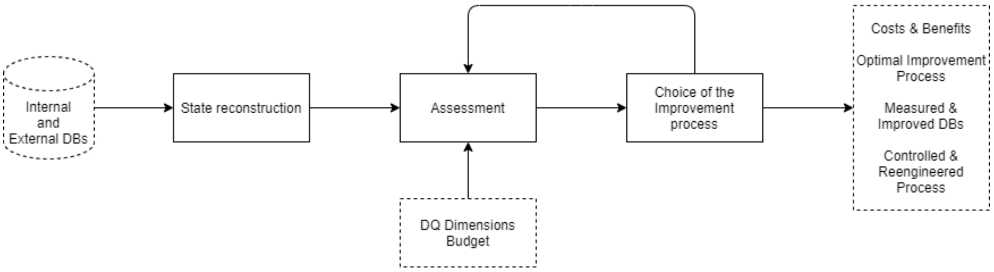

Internal and External DBs
State reconstruction
Assessment
Choice of the Improvement process
DQ Dimensions Budget
Costs & Benefits
Optimal Improvement Process
Measured & Improved DBs
Controlled & Reengineered Process

Figure 2.3: Diagram of the CDQ methodology [3]

2.3.1 State reconstruction

In the state reconstruction phase, the business/organizational context linked to internal and external data is modelled in terms of organizational units, processes, and rules [3]. This phase offers an overview of data providers and users, the flow of data and the use of data between them [3].

The state of data and their use-cases are recreated int he first step. For a meaningful representation of this knowledge, two matrices are used. The first one is Data Organizational Unit matrix. The matrix's cells indicate whether an organizational unit generates (i.e., owns) or utilizes a collection of data. The second one is Dataflow Organizational Unit matrix. In this case, each cell of the matrix indicates whether an entity is a data flow consumer or provider [3].

In the second step, the Process Organizational Unit matrix identifies and describes the owner and contributing units for each process. This matrix assists in the delegation of responsibility for quality improvement activities [3].

This step helps in provisioning of a comprehensive view of organizational processes and, as a result, aids in the decision-making process for quality improvement activities [3]. The Service Norm Process matrix is built to provide the information on how each macroprocess produce services for the clients and how the processes cooperate in the production of those services [3].

2.3.2 Assessment

In the assessment phase, internal and external users are involved to identify relevant DQ issues. After obtaining information about the DQ issues, it is necessary to define quantitative metrics to evaluate the severity of DQ problems.

Quality dimension/database	Duplicate objects	Matching objects	Accuracy	Currency
Social Security DB	5%		98%	3 months delay
Accident Insurance DB	8%		95%	5 months delay
Chamber of Commerce DB	1%		98%	10 months delay
The three databases		98%		

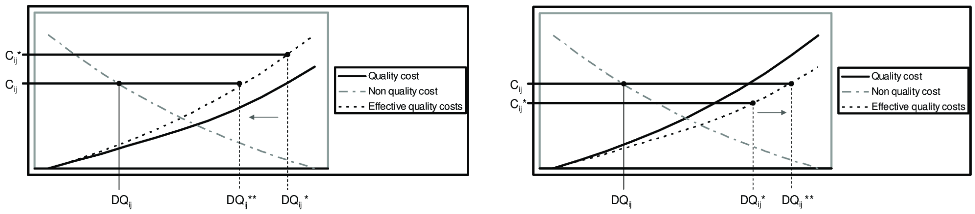
Table 2.2: Example databases, quality dimensions and metrics [3]

2.3.3 Choice of the optimal improvement process

The organisation must set target quality values DQ_{ij}^{*}, based on actual quality values DQ_{ij} linked with the *i-th* dataset and the *j-th* quality dimension, to be achieved through the improvement process [3]. DQ targets are defined by performing a *process-oriented* and a *cost-oriented* analysis [3].

In the *cost-oriented* analysis, the economic costs – that the business can afford for the DQ improvement process – needs to be defined. A major obstacle is the difficulty of estimating costs and benefits in advance.

A *non-quality costs* (C_{ij}) are the costs associated with poor data quality and, therefore, with all the inevitable activities to correct errors and re-execute tasks [3]. The evaluation of quality targets is in the Figure 2.4.



(a) Evaluation of less ambitious quality targets

(b) Evaluation of less ambitious quality targets

Figure 2.4: Evaluation of quality targets [3]

2.4 Business Oriented Data Quality

Otto et al. (2011) developed a design process for the identification of business oriented DQ metrics [29]. The paper does not present any concrete DQ metrics even though they studied data quality problems in three companies. Instead, those three companies' data problems were used to create an assumption that data defects cause business problems [29]. According to Otto et al. (2011), the identification of DQ metrics therefore should be based on how the data impacts process metrics [29].

A method engineering (ME) is used to design the framework. Methodology therefore consists of five components: (i) design activities, (ii) design results, (iii) meta-model, (iv) roles and (v) techniques.

2.4.1 Meta-model

Otto et al. (2011) describe entities and relations used to characterize the activities of the procedure model [29].

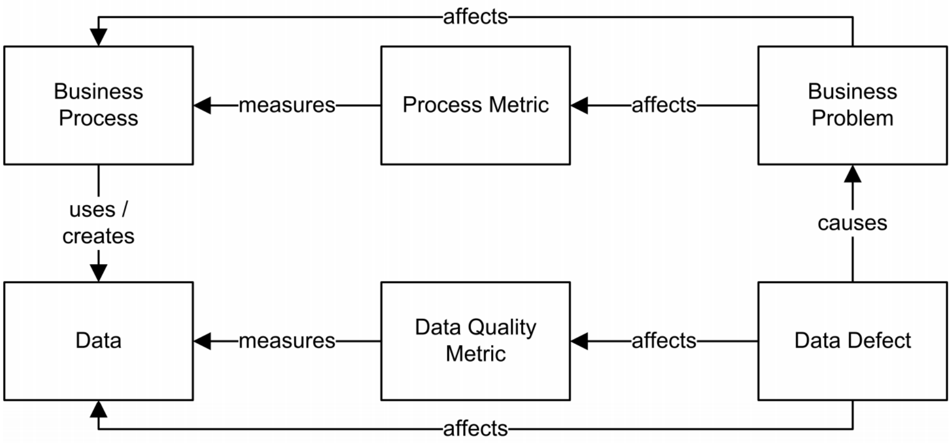


Figure 2.5: Entities and relations of a business oriented data quality metric [29]

Business Problem

Business problem is either system state (e.g. the package cannot be delivered) or incident (e.g. scrap parts production) causing decrease of system performance, therefore impacts *process metrics* results. It directly impacts business process and is defined by *probability of occurence*[1]and *intensity of impact*[2].

[1] Probability of occurence of event E can be denoted as $P(E) = \frac{r}{n}$, where r is number of ways E can happen from all possible ways n, $P(E) \in [0, 1]$.

[2] Intensity of impact is a measure of the time-averaged power density of a wave at a particular location. In our case, intesity should be defined as $I = \frac{\langle BC \rangle}{BA}$, where $\langle BC \rangle$ is time-averaged business cost of problem and BA business area through which the problem propagates during certain time frame, $I \in [0, \inf]$. If we define business area as sum of employees impacted by problem and their time spend to solve it, the unit of intensity would be costs per hour.

Business Process

By business process is meant sequence of tasks intended to generate value for customer and profit for the company. The business process is controlled and defined as part of a business strategy with corresponding modeling and measuring tools such as BPMN 2.0 or Key Performance Indicators (KPIs).

Process Metric

Quantitative measure of the degree to which a process fulfill a given quality attribute (e.g. scrap rate).

Data

Data is representation of objects and object relations.

Data Defect

It is an incident (e.g. wrong entered data), causing value decrease of data quality metrics. As well as *business problem*, a data defect poses a risk in terms of *probability of occurence* and *intensity of impact*.

Data Quality Metric

Quantitative measure of the degree to which data fulfill a given quality attribute (e.g. accuracy, consistency, currency,...).

2.4.2 Procedure Model

Procedure model defined by Otto et al. (2011) consists of three phases and seven activities. Activity flow model is shown in the figure 2.6. Letter color codes under the activities indicate degree of usage in the respective companies mentioned in the paper. Black color means that activity was fully used, grey color means partial usage and white indicates no use at all.

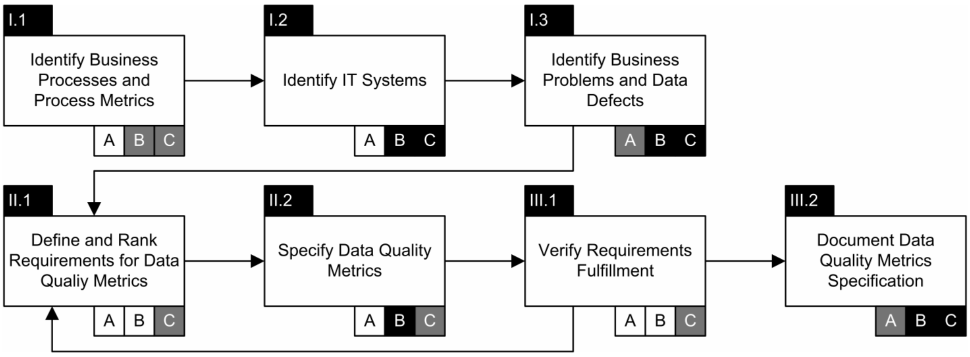

I.1
Identify Business Processes and Process Metrics
A B C
I.2
Identify IT Systems
A B C
I.3
Identify Business Problems and Data Defects
A B C
II.1
Define and Rank Requirements for Data Qualiy Metrics
A B C
II.2
Specify Data Quality Metrics
A B C
III.1
Verify Requirements Fulfillment
A B C
III.2
Document Data Quality Metrics Specification
A B C

Figure 2.6: Procedure model and degree of usage of activities in each case [29]

Phase 1

First phase is used to collect information. It consists of three activities:

1. Identify Business Processes and Process Metrics,
2. Identify IT Systems,
3. Identify Business Problems and Data Defects.

Phase 2

Second phase is used to specify requirements and design data quality mestrics. It consists of two activities:

1. Define and Rank Requirements for Data Quality Metrics,
2. Specify Data Quality Metrics.

Phase 3

Third phase is intended to approve and decument results. As well assecond phase, this one consists of two activities:

1. Verify Requirements Fulfillment,
2. Document Data Quality Metrics Specification.

2.4.3 Roles

In the last part, the authors declare six roles and their assignment to activities from section 2.4.2. Those roles are: (i) Chief Data Steward, (ii) Business Data Steward, (iii) Technical Data Steward, (iv) Process Owner, (v) Process user and (vi) Sponsor.

2.5 ORME

Batini et al. (2007) provided DQ assessment methodology called ORME (from italian word “orme” meaning track or trace). The methodology consists of four Data Quality Risk evaluation phases:

1. prioritization,
2. identification,
3. measurement,
4. monitoring [4].

The authors provided a comprehensive classification of the costs of poor data quality in their work. In short, they classified costs into three categories:

- current cost of insufficient data quality,
- cost of IT/DQ initiative to improve current quality status,
- benefits gained from improvement initiative implementation [4].

2.5.1 Prioritization

In this phase the model reconstruction happen. All the relationships among organization units, processes, services and data are put together and organized e.g. in the form of matrices (database/organization matrix, dataflow/organization matrix, database/process matrix) [4]. The main goal is to provide map of the main data use across data providers, consumers and flows [4].

2.5.2 Identification

This phase main focus is on identification of loss events and definition of overall economic loss metrics [4]. In this case, loss can be expressed in (i) absolute values (e.g. 100 USD), (ii) a percentage with respect to reference variables (e.g. 10% of GDP), or (iii) a qualitative evaluation (e.g. low-medium-high) [4].

2.5.3 Measurement

In this phase actual qualitative and quantitative assessment of data quality is conducted.

