Collaborative Network for Industry,
Manufacturing, Business and Logistics in Europe

D6.4
Information Quality Management

Project Acronym	NIMBLE
Project Title	Collaboration Network for Industry, Manufacturing, Business and Logistics in Europe
Project Number	723810
Work Package	WP6
Lead Beneficiary	UNI-HB
Editor	Stefan Wellsandt (UNI-HB)
Reviewers	SRFG
Contributors	SRFG, SRDC, Holonix, ENEA
Dissemination Level	PU
Contractual Delivery Date	30/09/2019
Actual Delivery Date	30/09/2019
Version	1.1

NIMBLE Collaboration Network for Industry, Manufacturing, Business and Logistics in Europe

Abstract
This report presents the "information quality management" guidebook for the NIMBLE platform. In the introduction, it outlines why the quality of data and information matters for a platform. The second chapter provides background about key terms and concepts. Chapter 3 proposes the NIMBLE information quality management. It explains the basic structure and the focused approach to create awareness for quality problems. Chapter 4 presents the software tool that NIMBLE uses to raise its users' awareness. The last part of this deliverable covers the relation between cyber-threats and information quality.

NIMBLE in a Nutshell
NIMBLE stands for the collaborative **N**etwork for **I**ndustry, **M**anufacturing, **B**usiness and **L**ogistics in **E**urope. It develops the infrastructure for a cloud-based, Industry 4.0, Internet-of-Things-enabled B2B platform on which European manufacturing firms can register, publish machine-readable catalogues for products and services, search for suitable supply chain partners, negotiate contracts and supply logistics. Participating companies can establish private and secure B2B and M2M information exchange channels to optimise business workflows. The infrastructure is developed as open source software under an Apache, permissive license. The governance model is a federation of platforms for multi-sided trade, with mandatory interoperation functions and optional added-value business functions that can be provided by third parties. This fosters the growth of a net-centric business ecosystem for sustainable innovation and fair competition as envisaged by the Digital Agenda 2020. Prospective NIMBLE providers can take the open source infrastructure and bundle it with sectorial, regional or functional added value services and launch a new platform in the federation. The project started in October 2016 and will last for 42 months.

Document History

Versions	Dates	Comments
0.1	29.07.2019	First draft with structure and preliminary contents.
0.3	14.08.2019	Advanced draft with measures and QualiExplore contents
0.7	13.09.2019	Final draft with QualiExplore implementation
1.0	30.09.2019	Final version of document and QualiExplore implementation

Table of Contents

List of Figures

List of Tables

1 Introduction

The objective of Task 6.4 was to develop an information quality approach for the NIMBLE platform, including the view on data integrity as one of the three cornerstones of the cybersecurity model called CIA (Confidentiality, Integrity, Availability). Data integrity is related to data security, both giving different but complementary perspectives on the data. Data security deals with, for instance, the protection of data against unauthorized data manipulations and ensures the data integrity, while data integrity checks for validity and accuracy of data referring to possible security related data corruption.

Information quality and data integrity are important for B2B software platforms because accurate information is a foundation of informed business decisions. Problems such as inaccurate, imprecise, ambiguous, delayed, and inaccessible information can result in decisions that are costlier than expected because of corrective actions, delays, damages and missed opportunities. The risks that emerge from low information quality affect the users' trust in the platform and thus the platforms profitability. The platform operator has a strong interest to offer measures and software tools to avoid quality problems. One approach to achieve this is to implement **Information Quality Management** (IQM).

IQM is a comparably young domain that first emerged in the 1990s. Today, it adopts the conceptual background of the ISO 9000 quality management standard family, defining quality as the match between an object's characteristics and pre-defined requirements. It also assumes that management includes the actions "plan, do, study, and act" defined by Edwards Deming.[1]

[1] https://deming.org/explore/p-ds-a

This so-called **Deming Cycle** begins with the *Plan* step that includes the identification of goals, formulating a theory, and defining success metrics. The *Do* step realizes the plan while the *Study* step monitors its outcomes to test the validity of the theory regarding success, progress, problems, and potential improvement. Finally, the *Act* step transfers what the managers learned from the process into changes of the goals, methods, theory, and other aspects of improvement. In NIMBLE, the IQM grounds on the Deming Cycle and the idea that platform operators should identify and avoid quality problems.

Chapter 2 explains the basic paradigms in data quality. Chapter 3 presents the background of IQM. Chapter 4 introduces the proposed solution. Chapter 5 presents the QualiExplore tool. Chapter 6 covers cyber-threats and chapter 7 the conclusion remarks.