2.5.4 Monitoring

The last phase establishes a feedback loop and threshold in the DQ assessment process. DQ dimensions should be, according to the authors, evaluated periodically. Therefore quality rule violation allerts and automatic processes should be defined in order to ensure required DQ levels [4].

Authors suggest discriminant analysis as an easy and effective way of loss event identification. The goal is to identify loss event based on set of new values in the data source. The model is build on a training set, with two classes (*loss* and *no loss*) in consideration. A set of linear functions from predictors is constructed,

$$L = b_1x_1 + b_2x_2 + \ldots + b_3x_3 + c$$

where b_k are discriminant coeficient, x_k are input variables (predictors) and c is a constant [4].

2.6 Data Quality and Security

Given the continuous risk of data braches, we should consider the impact of security mechanisms on data quality. This topic is very timely, especially with the need to comply with the GDRP regulations. Data Quality and Data Security are two key issues that address various problems in Data Engineering, such as large volumes and diversity of data, credibility of data and their sources, data collection and processing speed, and so on [33]. Confidentiality, Integrity, and Availability are the three key security assets defined in terms of data security [33]. The ISO/IEC 25010 standard defines each of these properties in detail [22]. Whereas in terms of data quality, there is no consensus on any of the properties that define data quality or the precise definition of each property [5].

The main point of data security principle, particularly confidentiality and integrity, is to protect data from unauthorized access. However, implementing a comprehensive data quality management system necessarily requires unrestricted read and write access to all data [33]. Since the data quality system can share data with other systems or be accessed by individuals with different business interests, this requirement can lead to plenty of security issues [33]. As a result, data privacy can be a major obstacle to data quality. Security exceptions may be required by a quality management system, which poses potential security risks. This tensions between the two systems complicates their development and necessitates the emergence of new access control policies that allow quality processes to access the data they need without jeopardizing their security [33]. As Talha (2019) mentions, a sturdy *access control model* such as TBAC (Task Based Access Control), RBAC (Role Based Access Control), ABAC (Attribute Based Access Control), OrBAC (Organization Based Access Control), PuRBAC (Purpose-Aware Role-Based Access Control) must be used to fulfill the policy.

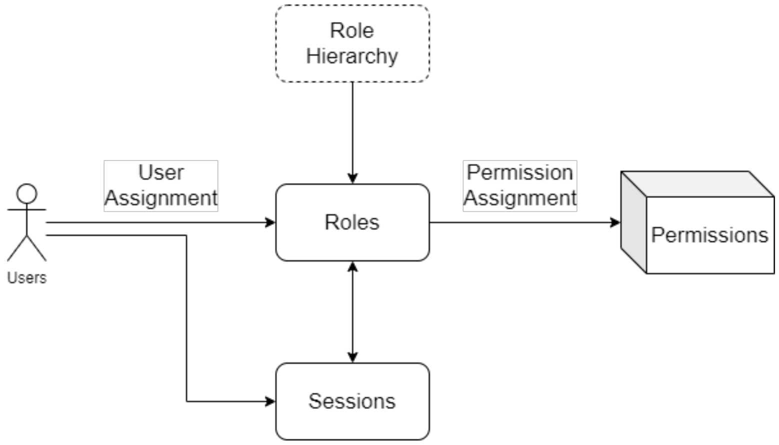


Figure 2.7: Role Based Access Control Model

Differential privacy has risen to prominence in applied mathematics as a leading data security technique, allowing accurate data analysis while preserving formal privacy guarantees [11]. Data often contains *sensitive attributes*, users' personal information, in a form of *personal identifiers* or *quasi-identifiers*. *Personal Identifiers* (PID) are data elements that identify a unique user in the dataset and allow another person to make the assumption of person's identity without their knowledge or consent (e.g., ID Number, Bank Account Number,...). A *quasi-identifier* is a set of attributes that, when combined with external information, can be used to reidentify (or reduce uncertainty about) all or some of the entities to whom information is being referred (e.g., gender, postal code, age or nationality) [30].

The personal data can be obscured using *anonymization* techniques, allowing for accurate data analysis. Narayanan (2008) proved, that *anonymized* data can be "easily" recovered using *linkage attack* (combining pieces of anonimized data to reveal one's identity). To prevent data misuse (re-identification of users) after a security breach, several more advanced models – *optimal k-anonymity*, *l-diversity*, *t-closeness* and *differential privacy* – exists to safeguard individuals' personal information in datasets.

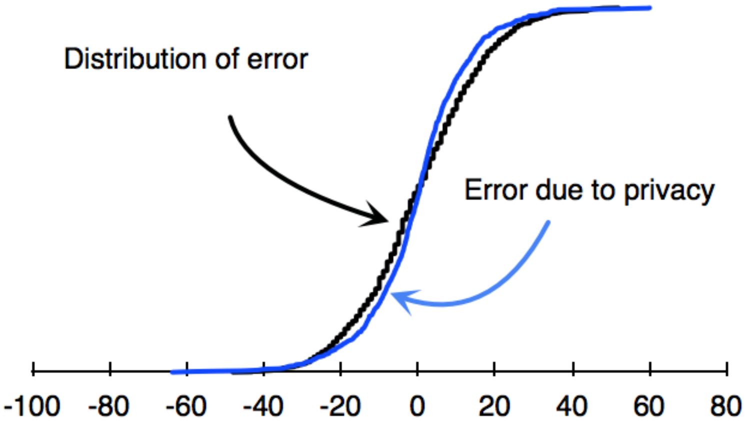


Figure 2.8: Simplified explanation of differential privacy error [14]

Differential privacy algorithm stores complete and trusted data in some cases only. In a certain subset of data, statistical noise is added, which compromises individual records (the record may or may not be true), but still allows accurate statistical analysis over the entire dataset [14]. By leveraging of Laplace distribution to spread data over a large *state space* and to increase the level of anonymity, the differential privacy model ensures that even if someone has full information on 99 of 100 people in a data collection, they will not be able to deduce information about the final user [41, 18]. This mechanism is interesting because it certainly affects the quality of the data, but in a different way than we would expect – making impossible to look at a specific record, but allowing analysis of the whole and the trend.

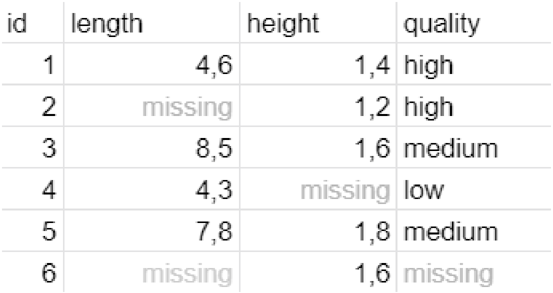
2.7 Business Problems and Data Defects

It is impossible to discuss data quality and methodologies for ensuring it without mentioning the specific types of issues that arise within the topic. In the field of data engineering and data science, there are a number of common data defects.

Missing Data

This is data that does not reach the destination data store [9]. This problem usually occurs when handling the data needed to clean up in the source database; by operating with invalid or incorrect lookup table in the transformation logic; or by invalid table joins. An example of missing data is shown in Figure 2.9.

Example We transform data from Task Management Solution. Lookup table should contain a field value of “Minor” which maps to “Low”. However, source data field contains “Mino” - missing the *r* and fails the lookup, resulting in the target data field containing null. If this occurs on a key field, a possible join would be missed and the entire row could fall out.



id	length	height	quality
1	4,6	1,4	high
2	missing	1,2	high
3	8,5	1,6	medium
4	4,3	missing	low
5	7,8	1,8	medium
6	missing	1,6	missing

Figure 2.9: Missing Data

Truncation of Data

Many data is being lost by truncation of the data fields. This happens when there are invalid field lengths on target database or by transformation logic not taking into account field lengths from the source [9].

Example We transform financial data with complete exchange-traded fund (ETF) names. Source field value “iShares Global High Yield Corp Bond UCITS ETF” is being truncated to *varchar(32).* since the source data field did not have the correct length to capture the entire field, only “iShares Global High Yield Corp B” is stored.

Data Type Mismatch

Data types not setup correctly on target database cause serious problems. This usually happens when using ETL pipeline with an automatic or semi-automatic column type recognition [9]. The data engineer relies on error-free data type recognition and does not check the accuracy of the output tables.

Example Source data field was required to be a *varchar*, however, when initially configured, was setup as a *date*.

Null Translation

In the source dataset, *null* values are not being transformed to correct target values [9]. Development team did not include the *null* translation in the ETL process.

Example A “Null” source data field was supposed to be transformed to “None” in the target data field. However, the logic was not implemented, resulting in the target data field containing “null” values[3].

Wrong Translation

Wrong translations happen when a source data field for *null* was supposed to be transformed to “None” in the target data field, but was not transformed correctly [9]. The logic was not implemented, resulting in the target data field containing *null* values. Wrong translation is the exact opposite to *Null Translation*.

Example Target field should only be populated when the source field contains certain values, otherwise should be set to null. Let’s look at a very basic example. During analytical processing of medical data (e.g., list of patients with oncological finding), we need to set target field to *true* if the one or multiple source values indicate certain treatment. However, the target field is populated (either with blank charater or other values) although source values do not correspond to the required logic.

Misplaced Data

If the source data fields are not being transformed to the correct target data fields, we call the issue “Misplaced Data” [9]. One of the possible causes is that development team inadvertently mapped the source data field to the wrong target data field.

Example A source data field was supposed to be transformed to target data field “Last_Update”. However, the development team inadvertently mapped the source data field to “Date_Created”.

Extra Records

Records which should be excluded in the ETL are included in the ETL. This happens when developers do not include filter in their code [9].

Example If a record has the deleted field populated, the record and any data related to that record should not be in any ETL.

[3] None is a concept that describes the absence of anything at all (nothingness), while Null means *unknown* (we do not know if there is a value or not).

Not Enough Records

Records which should be in the ETL are not included in the ETL. Development team had a filter in their code which should not have been there [9].

Example If a record was in a certain state, it should be sent through ETL pipeline over to the data warehouse.

Transformation Logic Errors

Testing sometimes can lead to finding "holes" in the transformation logic or realizing the logic is unclear [9].

Sometimes the processes are overly complicated, and the development team fails to account for special cases. Most cases fall into a certain branch of logic for a transformation, but a small subset of cases (sometimes with unusual data) may not fall into any branches [9]. How the analytics and developers handles these cases could be different (and may both end up being wrong) and the logic is changed to accommodate the cases. The next reason why this happens is that analytic and developer have different interpretation of transformation logic, which results in different values [9]. As a result, the logic is rewritten to make it more clear.

Example Foreign country cities that contain special language specific characters might need to be dealt with in the ETL code (e.g., Århus).

Simple and Small Errors

Capitalization, spacing and other small errors cause problems with data. Data inconsistencies are easy to fix, but happen often [9]. The only real solution is to always double check data and ETL procedure [9].

Sequence Generator

Ensuring that the sequence number of reports are in the correct order is very important when processing follow up reports or answering to an audit. If the sequence generator is not configured correctly, procedure results in records with a duplicate sequence number [9].

Example Duplicate records in the sales report were doubling up several sales transactions which skewed the report significantly.

Undocumented Requirements

During ETL development, sometimes certain requirements are found, that are "understood" but are not actually documented anywhere. This causes issues when members of the development team do not understand or misunderstood the undocumented requirements [9].

Example ETL pipeline contains a restriction in the "where" clause, limiting how certain reports are brought over. Moreover, there were used mappings that were understood to be necessary, but were not actually in the requirements. Occasionally, it turns out that the understood requirements are not what the business wanted.

Duplicate Records

Duplicate records are two or more records that contain the same data. This issue happens when development team does not add the appropriate code to filter out duplicate records or there is some unexpected error in data generators.

Example Duplicate records in the sales report were doubling up several sales transactions which skewed the report significantly.

Numeric Field Precision

Numbers that are not formatted to the correct decimal point or not rounded per specifications cause precision problems. This has several causes, development team rounded the numbers to the wrong decimal point, used wrong rounding type or used wrong data type which lead to faulty rounding [9].

Example The sales data did not contain the correct precision and all sales were being rounded to the whole dollar.

Rejected Rows

Data rows that get rejected by ETL process due to data issues. Development team did not take into account data conditions that break the ETL for a particular row [9]. An example of an ETL process with rejected rows is shown in Figure 2.10.

Example Missing data rows on the sales table caused major issues with the end of year sales report.

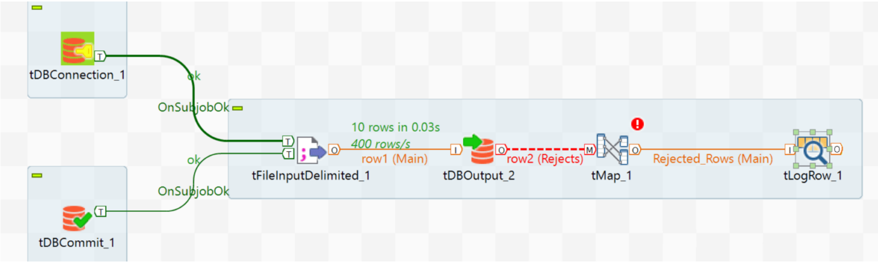


Figure 2.10: Talend ETL Rejects Rows

Chapter 3

Methodology

The proposed data quality methodology will have two major components, a model and supporting processes. The model defines the activities, their descriptions, goals, and the order in which they must be completed in order to ensure data quality. Support processes will then provide additional value by increasing the security and timeliness of datasets.

3.1 Model

The methodology has several important components that need to be identified or developed. The metamodel that covers the required components is as depicted in the Figure 3.1. The activities within the process model have a goal to develop those components.

Overall, the methodology consists of two main processes. The first one is **Specification Process**. The goal of this processs is to identify and define context specific ways to measure data quality. The second one is an **Execution Process**. Its main goal is to *collect* and *verify* data with output from *Specification Process* taken into account.

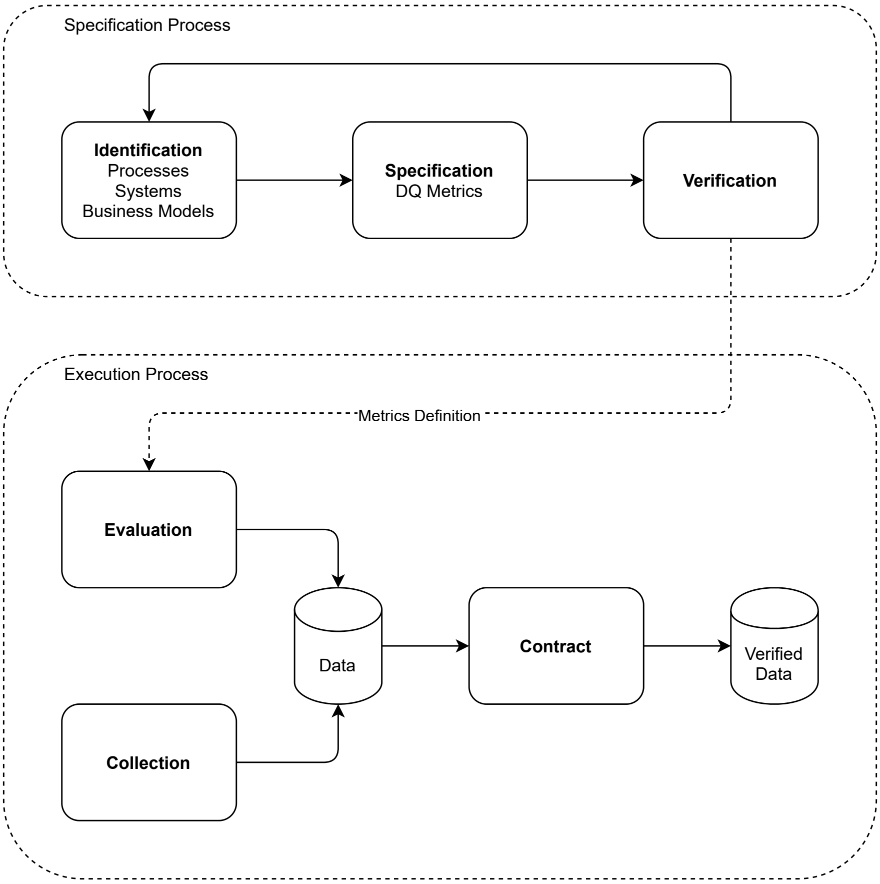


Figure 3.1: Methodology Metamodel

3.1.1 Specification Process

The specification process serves as a tool for defining qualitative and quantifiable quality requirements. This is a key part of the system. However, it is also the only part of the process that requires the necessary initiative of the analyst or analytical team. Now, we describe each part of the process.

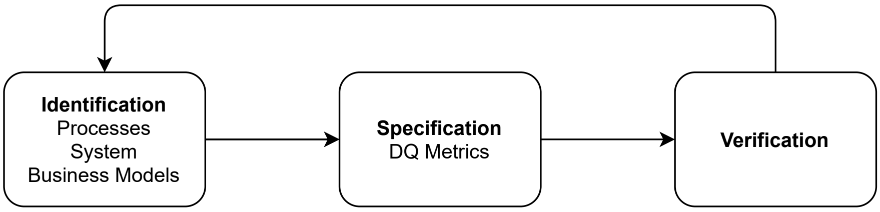


Figure 3.2: Specification Process

Identification

This activity focuses on identification of systems, processes and business schemes generatig data. By identifying weak points and bottlenecks in those processes, we can find causes of poor data. Also, we need to identify the subprocesses or activities that are mostly affected by the product data quality.

Metrics Specification

The goal of this activity is to identify the process metrics or KPIs. Measuring data quality is all about understanding what data quality attributes are, and choosing the correct data quality metrics. A comprehensive list of Data Quality Attributes by Eppler (2006) is available in appendix A. Specific attributes will be further discussed in Chapter 4.

Verification

The last part of current process is verification. This activity has to ensure that selected metrics are meaningful enough, capturing the actual condition of data.

3.1.2 Execution Process

The second main component is the execution process. This includes the actual collection and validation of data against the requirements obtained by the analysis from the first process. Ideally, in a semi-automated information system, this part runs independently, without human intervention. However, we are aware that in many cases it is not possible to implement a fully automated system, either due to the information complexity of the task or the financial costs of system development.

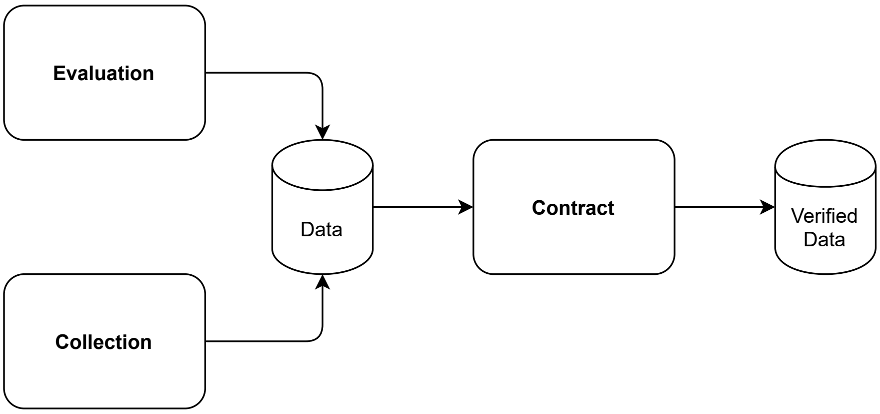


Figure 3.3: Execution Process

Collection

Data collection is a systematic process of gathering observations or measurements. Data collector can be either *Information System*, computer program or a human. Before the beginning of collecting data, we need to consider:

- the type of data we will collect;
- the methods and procedures we will use to collect, store and process data.

Verification

In our general case, verification is based on actual reliability of data, computed using DQ metrics. In other scenarios, the verification could be based on data redundancies, therefore based on the comparison of the collected data from two or more different collectors. If all data match, the data will be considered as valid. If not, the data remains invalid until a further collector validates it.

Artificial Intelligence and Machine Learning could be used to further ease and optimize data verification. Especially when processing image data and data with a high level of abstraction.

Contract

The contractual process is a subprocess that has the task of marking data as trustworthy if all the necessary requirements are met. This is the same concept as the so-called "smart contracts". Smart contracts are essentially blockchain programs that are processed when mandatory conditions are fulfilled. They are commonly used to simplify agreement implementation so that all parties can be sure of the result instantly, without intermediary intervention or time loss. This leads to workflow automation, initiating the next step if all conditions have been satisfied.

The contracts work by following simple "if-then" statements. This mechanism might include allocation of funds to the appropriate parties, sending notifications, or releasing a ticket.

3.2 Supporting Techniques

There are a several data quality rules one can deduce from a Feedback-Control Systems view of information systems reviewed by Orr (1998):

1. unused data cannot remain correct for very long;
2. data quality in an information system is a function of its use, not its collection;
3. data quality cannot be better than its most strict use;
4. data quality problems tend to become worse as the system ages;
5. the less likely some data attribute is to change, the harder it will be to change it when the time comes;
6. laws of data quality apply equally to data and metadata [28].

To prevent the consequences of these rules and the unauthorized creation of data, we present two additianal concepts. These concepts should be incorporated into the design of the information system respecting the proposed methodology.

3.2.1 Proof of Constancy

Proof of Constant Data, alias Proof of Constancy, is a way to assure a constant accuracy of data [13]. Data have to be regularly updated to keep the accuracy rate high. Data accuracy rate will decrease progressively based on a specific time frame basis (e.g., X% per month) [13]. This percentage is different depending on the type of data. Datasets more sensitive to changes may see this rate decrease by 5% to 10% per month or day depending on the circumstances [13]. On the other hand, established, well-known sets, will see their rate decrease by 0.1% per month or even year. A scale of discount rates will have to be established based on the areas of interest and actual items collected.

3.2.2 Proof of Trust

Proof of Trust is an instrument for data collector evaluation [13]. The collector or generator will get 'quality score' for his/her or its collection actions [13]. The more collectors initiate, update and verify data correctly, the higher their 'quality score' will be [13]. A higher quality score leads to a higher level of 'trust'. Incorrect collection, on the other hand, results in a retroactive decrease in the collector's quality score [13].

3.3 Use Cases

In this part, we will present several use cases to illustrate versatile use of the presented framework.

3.3.1 Enterprise Information System

Enterprises suffer from poor data quality. We propose, following the methodology, to introduce a central register of data sources. This central register should be supported by a set of services and a central data repository.

After a thorough analysis of data requirements and their quality, a defined set of metrics and key performance indicators parameterizes the verification chain of activities. If the predefined quality limit is not met, the data will either be rejected or saved with an error flag. If the data meets the required level of error, they go through the contracting process and are considered as a reference until their latest version is qualitatively degraded by the ordered process (e.g., Proof of Constancy) and marked as untrusted.

A penalty for poor quality would be automatic reporting to the company's senior management. Management could then impose sanctions on those responsible for specific datasets and data flows in the form of reductions or cancellations of personal rewards.

3.3.2 IoT Cluster

Based on the domain and usage of the IoT devices, the data repository could be either centralized (e.g., nuclear power plant cluster of secondary senzors) or decentralized (e.g., community weather stations).

The verification algorithm would - in this case - consist from two general authorities. The first authority being k nearest neighbours of the same sensors (or IoT devices in general), and the second one being the set of domain rules. Nearest neighbors provide redundancy by which data can be verified. And, of course, the data itself must meet the criteria restrictions set by the domain of use.

Poor quality would reduce the importance of the sensor in the cluster, or its temporary or complete decommissioning. This system would also create a very effective defense barrier against attacks, especially against data poisoning.

Data poisoning is a class of attacks on machine learning algorithm where an adversary alters a fraction of the training data in order to impair the intended function of the system. Objective can be to degrade the overall accuracy of the trained classifier, escaping security detection or to favor one product over the another. Machine Learning systems are usually retrained after deployment to adapt to changes in input distribution, so data poisoning represents serious danger.

Qualitative degradation of data by Proof of Constancy would not be the so important, because we expect very high update frequency. However, lower update frequancy of IoT device would suggest an error within a system, which could serve as a warning to network operators about a faulty device. Data from defective equipment should also not be taken into account in many cases.

3.3.3 Open Data Library

The final use case demonstrates the use of a completely decentralized solution. The system would reward those who collect and generate data, and the data would be available for use in a decentralized marketplace. The decentralized network would democratize data access while rewarding data creators [13].

Data would be collected using an application (system) that would be used by a community of collectors who would be rewarded for their efforts. The reward should be determined by a ‘collection value’.

The collection value would be calculated using an algorithm that considers a number of factors, including:

- demand and rarity,
- availability and accessibility,
- data licensing and
- market value [13].

To maintain a high level of dependability, each collector receives a quality score. The verified data is then made available (via contract) on the decentralized marketplace and is updated on a regular basis to ensure its accuracy.

Automation of the verification process is nearly impossible due to the variety of open data. The data must be manually verified. As a result, a collector serves two purposes:

- initiate the data collection (input and update data),
- verify the collected data (check a collected data not yet verified) [13].

The process guarantees that the reward for data collection is split equally between the collector and the verifier [13]. For example, the collector who initiated the data collection would receive 60-80% of the reward [13]. The remaining 20-40% would be obtained by the verifier [13].

Depending on the data, a decentralized network like filecoin could be used as data storage. The use of blockchain renders the data unalterable, ensuring the transparency and traceability of the validation process (collection, verification, update). The blockchain (which is tamper-proof, immutable, and decentralized) ensures the integrity and verification of the data on the marketplace. This gives data users confidence and security. The use of **Smart Contract** technology also ensures the collectors' rewards.

Chapter 4

Quality Classification System

The original idea was to leverage some Machine Learning classification algorithm to automatically classify datasets. During thesis elaboration the referential materials turned out to be insufficient in providing usefull information on the topic, hence different technique was chosen (composite score-card evaluation). Shortcoming of white-papers about Machine Learning supported DQ classification probably results from the absence of well-defined general DQA algorithm and output classes. Complexity of developing all-embracing method for DQA competes with unfolding general artificial intelligence, indeed.

4.1 Data Quality Dimensions

In order to provide an objective way to measure data quality, we have to choose some DQ metrics and define formal way to compute them. The list of all candidates can be seen in the Figure A.1. Many of these candidates are very suitable for specific cases, but completely inappropriate for general use. After narrowing the selection due to general applicability, we get this list of metrics: completeness, uniqueness, timeliness, validity, accuracy and consistency [34].

4.1.1 Uniqueness

Uniqueness indicates that each data record should be unique, or else the risk of accessing obsolete information rises. Just one instance of each real-world object should be recorded in a dataset. We may have two rows with objects “John Doe” and “Jonathan Doe”, who are the same person, but the latter has the most up-to-date information. Any metrics involving those object instances (e.g., customer count, average spend per customer and sales frequency) would return incorrect results. Identifying a suitable primary key is the first step in resolving this issue. In the example, having different names and Customer IDs, but matching email addresses is a good indicator that they are in fact the same individual. This means that before any analysis or modeling, an additional phase of data inspection is required to consolidate these records.

Leveraging information theory enables us to move the idea forward. The supporting method for calculating uniqueness could be the calculation of entropy for each record and comparison through distance statistics [35]. For each key, we could compute the Shannon entropy H of the values. The higher the entropy, the more diverse the key's values are. Entities with similar entropy are likely to be the same objects.

4.1.2 Validity

Validity is a quality dimension that refers to information that does not obey business standards or conforms to a particular format. For example, surname must be a sequence of alphabetic characters and telephone numbers must be composed of numeric characters and must comply with specific regional rules. Regular expressions can be used to check for validity in a variety of contexts. Databases containing regular expressions for many common data types are available online. For discrete data types, simple frequency statistics can tell whether there is a validity issue (e.g., school grades data type with more than 4-5 elements). It basically becomes a completeness problem once invalid data is found.

4.1.3 Accuracy

Accuracy shows how reliable the data reflect the object or event in the real world. For example, if a temperature in the room is 21°C, but the thermometer says it is 25°C, that information is inaccurate. Probably the easiest way to improve accuracy is to introduce redundancy into the system. An additional check of the acquired data will help to identify discrepancies before entering the system.

4.1.4 Completeness

Blake and Mangiameli (2011) defined completeness as follows. On the level of data values, a data value is incomplete (i.e., the metric value is zero) if and only if it is 'NULL', otherwise it is complete (i.e., the metric value is one). A tuple in a relation is defined as complete if all data values are complete (i.e., none of its data values is 'NULL'). For a relation R, let T_R be the number of tuples in R which have at least one 'NULL'-value and let N_R be the total number of tuples in R. Then, the completeness C of R is defined as follows [6].

$$C = 1 - \frac{T_R}{N_R} = \frac{N_R - T_R}{N_R} \tag{4.1}$$

This definition of *completeness* meets the requirements for metrics according to Heinrich et al. (2018). The metric values are within the bounded interval $[0;1]$ for all aggregation levels. The minimum value represents perfectly poor data quality and vice versa. To archieve full score, no tuple must not contain a null value, as well as relations must not contain any tuple with data values which equal 'NULL'.

The metric is reliable because all configuration parameters of the metric can be determined by a database query. Due to the existence of a mathematical formula, the metric is objective and because the metric quantifies a dimension at all quality levels according to the corresponding definition, the determination of the metric value is also valid. The metric formula is applicable to single data values as well as to sets of data values.

4.1.5 Consistency

There are several forms of data consistency. The **first form** is actual wide or narrow distribution of data. In this way, consistency of data can be viewed as *stability*, *uniformity* or *constancy.*

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Figure 4.1: Population Standard Deviation formula

Typical measures include statistics such as the *range* (i.e., the largest value minus the smallest value among a distribution of data), the *variance* (i.e., the sum of the squared deviations of each value in a distribution from the mean value in a distribution divided by the number of values in a distribution) and the *standard deviation* (i.e., the square root of the variance).

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Figure 4.2: Sample Standard Deviation formula

The *standard error of the mean* (i.e., the standard deviation of the sampled population divided by the square root of the sample size) is frequently examined when evaluating the consistency of data drawn in a sample from a population. Finally, the constancy of data produced by instruments and tests is typically measured by estimating the reliability of obtained scores. Reliability estimates include test-retest coefficients, split-half measures and Kuder-Richardson Formula №20 indexes [39]. For Time Series data, stationary analysis can be done. If the data is non-stationary then it is likely to have some degree of inconsistency.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Figure 4.3: Standard Error of the Mean formula

Then there is the **second form** of data consistency, which is whether the data are uniformly defined across the dataset, that is, across variables and over time. For example, suppose we want to use the data to estimate real estate sales per year to see how that number has changed over time. In this case, we have to make sure the estimates of real estate sales are uniformly defined over time. Specifically, does the data series always either include apartments or exclude apartments from the counts? Does it always either include houses or exclude houses from the counts? If the data sometimes include apartments, but not always, or if the data sometimes include houses, but not always, then the data are inconsistent.

The **third form** of consistency tightly coupled with relational databases and their referential integrity. A relational database is said to be ACID (vs non-relational BASE), meaning (i) atomicity, (ii) consistency, (iii) isolation and (iv) durability. The term onsistency there refers to the requirement that any given database transaction must affect data only in allowed ways, therefore data must be valid according to all defined rules, including constraints, cascades, triggers, and any combination thereof.

Inconsistencies in data can be due to changes over time and/or across variables for example, in (i) vintages or time periods, (ii) units, (iii) levels of accuracy, (iv) levels of completeness, (v) inclusions and exclusions. Those inconsistencies occur most often when merging or aggregating datasets, therefore the user has to make sure data are consistently defined throughout.

4.1.6 Timeliness

Timeliness is another one of the major dimensions in the field of data quality. Obsolete data suppress innovation, therefore businesses and startups want to trust the data publisher that the data will remain available and relevant, especially when using open data or reference data from central registers. A measure of timeliness has to focus on the update cycle. Automation must be a key part of this process, leading to the efficiency in publishing and processing of data. Meeting all these points is a necessary, but not sufficient condition to create a sustainable data ecosystem.

Atz (2014) proposed an unique metric for measuring the timeliness of data. The research defines timely dataset as a product of function of the forecast update frequency (a dataset released annualy will be updated only once a year) [2]. The concept of timeliness T can be expressed by the equation 4.2.

$$T = I \frac{f_U}{today - lastupdate} \tag{4.2}$$

In the equation 4.2, the I is an indicator function causing *Heaviside* step function effect returning 1 when the ratio is greater than one and 0 otherwise. For example, a dataset with a daily cycle and a last major update last month would result in a 0. On the other hand, a dataset with monthly cycle and an update in last two weeks, would yield 1. In the equation f_U represents *update frequency*; the terms *today* and *last update* are timepoints corresponding to the names.

The reason for the presence of the indicator function is that we do not have a tool to evaluate the aging of the dataset. Data can become obsolete linearly and continuously, but also non-linearly and discontinuously. This functional dependence is hidden from us, so we consider the data to be either current or obsolete.

Atz (2014) introduced a metric for measuring data catalogue timeliness, τ (equation 4.3). In general, minor changes such as typo correction should not be considered as an update.

$$\tau = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{f_{U_i} \cdot \lambda + \delta}{today - lastupdate_i}\right) \tag{4.3}$$

The τ of a data catalogue is the average across datasets, indicated by the subscript i [2]. The number of datasets in the catalog is denoted by the N.

Two parameters in a linear form have been introduced to the core of the expression. The lambda (λ) is degree of freedom relative to the update frequency; the days we allow the update of the data catalog. For example, considering 5% of the time reserve (e.g., due to ETL delays), the annual renewal dataset is going have a buffer of 0.6 months and for a monthly dataset it implies 1.5 days in tolerance. The delta (δ) is a fixed number of days applicable to all datasets, for example one day for processing [2].

τ	Data Timeliness
0.9-1	exemplar
0.7-0.9	standard
0.5-0.7	ok
0.25-0.5	poor
0-0.25	obsolete

Table 4.1: Proposed benchmarks for different levels of τ [2]

The trivial case (only one dataset in the catalog) is constrained, by design, to a binary classification, data are either up-to-date or not. This means that a data catalogue that is one day late is considered the same as one that fails to fully update the datasets. However, the advantages of simplicity outweigh the disadvantages of a more complex method.

4.2 Weighted Aggregation

Probably the best known project dealing with data classification is “5-star Open Data”. This classification system defines quality in terms of how well they provide the context in which the data is located as well as in how well machine-readable the data is [36]. The highest quality data are those that have a fully defined ontology and are connected to other datasets [1]. The RDF schema and the SPARQL language are used for this case [1]. Automatic classification of a dataset can be performed by testing the existence of hyperlinks. The use of this framework is in the context of the Internet, so it is not very suitable for our purpose.

Considering that we want to achieve the relative objectivity of the framework, we will have to use the general metrics discussed in the previous section. The resulting system should therefore be similar to machine learning ensemble voting. A Regression Voting Ensemble is a machine learning model that combines the average predictions of contributing models to improve model performance. Model is shown in the picture 4.4.

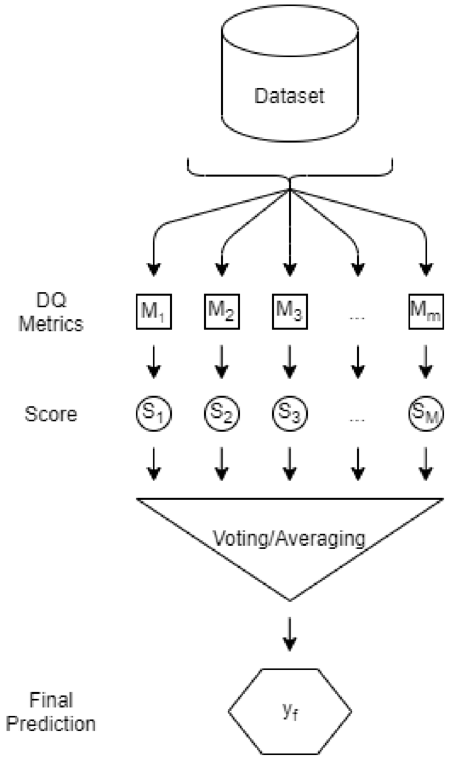


Figure 4.4: Voting Model

The model can be expressed using the equation 4.4.

$$y_f = \sum_{i=1}^{N} w_i y_i \tag{4.4}$$

In the following chapter, we will use this concept as a base to evaluate the performance of the data catalog.

4.3 Data Quality Framework Evaluation Instrument

The Data Quality Framework Evaluation Instrument is a tool developed by Long et al. (2004). The instrument's primary goal is to promote continuous data quality enhancement and the user limitations tracking [25]. The instrument is structured as a four-level conceptual framework, with 88 requirements at its root [25]. Using the instrument algorithm, the criteria can be folded into the second level of 24 data quality characteristics which can then be rolled up into 6 data quality dimensions [25]. The five dimensions can be combined into a single overall data collection evaluation [25].

The instrument is designed for the evaluation of medical datasets, we will try to use it unchanged for COVID-19 datasets.

4.3.1 Instrument Usage

Each of the 88 criteria is meant to be scored as “unknown (1)”, “not met (2)”, or “met (3)” [25]. Each score must be briefly substantiated in writing. All dimensions and characteristics are scored as “unknown (1)”, “not acceptable (2)”, “marginal (3)”, and “appropriate (4)” [25].

Long et al. (2004) proposed the following procedure for aggregating criteria into characteristics.

1. If one of the criteria within a characteristic is “unknown” then the characteristic is scored as “unknown” [25].
2. If the status of all the criteria is known and more than half are “not acceptable” then the characteristic is “not acceptable” [25].
3. If the status of all the criteria is known and half are “met” then the characteristic is scored as “marginal” [25].
4. If all the criteria are “met” then the characteristic is “appropriate” [25].

The procedure for characteristic-dimensions aggregation is as follows.

1. If one of the characteristics within a dimension is “unknown” then the dimension is scored as “unknown” [25].
2. If the status of all the characteristics is known and at least one is “not acceptable” then the dimension is “not acceptable” [25].
3. If the status of all the characteristics is known and they are a combination of “marginal” and “appropriate” then the dimension is scored as “marginal” [25].
4. If all the characteristics are “appropriate”, then the dimension is “appropriate” [25].

The algorithm proposed by the authors of the article is suitable for datasets which we have complete information about. However, this will not be our case. There will be too many questions that we cannot answer directly in the questionnaire. This would probably lead to a final evaluation of the dimensions as unknown. The algorithm also assumes that all dimensions, characteristics, and criteria have the same weight, which generally may not be the case.

For this reason, we propose to adjust the score scale and modify the algorithm. Newly, the criteria will be evaluated as “unknown (0)”, “not met (2)”, or “met (3)”. The dimensions and characteristics will be evaluated as “unknown (0)”, “not acceptable (1)”, “marginal (2)”, and “appropriate (3)” The aggregation algorithm will be simplified by averaging the values and then rounding to the nearest score on the scale.

1. Evaluate all criteria within characteristics.
2. Average the rating of the criteria within characteristics and round the result to the nearest integer.
3. Repeat the procedure for aggregation at the dimension level.

By stretching the criteria scores, we are able to use pure mathematical tools rather than analytical ones. Weighing of dimensions and characteristics can be added to the proposed procedure. In this way, we obtain a model that can express the diversity of datasets.

The advantage of the presented approach is that it can be used both for individual datasets and for the data catalog. The disadvantage of the approach is that the evaluation of criteria in a more specialized field will require the cooperation of industry and IT specialists. The main limitation is that the questions in the questionnaire provide only a vague idea of the state of the data in the database (e.g., integrity). On the other hand, they provide a comprehensive overview of the status of data, including the process of data collection and staff training. Fortunately, this condition is suitable for the COVID-19 dataset.

Additional criteria can be added to the questionnaire. However, such an intervention requires a field specialist and subsequent testing of the evaluation process.

Chapter 5

Case Study

In December 2019, a virus known as COVID-19 was first identified in Wuhan, China [23]. Three months later, on March 1, 2020, the first three cases of the disease were confirmed in the Czech Republic [23]. The disease has shown and continues to show the shortcomings of social and political environment worldwide. But the disease, although very serious, has given us many opportunities. One such opportunity is open datasets made available by state institutions.

In the following part of the work we will try to analyze the state of datasets provided by the institutions of the Czech Republic. We apply the metrics defined in Chapter 3 to the data in order to objectively measure their quality and comment on the results. Available datasets as of April 1, 2020 from the URL address https://onemocneni-aktualne.mzcr.cz/ are listed in Appendix B.

5.1 Technical Analysis

The data can be downloaded via the REST API in JSON, in addition, the data can be downloaded in CSV together with metadata also in JSON format. The format of the JSON data file can be seen in Figure 5.1.

```
{
    "modified": "2021-04-18T12:28:42+02:00",
    "source": "https:\/\/onemocneni-aktualne.mzcr.cz\/",
    "data": [
        {
            ...
        }
    ]
}
```

Figure 5.1: Data file structure

All data files contain one main object, with three keys (modified, source and data). The “modified” key contains the date and time of the dataset update in ISO 8601 format, offset from UTC (Coordinated Universal Time). The “source” key contains the dataset’s URL, especially the protocol and domain name. The last key, “data”, contains an array of json objects with a structure specified by the metadata. This is an array of objects even if the array contains only one object. Figure 5.2 depicts a sample structure of the objects in the list.

```
[
    {
    "datum": "2020-01-01",
    "vek": 50,
    "pohlavi": "M",
    "kraj_nuts_kod": "CZ000",
    "okres_lau_kod": "CZ0000"
    },...
]
```

Figure 5.2: Sample data with information on deceased patients

An example of the structure of the metadata file can be found in the appendix C.

5.2 Data Quality Criteria Analysis

Long et al. (2004) wrote a comprehesive data quality checklist for emergency medical services data [25]. We will use this checklist as a base for our own data evaluation. The OECD's Health Care Quality Framework serves as the foundation for the questionnaire. The OECD identifies an unique set of dimensions that differs from the set we defined in Section 4.1 as dimensions suitable for general use. However, this is insignificant because the defined dimensions are still appropriate for the quantitative determination of quality, from which we were forced to shift due to the nature of the data to the qualitative determination of quality.

5.2.1 Relevance

Relevance is the degree of correspondence between the data's content and the user's areas of interest [32]. In other words, the extent to which data answers provide insight into the individual user's question [32].

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Relevance	Adaptability	3	3	3	3	3	0	3	3		
	Value	3	0								

Table 5.1: Evaluation of criteria for dimension Relevance

Value

The characteristic is about how valuable the data is to its users and whether or not user requests and comments are taken into account. Although evaluation is highly subjective, it is crucial. There is no point in publishing data if it does not meet the expectations of users.

a. **The purpose of the data is clear.**

   Yes, the data's overall purpose is clearly evident. The data is provided by the Institute of Health Information and Statistics of the Czech Republic (IHIS CR). Individual datasets have a description that specifies their purpose.

b. **The data shed light on the issues of most importance to users and the data are used in policy formulation and decision making.**

   Yes, the data provide information on the current development of the epidemic situation in the Czech Republic. The dynamics of the *Compartmental Model* can be derived from the data and thus predictive analysis can be done. The data are used for policy formulation and decision making.

c. **There are no other more valid sources for the data.**

   Yes, the data is provided by a government agency. There is no other source of data for the public.

d. **Client liaison mechanisms are in place and client needs are monitored.**

   Yes, IHIS CR does have mechanisms in place for contact with clients (usually in the form of press conferences). Client needs are monitored in some ways, as IHIS expands the data catalog in response to these needs [10].

e. **How the data are used is known and well understood.**

   The Institute of Health Information and Statistics of the Czech Republic publishes summaries of the epidemiological situation and other reports, including descriptions of methodologies for using data for predictions. Formally, this point is met.

f. **Client evaluations are conducted and reviewed.**

   We were unable to find sufficient information to answer this question.

g. **The data are found to meet the needs of its users.**

   In this case, as "needs of the users" is considered the awareness of the general public. From this point of view, the question is met.

h. **The data are found to be worth the resources dedicated to its production.**

   Yes, these data are necessary to inform the public about the pandemic's current state of development. In this case, almost any expenditure is worthwhile.

Adaptability

Adaptability describes how the data and the team behind it can respond to changes in the external environment.

a. **The data can be used to inform emerging issues and can adapt to change.**

Yes, this has been demonstrated in the past by the introduction of anti-epidemic measures and the changes that have been introduced in the datasets.

b. **Ongoing explicit program review and priority determination are conducted.**

The degree of review by the data issuer is **unknown**.

5.2.2 Accuracy

The degree to which data accurately estimate or describe the quantities or characteristics that they are designed to measure [27].

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Accuracy	Form Design and Completion	3	3	3	2	0	0	0	2	2	0
	Frame	3	3								
	Over Coverage	0	0								
	Under Coverage	0	2								
	Response	2	3								
	Completeness	0	0	0	0						
	Bias	0	2	0	0	0	0				
	Validity	3	0	0	2						
	Reliability	0	2	0							
	Collection	0	0	3	0	0	0	2	3	2	2
	Processing	0	0	2	0						
	Imputation	0	3								
	Analysis	0	0	0	2						

Table 5.2: Evaluation of criteria for dimension Accuracy

Form Design and Completion

How data collection forms and principles are designed, what information is provided with the forms, and how employees are trained.

a. **The data retrieval form was designed by a team that includes methodologists, a form design expert, representatives for those who are responsible for completing the form, as well as other subject matter experts.**

Data are issued by IHIS CR. The criterion is **met**.

b. **The purpose and population of interest are clear and well documented.**

The purpose of the data is clear. The population to be monitored is precisely defined, individuals infected with COVID-19 and individuals tested or vaccinated.

c. **There is adequate justification for each field gathered.**

The accessible data is preprocessed and therefore each field in the dataset has a purpose. The purpose may not be obvious at first glance, but all fields are defined and annotated in the metadata. We consider this criterion to be **met**.

d. **The form is user-friendly and is accompanied by a clear, readily accessible, and user-friendly manual that describes in detail the data collection guidelines including when and how to complete the form and defines each field in detail.**

Data is available through a well-defined API. Although instructions for data collection are either missing or very difficult to find. As this absolutely essential requirement is violated, we consider the criterion to be **not met**.

e. **Those responsible for completing the form receive training so that they are able to properly complete the form.**

Unknown.

f. **As part of training, the importance of the data is conveyed and those responsible for completing the form are tested and immediate feedback is provided regarding the reliability and validity of their performance.**

Unknown.

g. **Those responsible for completing the form are allocated the time and have the motivation to do so, as well as confidence in the form completion process.**

Unknown.

h. **The data are monitored for outliers, logical errors, completeness, and consistency and ongoing monitoring and constructive feedback is provided to the primary collectors where necessary.**

Data monitoring information is unknown. But the data contains logical errors. Feedback to primary collectors (hospitals, medical facilities and sanitation stations) is very difficult to implement. We consider the criterion to be **not met**.

i. **Any major revisions to the original form design (purpose, structure, etc.) of the database and the dates of any major revisions are known, documented, and readily available. Moreover, an explicit consideration of overall trade-offs between accuracy, cost, timeliness, and respondent burden was conducted at the design stage.**

Information on the design phase is unknown. To our knowledge, changes to the API design are not available to the public, if documented at all. For this reason, we consider the criterion to be **not met**.

j. **A revised form is pilot tested until high standards of reliability and validity are met and the pilot test results are readily available.**

Unknown.

Logical Errors One of the most serious errors in the data catalog is the absence of categorization of deaths. In the dataset, which contains data on patients' deaths, we learn the date of death, the patient's age, gender and the codes of the region and municipality from which the patient came. It does not take into account whether the patient was very ill (e.g., a patient with a stage 4 pancreatic cancer with primary tumor and metastasis outside pancreas) and therefore had a high probability of death, or if the patient was a completely healthy person. No consideration of this fact leads to the introduction of virtually immeasurable systematic (logical) error.

A similar issue can be observed in the data on new COVID-19 cases (Figure 5.3) and the daily number of tests performed. In the first case we are unable to determine whether there are any percentage of patients in whom the disease manifested itself repeatedly. In the second case, information on retested patients is missing, as well as information on whether the tested patient showed symptoms of the disease. All these problems lead to a reduction in data quality, which we are not able to measure effectively.

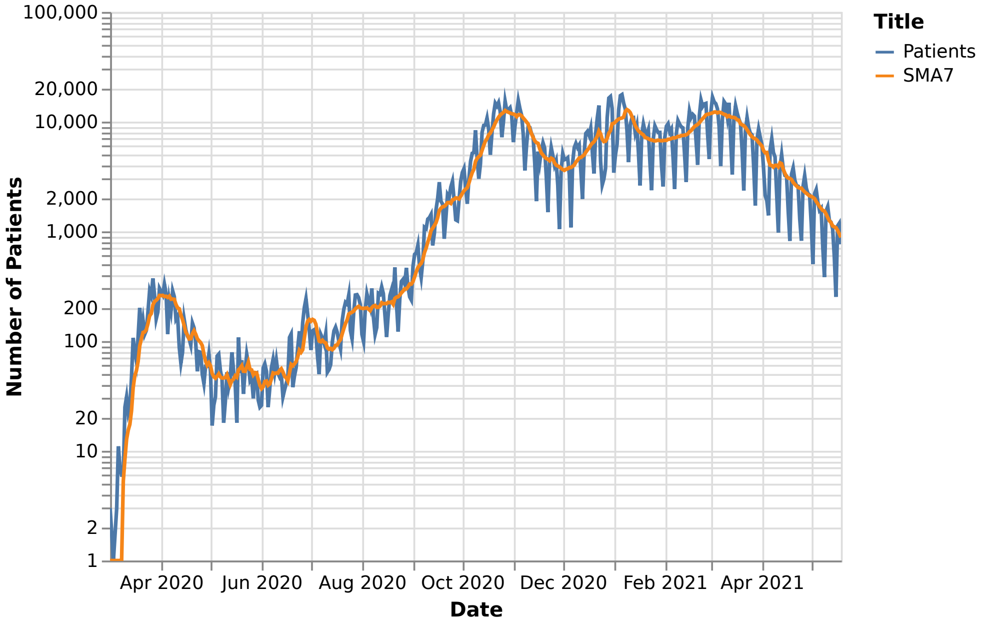


Figure 5.3: Daily confirmed cases of infection and weekly moving average.

Frame

By the "frame" it is meant the structure of files containing data. The COVID-19 API is well documented and easy to understand. During the epidemic, the API was shown to be continuously updated.

a. **The frame is known and documented.**

   The criterion is **met**.

b. **The frame is maintained in an ongoing manner.**

The criterion is **met**.

Over Coverage

Over-coverage is caused by the existence of objects not belonging to the frame, as well as units that appear in the target population multiple times [16].

a. **Only qualifying data suppliers are on the frame.**

Information on data suppliers is not available. The criterion is **unknown**.

b. **The data are checked for duplicates and erroneous entries from qualifying suppliers.**

Unknown.

Under Coverage

Over-coverage is caused by exclussion of objects belonging to the frame [17].

a. **Data are received from all qualifying data suppliers on the frame.**

Unknown.

b. **The list of those actually sending data are compared to independent lists.**

Given the nature of the data, the existence of third party lists can be questioned. We therefore consider the criterion not to be met.

Response

System response during data deliveries. Check that the data is complete and that a critical component of the data is not damaged.

a. **The overall expected and actually received number of records are known and tracked per year and response across month is checked and compared against previous years.**

This criterion is in principle difficult to meet. The number of records is not known in advance and, due to the relatively short existence of the dataset, it was not possible until recently to compare the data with previous time periods. We consider the criterion to be **not met**, although the question remains whether it is even valid for our case.

b. **The amount of missing data per record is known and tracked per field per year and key fields (e.g., age, gender, and clinical code) are at least 98% complete.**

Key data has been remeasured, the criterion is **met**.

Completeness

Data completeness in terms of creation. In this case, in terms of data collection from patients, which is not entirely relevant to our data because we do not have such information and cannot obtain it.

a. **The patient service form/chart used for data retrieval from abstraction is easy to understand.**

   There is no information available on internal procedures.

b. **The form/chart used for abstraction is complete.**

   There is no information available on internal procedures.

c. **All patient encounters/visits are abstracted and represented in the database.**

   There is no information available on internal procedures.

d. **All fields are systematically completed per patient record.**

   There is no information available on internal procedures.

Bias

The section of the questionnaire that checks for points that may prevent bias from being introduced.

a. **Explicit standard guidelines are in place and adherence is monitored for data collection.**

   Information on internal procedures is **not known**.

b. **Clear guidelines and training eliminate as much as possible the need for interpretation.**

   The IHIS CR does not provide data training to external users, and we do not have information on internal employee training. The data are clear, but due to their sensitivity, they should be interpreted. As a result, we consider the criterion to be unsatisfactory.

c. **For data that need to be classified, clear coding standards are available.**

   Information on internal procedures is **not known**.

d. **For data that need to be classified, the available standards are adhered to.**

   Information on internal procedures is **not known**.

e. **For data that need to be classified, only highly trained certified staff classify the data.**

   Information on internal procedures is **not known**.

f. **Sources of bias (e.g., upcoding) are understood and eliminated if possible and ongoing quality assurance tests ensure that data collection, abstraction, and entry are conducted in a standard manner according to guidelines.**

   Information on internal procedures is **not known**.

Validity

What are the processes in place on the data supplier's end to ensure the data's validity?

a. **The patient service form/chart is complete and reflects the patient encounter and the codesheet or abstract that is based on the form/chart reflects what is in the form/chart.**

   Since we do not work with complete input data, this question is irrelevant. However, we have a dataset with patient data (age, gender, residence) that directly corresponds to the input data. For this reason, we consider the requirement to be **met**.

b. **Adequate resources are in place to ensure valid timely data and ongoing database improvement.**

   Information on internal procedures is **not known**.

c. **Random audits and/or reabstraction studies are conducted and the data are compared to external sources of the same or similar data (if possible).**

   Information on internal procedures is **not known**.

d. **Validity coefficients are available and are greater than or equal to 0.8 for key data elements (i.e., postal code, patient age, most responsible diagnosis, procedures, and comorbities).**

   The IHIS CR itself does not provide any validity metrics. This requirement is **not met**.

Reliability

The level of agreement between repeated administrations of a diagnostic test by evaluators [20].

a. **Reliability studies of key data elements (e.g., age, gender, and clinical code) are conducted at regular intervals.**

   Information on internal procedures is **not known**.

b. **Intra rater coefficients are available.**

   The IHIS CR itself does not provide any reliability metrics. This requirement is **not met**.

c. **Evaluation coefficients for key data elements are greater than or equal to 0.8. (i.e., postal code, patient age, most responsible diagnosis, procedures, and comorbities).**

   Intra-rater coefficients are not available. The criterion is **not known**.

Collection

Information on the data collection processes and tools used.

a. **Standard data retrieval form is in place.**

Information on internal procedures is **not known**.

b. **Range checks are place for all fields at data entry and key logic checks are run (e.g., checks for clinical impossibilities or date of birth greater than call date).**

Information on internal procedures is **not known**.

c. **Standard data specifications are provided to vendor(s).**

The institute provides a dictionary with metadata. This requirement is **met**.

d. **Standard test data are used to test edits.**

Information on internal procedures is **not known**.

e. **Data entry software and equipment are user friendly.**

Information on internal software and hardware is **not known**.

f. **Staff is available and motivated to enter the data and data entry is monitored and constructive feedback is provide to staff.**

Information on internal procedures is **not known**.

g. **Edit errors are set aside and made available for analysis.**

The institute does not provide any dataset regarding data errors. This requirement is **not met**.

h. **Data entry of abstracted data takes place in close proximity to the original data (original forms/charts).**

Although no information is available about internal processes, the data is prone to factual accuracy. Let us consider this requirement to be **met**.

i. **Error detection reports are generated.**

The institute does not provide any documentation regarding data errors. This requirement is **not met**.

j. **Error correction is documented.**

The institute does not provide any documentation regarding data errors. This requirement is **not met**.

Processing

Information on transformation processes, how they are used and tested.

a. **All programming is tested and the results are documented.**

Information on internal procedures is **not known**.

b. **Ongoing quality control checks are conducted on electronically extracted data.**

Information on internal procedures is **not known**.

c. **Documentation on how the various systems involved interact, extract, change, and/or append the data exists and is available.**

The information mentioned in the request is not known to the public. This requirement is **not met**.

d. **Ongoing tests are run to ensure all systems are interacting properly.**

Information on internal procedures is **not known**.

Imputation

The replacement of approximate values for incomplete or conflicting data elements is known as data imputation [15]. The substituted values are meant to produce a data record that is not subject to edit failure [15].

a. **Imputation is automatically derived.**

Information on internal procedures is **not known**.

b. **The raw data are preserved.**

Although information on internal processes is not available, it can be assumed with great certainty that the original data is not modified in any way. Let us consider this requirement to be **met**.

Analysis

Is there any information available about outliers and data errors?

a. **Edit errors are analyzed.**

Information on internal procedures is **not known**.

b. **Error detection analyses are conducted and the data are checked for missing data.**

Information on internal procedures is **not known**.

c. **Outliers or other suspicious data are investigated.**

Information on internal procedures is **not known**.

d. **Regular standard summary analyses are conducted and made available.**

Although we have summary data in the form of a graphical report, it is not a statistical report summarizing the qualitative status of the data. For this reason, we consider the requirement to be **not met**.

5.2.3 Timeliness

This aspect is already covered in Chapter 4. Although we have described its quantitative features and are now dealing with its qualitative ones, there is significant overlap between them, and some of the questions in the questionnaire could be quantified.

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Timeliness	Data Currency	3	3	3	3						
	System Efficiency	0	0								

Table 5.3: Evaluation of criteria for dimension Timeliness

Data Currency

Are data updates delivered on time and on a consistent basis?

a. The time between original form or chart completion and data abstraction is reasonably brief.

Most datasets are updated daily, with values from the previous day. We consider this requirement to be **met**.

b. The time between the end of the reference period to which the data pertain and data release is reasonably brief.

Most datasets are updated daily, with values from the previous day. We consider this requirement to be **met**.

c. The official date of release was announced in advance of the release.

The datasets are published at approximately the same time, with a given update frequency. We consider this requirement to be **met**.

d. The official date of release was achieved.

The data collection is continuously updated. We consider this requirement to be **met**.

System Efficiency

Information on the effectiveness of database and distribution processes.

a. Database methods are regularly reviewed for efficiency.

Information on internal procedures is **not known**.

b. Processing methods are regularly reviewed for efficiency.

Information on internal procedures is **not known**.

5.2.4 Accessibility

The ease with which data products can be located and accessed within data holdings is reflected in Accessibility [27].

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Accessibility	Awareness	3	3								
	Ease of Access	3	3								

Table 5.4: Evaluation of criteria for dimension Accessibility

Awareness

Data awareness is about providing data visibility and intelligence about the content, users, and activity [37].

a. **The existence of the data can be ascertained.**

   Yes, the public cannot look at reference data sources, but patients certainly exist, so data about them also exist. We consider this requirement to be **met**.

b. **Standard tables and analyses are produced and made available per reference period.**

   Data (and therefore visualizations) are updated daily. We consider this requirement to be **met**.

Ease of Access

The ease with which data can be accessed.

a. **The data are well organized and readily available for users.**

   The data is published in CSV and JSON files. The structure of data files is known, unchanging and well defined. We consider this requirement to be **met**.

b. **Privacy and confidentiality rules related to accessibility are adhered to.**

   The data is anonymized and most of them are free to access. To access the non-public part of the data sources, it is necessary to fill in an official application. We consider this requirement to be **met**.

5.2.5 Interpretability

The ease with which users can understand, use, and analyze data is reflected in interpretability [27]. The degree of interpretability is largely determined by the adequacy of the definitions of concepts, target populations, variables, and terminology underlying the data, as well as information describing the data's limitations, if any [27].

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Interpretability	Documentation	2	3								
	Education	3	2								

Table 5.5: Evaluation of criteria for dimension Interpretability

Documentation

Level of data documentation, metadata and its comprehensibility for the public.

a. **The limitations of the data are documented for users using a standard format and the documentation is readily available for users.**

   We have been unable to discover any documentation that satisfies the requirement. This criterion has not been fulfilled.

b. **The supplementary information and metadata necessary to interpret and utilize the data appropriately are kept up to date and are readily available.**

   Metadata is available and the institute publishes documents that interpret the data. We consider this requirement to be **met**.

Education

The level of employee training and the availability of educational materials.

a. **Examples of how the data can be used appropriately are provided.**

   The application https://onemocneni-aktualne.mzcr.cz/covid-19 can be considered as an educational material. We consider this requirement to be **met**.

b. **Staff is available to answer questions about the data and to aid interpretation.**

   The institute does not provide official online support regarding the data. We consider this requirement to be **not met**.

5.2.6 Coherence

The degree to which the data are logically connected and mutually consistent is reflected in coherence [27]. Within a dataset, coherence means that the basic data items are based on compatible principles and can be meaningfully integrated [27].

Dimension	Characteristics	Criteria									
		a	b	c	d	e	f	g	h	i	j
Coherence	Standardization	0	3	3	2						
	Linkage	3	3	3	3						
	Historical Comparability	3	2								

Table 5.6: Evaluation of criteria for dimension Coherence

Standardization

Data standardization is the process of converting data into a standardized format that enables collaborative research and large-scale analytics [12]. It is a critical component of the data collection process because data, particularly healthcare data, can vary greatly from one organization to the next [12].

a. **All data elements are compared to a standard data dictionary in an ongoing manner and for classified data, standard classification methodologies are used (e.g., International Classification of Diseases – ICD10).**

   Information on internal procedures is **not known**.

b. **As many data elements as possible conform to a standard data dictionary.**

   The data that can be compared with the data dictionary are mainly territorial identification codes. These codes correspond to the *RÚIAN* (register of territorial identification, addresses and real estate). We consider this requirement to be **met**.

c. **Data are collected at the finest level of detail as is practical.**

   Some datasets contain anonymized data about individual patients. We consider this requirement to be **met**.

d. **For any derived variable, the original variable or variables are also maintained.**

   The dataset is available in a very consolidated form. Derived variables are preserved in some cases (total numbers of infected can be manually summed from another dataset), but this may not apply to all of them. We consider this requirement to be **not met**.

Linkage

Data linkage is a technique for combining information on the same person or object from various sources to generate a new, more comprehensive dataset [38]. The linking of data from many sources allows for the building of chronological sequences of events, which, when leveraged at the macro level, provides useful information for policy and study into the population's health and well-being [38]. In this case, the questionnaire seeks the level of standardization for the feasibility of forming a relation.

a. **Standard Geographical Classifications (SGC) can be used.**

   Yes, the data can be referenced by the local classification system. We consider this requirement to be **met**.

b. **Data are collected using a consistent time frame (e.g., fiscal year).**

   Technically, data is collected on a daily basis. We consider this requirement to be **met**.

c. **Codes are used to uniquely identify institutions (e.g., hospital numbers) and persons (e.g., health insurance number).**

   This requirement is **met**.

d. **Privacy and confidentiality rules related to record linkage are adhered to.**

   This requirement is **met**.

Historical Comparability

The term "historical comparability" refers to whether the data can be compared to previous values and whether changes in data collecting and management techniques can be traced.

a. **Trend analysis is used to examine changes in key data elements over time, and breaks in the series are explained.**

   The data itself is used to analyze the trend. The interruption of the series has not yet occurred in the data. We consider this requirement to be **met**.

b. **Documentation of changes in concepts or methods is available and easily accessible.**

   We were unable to identify any documentation that satisfies the requirement. This criterion has not been met.

5.3 Characteristics Evaluation

The algorithm in Section 4.3 was used to calculate the evaluation of the characteristics. Table 5.7 presents the results. The table summarizes the quality scores for each of the defined *characteristics* across all *dimensions*. The term "calculated score" refers to the average value calculated from the evaluation of *criteria* for a given *characteristic*. The "final score" is the value converted to the scale defined in the previous chapter.

Dimension	Characteristics	Calculated Score	Final Score
Relevance	Adaptability	2.0	(2) marginal
	Value	2.63	(3) appropriate
Accuracy	Form Design and Completion	1.5	(2) marginal
	Frame	3.0	(3) appropriate
	Over Coverage	0.0	(0) unknown
	Under Coverage	1.0	(1) not acceptable
	Response	2.5	(3) appropriate
	Completeness	0.0	(0) unknown
	Bias	0.33	(0) unknown
	Validity	1.25	(1) not acceptable
	Reliability	0.67	(1) not acceptable
	Collection	1.2	(1) not acceptable
	Processing	0.5	(1) not acceptable
	Imputation	1.5	(2) marginal
	Analysis	0.5	(1) not acceptable
Timeliness	Data Currency	3.0	(3) appropriate
	System Efficiency	0.0	(0) unknown
Accessibility	Awareness	3.0	(3) appropriate
	Ease of Access	3.0	(3) appropriate
Interpretability	Documentation	2.5	(3) appropriate
	Education	2.5	(3) appropriate
Coherence	Standardization	2.0	(2) marginal
	Linkage	3.0	(3) appropriate
	Historical Comparability	2.5	(3) appropriate

Table 5.7: Total Score of DQ Characteristics

5.4 Dimensions Evaluation

Table 5.8 displays the "calculated score" and "final score" values for the *dimensions* derived from *characteristic* values using the drill-up approach. Dimensions can be further grouped to obtain the evaluation data catalog's final score.

Dimension	Calculated Score	Final Score
Relevance	2.5	(3) appropriate
Accuracy	1.2	(1) not acceptable
Timeliness	1.5	(2) marginal
Accessibility	3.0	(3) appropriate
Interpretability	3.0	(3) appropriate
Coherence	2.7	(3) appropriate

Table 5.8: Total Score of DQ Dimensions

5.5 Data Quality Evaluation

Table 5.8 provides an evaluation of the dimensions. All dimensions will be of the same weight, because we are not able to objectively assess the importance of individual dimensions. By averaging the results, we obtain the value 2.5, which belongs to the score "appropriate (3)". The data can thus be improved in many ways, but the end result corresponds to our perception of the catalog's qualitative condition.

5.6 Application of the Methodology

Returning to the methodology discussed in Chapter 3, we will implement the framework to our use case. The entire specification process was solved by utilizing the completed questionnaire from Long et al. (2004). The **identification** activity was completed by analyzing the state of Emergency Medical Services (EMS) data in the Canadian health care system. A questionnaire, which is part of this work, is used to **specify metrics**. Metrics were **verified** using EMS administrative data [25]. We don't know if the *specification process* was iterative (in the feedback loop), but it's safe to assume it was. We would introduce feedback into the process by verifying metrics ourselves. However, we believe that metric verification is beyond the scope of this paper because it would necessitate interdisciplinary collaboration with health data experts.

The data **collection** activity, which is part of the *execution process*, is completely covered by IHIS CR data. The main part of this chapter is data **verification**, in which the data catalog is evaluated in terms of data quality. The analysis result should be attached to the data, and the catalog should be marked as verified or rejected as part of the **contract** activity.

5.7 Summary

The possibilities of statistical analysis of the selected dataset are limited and determining their quality is very difficult. The main reason for the difficulties is the nature of the data, which corresponds in structure to econometric data – already processed in a certain way – but without the freedom to use the usual statistical methods. Thus, an available data catalog is in practice a set of results without background data. Moreover, we tried to draw conclusions from existing data and tried to learn their quality without knowledge of the criterial limitations and general description of the data, which should be determined by the methodological guideline. However, the instructions providing the necessary information about the data is missing (or is insufficiently informative) in these cases.

Nevertheless, we assessed the 88 questions posed in the article by Long et al. (2004). Through the evaluation, we drew attention to possible problems with the data and their interpretation. Despite all efforts to be objective, the data are very extensive and it is therefore possible that part of the questionnaire was not properly evaluated due to insufficient analytical coverage.

Chapter 6

Conclusion and future work

6.1 Future work

We opened many topics in the paper but were unable to explore them in depth. There are numerous ways in which the paper could be improved.

One topic that deserves more attention is the issue of data quality in relation to security constraints. As discussed in Section 2.6, data security and anonymization algorithms have an impact on data quality. It would be interesting to measure the influence of these algorithms and compare them to reports on unclassified data in future work.

The use of quantitative methods for evaluating the quality of datasets is the second topic that was not sufficiently researched in the work. Quantitative dimensions of quality were partially examined in the Section 4.1. These dimensions were even put to the test on a portion of the data catalog. However, during the writing of the work, the approach proved to be a dead end because we couldn't objectively assess the current state of quality of the data catalog based on the measured values.

6.2 Conclusion

The subject of data quality is extremely broad and diverse. Although it may appear to be sufficiently researched at first glance, the application of specific (often very abstract) methodologies proves to be very problematic in practice. It is necessary to ensure a certain level of data source quality. Whether it is for the needs of the company's management or for the needs of a machine learning algorithm.

The first contribution of this work is the definition of a new methodology for evaluating the quality of datasets, as well as the presentation of its application on three examples across the data centralization spectrum. The second and most important contribution of this work is the evaluation of the catalog's quality using COVID-19 data from the Institute of Health Information and Statistics of the Czech Republic. The analysis was evaluated within the framework of the methodology that we defined.

We attempted an objective evaluation of the data catalog. However, because a portion of the questionnaire used could not be answered, a correction by a specialist in the field would be appropriate. This, however, does not diminish the importance of our work, which can now serve as a standard against which future work can be measured. It is critical to perform repeated evaluations in order to gradually improve data quality. However, due to the time consuming nature of such work, part of the questionnaire would need to be quantified and automated.

Appendix A

Data Quality Attributes

Eppler (2006) presented list of seventy of the most used data and information quality criteria explicitly defined in the literature. They provide the basis for most of the DQ frameworks.

1. Comprehensiveness
2. Accuracy
3. Clarity
4. Applicability
5. Conciseness
6. Consistency
7. Correctness
8. Currency
9. Convenience
10. Timeliness
11. Traceability
12. Interactivity
13. Accessibility
14. Security
15. Maintainability
16. Speed
17. Objectivity
18. Attributability
19. Value-added
20. Reputation (source)
21. Ease-of-use
22. Precision
23. Comprehensibility
24. Trustworthiness (source)
25. Reliability
26. Price
27. Verifiability
28. Testability
29. Provability
30. Performance
31. Ethics
32. Privacy
33. Helpfulness
34. Neutrality
35. Ease of Manipulation
36. Validity
37. Relevance
38. Coherence
39. Interpretability
40. Completeness
41. Learnability
42. Exclusivity
43. Right Amount
44. Existence of meta information
45. Appropriateness of meta information
46. Target group orientation
47. Reduction of complexity
48. Response time
49. Believability
50. Availability
51. Consistent Representation
52. Ability to represent null values
53. Semantic Consistency
54. Concise Representation
55. Obtainability
56. Stimulating
57. Attribute granularity
58. Flexibility
59. Reflexivity
60. Robustness
61. Equivalence of redundant or distributed data
62. Concurrency of redundant or distributed data
63. Nonduplication
64. Essentialness
65. Rightness
66. Usability
67. Cost
68. Ordering
69. Browsing
70. Error rate

Figure A.1: Data & Information Quality Criteria [19]

Appendix B

Dataset Collection

Available COVID-19 datasets with original, Czech titles and their categories.

B.1 Epidemiological Characteristics

1. Základní přehled
2. Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (v2)
3. Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří (v2)
4. Přehled vyléčených dle hlášení krajských hygienických stanic
5. Přehled úmrtí dle hlášení krajských hygienických stanic
6. Přehled hospitalizací
7. Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů (v2)
8. Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu
9. Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle ORP
10. Epidemiologická charakteristika obcí
11. Epidemiologická charakteristika městských částí hlavního města Prahy

B.2 Testing

1. Celkový (kumulativní) počet provedených testů (v2)
2. Přehled provedených testů podle typu a indikace

3. Celkový (kumulativní) počet provedených testů podle krajů a okresů ČR
4. Odběrová místa v ČR

B.3 Vaccination

1. Přehled vykázaných očkování podle krajů ČR
2. Přehled vykázaných očkování podle očkovacích míst ČR
3. Očkovací místa v ČR
4. Přehled spotřeby podle očkovacích míst ČR
5. Přehled distribuce očkovacích látek v ČR
6. Přehled registrací podle očkovacích míst ČR
7. Přehled rezervací podle očkovacích míst ČR
8. Přehled vykázaných očkování podle profesí

B.4 Other

1. Přehled distribuce ochranného materiálu dle krajů ČR (v2)

Appendix C

Metadata File Structure

```
{
    "@context": [
        "http://www.w3.org/ns/csvw",
        {
            "@language": "cs"
        }
    ],
    "url": "zakladni-prehled.csv",
    "dc:title": "COVID-19: Zakladni prehled",
    "dc:description": "...",
    "dc:source": "Krajske hygienicke stanice v CR",
    "dcat:keyword": ["COVID-19", "widget", "aktualni situace"],
    "dc:publisher": {
        "schema:name": "UZIS CR",
        "schema:url": {
            "@id": "https://www.uzis.cz/"
        }
    },
    "dc:license": {
        "@id": "https://data.gov.cz/podminky-uziti/volny-pristup/"
    },
    "dc:modified": {
        "@value": "2021-04-18",
        "@type": "xsd:date"
    },
    "tableSchema": {
        "columns": [
            {
                "name": "datum",
                "titles": "datum",
                "datatype": "date",
                "dc:description": "Datum vytvoreni aktualizace."
            },...
        ]
    }
}
```

Figure C.1: Metadata file structure

Bibliography

[1] *5-star Open Data.* Feb. 5, 2021. URL: https://5stardata.info/en/ (visited on 04/14/2021).

[2] Ulrich Atz. "The tau of data: A new metric to assess the timeliness of data in catalogues". In: *CeDEM14 Conference for E-Democracy and Open Government.* Vol. 22. 2014, pp. 147–162.

[3] Carlo Batini et al. "A Comprehensive Data Quality Methodology for Web and Structured Data". In: *Int. J. Innov. Comput. Appl.* 1.3 (July 2008), pp. 205–218. ISSN: 1751-648X. DOI: 10.1504/IJICA.2008.019688. URL: https://doi.org/10.1504/IJICA.2008.019688.

[4] Carlo Batini et al. "A Framework And A Methodology For Data Quality Assessment And Monitoring." In: Jan. 2007, pp. 333–346.

[5] Carlo Batini et al. "Methodologies for Data Quality Assessment and Improvement". In: *ACM Comput. Surv.* 41.3 (July 2009). ISSN: 0360-0300. DOI: 10.1145/1541880.1541883. URL: https://doi.org/10.1145/1541880.1541883.

[6] Roger Blake and Paul Mangiameli. "The Effects and Interactions of Data Quality and Problem Complexity on Classification". In: *J. Data and Information Quality* 2.2 (Feb. 2011). ISSN: 1936-1955. DOI: 10.1145/1891879.1891881. URL: https://doi.org/10.1145/1891879.1891881.

[7] Robert Blumberg and Shaku Atre. "The Problem with Unstructured Data". In: *DM Review* 13 (2003), pp. 42–46. URL: http://soquelgroup.com/wp-content/uploads/2010/01/dmreview_0203_problem.pdf.

[8] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. "An Overview of Business Intelligence Technology". In: *Commun. ACM* 54.8 (Aug. 2011), pp. 88–98. ISSN: 0001-0782. DOI: 10.1145/1978542.1978562. URL: https://doi.org/10.1145/1978542.1978562.

[9] *Common Defects in Big Data and Data Warehouses.* Mar. 30, 2021. URL: https://www-356.ibm.com/partnerworld/gsd/showimage.do?id=36622 (visited on 03/30/2021).

[10] ÚZIS ČR. *Rozšíření otevřených datových sad o očkování proti COVID-19 - Aktuality - ÚZIS ČR.* May 18, 2021. URL: https://www.uzis.cz/index.php?pg=aktuality&aid=8467 (visited on 05/18/2021).

[11] R. Cummings and D. Desai. "The Role of Differential Privacy in GDPR Compliance". In: 2018.

[12] *Data Standardization – OHDSI.* May 18, 2021. URL: https://www.ohdsi.org/data-standardization/ (visited on 05/18/2021).

[13] *Dataeum White Paper*. June 4, 2019. URL: https://dataeum.io/white-paper.pdf (visited on 03/20/2021).

[14] *Differential privacy: An illustrated primer*. May 5, 2021. URL: https://github.com/frankmcsherry/blog/blob/master/posts/2016-02-06.md (visited on 05/05/2021).

[15] OECD Statistics Directorate. *OECD Glossary of Statistical Terms - Data imputation Definition*. May 9, 2021. URL: https://stats.oecd.org/glossary/detail.asp?ID=3406 (visited on 05/09/2021).

[16] OECD Statistics Directorate. *OECD Glossary of Statistical Terms - Over-coverage Definition*. May 8, 2021. URL: https://stats.oecd.org/glossary/detail.asp?ID=4545 (visited on 05/08/2021).

[17] OECD Statistics Directorate. *OECD Glossary of Statistical Terms - Under-coverage Definition*. May 8, 2021. URL: https://stats.oecd.org/glossary/detail.asp?ID=5069 (visited on 05/08/2021).

[18] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3–4 (Aug. 2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

[19] Martin J. Eppler. *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. 2nd. Springer, 2006. ISBN: 3540314083, 9783540314080, 9783540322252. URL: https://www.springer.com/gp/book/9783540314080.

[20] Kilem Gwet. "Intrarater Reliability". In: vol. 2. Sept. 2008, pp. 473–485. ISBN: 9780471462422. DOI: 10.1002/9780471462422.eoct631.

[21] Anders Haug et al. "Master data quality barriers: An empirical investigation". In: *Industrial Management & Data Systems* 113 (Apr. 2013), pp. 243–249. DOI: 10.1108/02635571311303550.

[22] "ISO/IEC 25010: 2011 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models". In: 2013.

[23] *Koronavirus | Jak šel čas s covid-19: Čína-svět-Česko - Seznam Zprávy*. Apr. 18, 2021. URL: https://www.seznamzpravy.cz/clanek/koronavirus-covid-19-jak-sel-cas-91186 (visited on 04/18/2021).

[24] Yang W. Lee et al. "AIMQ: a methodology for information quality assessment". In: *Information & Management* 40.2 (2002), pp. 133–146. ISSN: 0378-7206. DOI: https://doi.org/10.1016/S0378-7206(02)00043-5. URL: https://www.sciencedirect.com/science/article/pii/S0378720602000435.

[25] Jennifer Long et al. "Where To Start? A Preliminary Data Quality Checklist For Emergency Medical Services Data." In: Jan. 2004, pp. 197–210.

[26] David Loshin. *Master Data Management (The MK OMG Press)*. 2008. ISBN: 0123742250, 9780123742254, 9780080921211. URL: https://www.sciencedirect.com/book/9780123742254/master-data-management.

[27] OECD. *OECD Handbook for Internationally Comparative Education Statistics 2018*. 2018, p. 148. DOI: https://doi.org/https://doi.org/10.1787/9789264304444-en. URL: https://www.oecd-ilibrary.org/content/publication/9789264304444-en.

[28] Ken Orr. “Data Quality and Systems Theory”. In: *Commun. ACM* 41.2 (Feb. 1998), pp. 66–71. ISSN: 0001-0782. DOI: 10.1145/269012.269023. URL: https://doi.org/10.1145/269012.269023.

[29] Boris Otto, Kai Hüner, and Hubert Österle. “Identification of Business Oriented Data Quality Metrics”. In: *Proceedings of the 14th International Conference on Information Quality, ICIQ 2009*. Ed. by Paul Bowen et al. Red Hook, NY: Curran, Nov. 2011, pp. 122–134. URL: https://www.alexandria.unisg.ch/71586/.

[30] *Quasi-Identifier | SpringerLink*. May 5, 2021. URL: https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-5906-5_763 (visited on 05/05/2021).

[31] Silvola Risto et al. “Managing one master data – challenges and preconditions”. In: *Industrial Management & Data Systems* 111.1 (Jan. 2011), pp. 146–162. ISSN: 0263-5577. DOI: 10.1108/02635571111099776. URL: https://doi.org/10.1108/02635571111099776.

[32] Stork.com. *Data relevance – Crucial in Asset Performance Management - Stork*. May 18, 2021. URL: https://www.stork.com/en/about-us/blog/data-relevance-crucial-in-asset-performance-management (visited on 05/18/2021).

[33] M. Talha, A. Kalam, and N. Elmarzouqi. “Big Data: Trade-off between Data Quality and Data Security”. In: *Procedia Computer Science* 151 (Jan. 2019), pp. 916–922. DOI: 10.1016/j.procs.2019.04.127.

[34] *The Six Dimensions of EHDI Data Quality Assessment*. Jan. 25, 2019. URL: https://www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf (visited on 04/01/2021).

[35] Kumar Venkatesh and Rajendran G. “Entropy Based Measurement of Text Dissimilarity for Duplicate – Detection”. In: *Modern Applied Science* 4 (Aug. 2010). DOI: 10.5539/mas.v4n9p142.

[36] *What is 5 Star Linked Data? | Webize Everything Community Group*. Apr. 14, 2021. URL: https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/ (visited on 04/14/2021).

[37] *What is Data Aware Storage?* May 19, 2021. URL: https://www.hubstor.net/blog/what-is-data-aware-storage/ (visited on 05/19/2021).

[38] *What is Data Linkage? - Menzies Institute for Medical Research | University of Tasmania*. May 4, 2021. URL: https://www.menzies.utas.edu.au/research/research-centres/data-linkage-unit/what-is-data-linkage (visited on 05/18/2021).

[39] *What is the best way to measure the consistency of data?* Mar. 10, 2021. URL: https://www.quora.com/What-is-the-best-way-to-measure-the-consistency-of-data (visited on 03/10/2021).

[40] Philip Woodall, Alexander Borek, and Ajith Kumar Parlikad. “Data quality assessment: The Hybrid Approach”. In: *Information & Management* 50.7 (2013), pp. 369–382. ISSN: 0378-7206. DOI: https://doi.org/10.1016/j.im.2013.05.009. URL: https://www.sciencedirect.com/science/article/pii/S0378720613000517.

[41] Tianqing Zhu. *Explainer: what is differential privacy and how can it protect your data?* Apr. 30, 2019. URL: https://theconversation.com/explainer-what-is-differential-privacy-and-how-can-it-protect-your-data-90686 (visited on 04/30/2019).