2 Background

This chapter introduces important terms and concepts needed to design and realize an IQM.

2.1 Data and information

The foundation of an IQM is the data and information. Researchers discuss these two concepts controversial. This paragraph adopts the viewpoints of Rowley (2007) and Wilson (2002). They examine the meaning of the data and information critically. *Data* are signs with no or little meaning that represent facts, such as a temperature or a system state. It is of little use for practical applications but the essential object of interest within data storage infrastructures and data analysis. *Information* is organized data, embedded in a context, and in general meaningful for someone or something. This report uses both terms synonymously as typically done in practical cases.

2.2 Data and information quality

Researchers investigated information quality intensely in the 1990s. Individuals and joint activities, such as the MIT Total Data Quality Management Programme[2], developed a rich body of knowledge for this topic. A key question of the early research was which characteristics refer to information quality (Wang and Strong, 1996).[3] This research led to the first empirically backed quality model with distinct characteristics that matter for information users. Many of these models use the quality concept described in the ISO 9000 standard family. It defines quality as the "*degree to which a set of inherent characteristics of an object fulfils requirements*". In IQM, the object is information. The identification of relevant characteristics and their corresponding requirements is less obvious. In general, one can determine relevant characteristics with three approaches (Liu and Chi, 2002):

- *Intuitive*: characteristics ground on an expert's experience.

[2] http://web.mit.edu/tdqm/

[3] The early topics do not clearly differentiate between data and information quality yet. Most of the research simply refers to data quality management.

- *Empirical*: derives characteristics from the information users.
- *Theoretical*: derives characteristics from a theory, e.g., mathematical theory of communication, information economy, and operations research.

Researchers from various domains created, adopted, revised, or extended information quality models. Today, the international standard ISO/IEC 25012:2008 (ISO, 2008) proposes 15 data quality dimensions.[4] They include amongst others accessibility, accuracy, completeness, credibility, precision, and understandability.

2.3 Data and information quality management

From the quality management perspective, data and information are not the same. *Data quality management* focuses on technical aspects – oftentimes related to the data storage in databases. It seeks to reduce problems, such as duplicates and syntax errors. IQM belongs to the organization's *information management* (IM) process. IM manages the processes, resources, technologies and policies in an organization that focus on information (Choo, 2002). It prepares, realizes and monitors information systems that supply the employees and stakeholders with information. The concept is much wider in comparison with data quality management. IQM promotes a user-centred view and emphasizes the understandability and usability of information. The wide scope of IM means that IQM must take into account a variety of factors that influence information quality. They include:

- collection, organization, distribution, and application of information (processes)
- employee behaviour and the available IT infrastructures (resources)
- advantages and disadvantages of data processing methods (technologies)
- security and privacy regulations and governance models (policies)

B2B software platforms are relevant in this context because they support information-based processes, employ resources, use technologies, and must meet requirements that emerge from business policies. In the B2B world, the sharing of information also has an influence on a platform's peer organizations and even their stakeholders. One organization's information output is the input for other organizations. Comparably simple quality problems, such as an interrupted data stream, can have a significant impact when many organizations consume this stream in their business processes. Figure 1 and Figure 2 illustrate these challenges.

[4] ISO reviewed their 25012:2008 standard in 2019. The 2008 revision is recent.

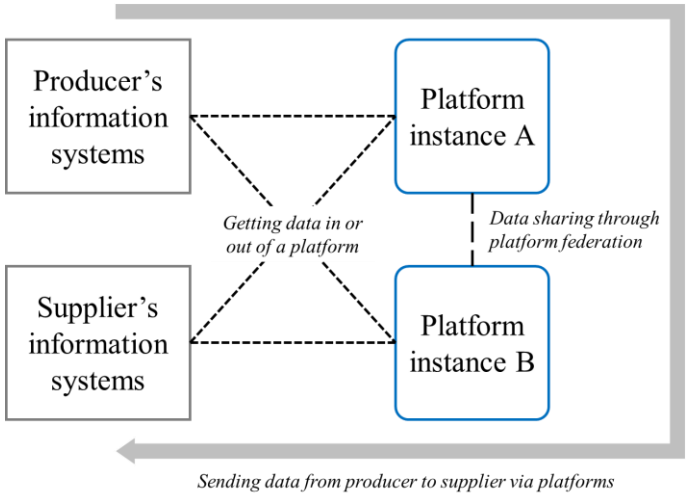

Figure 1: Data sharing through B2B platform instances

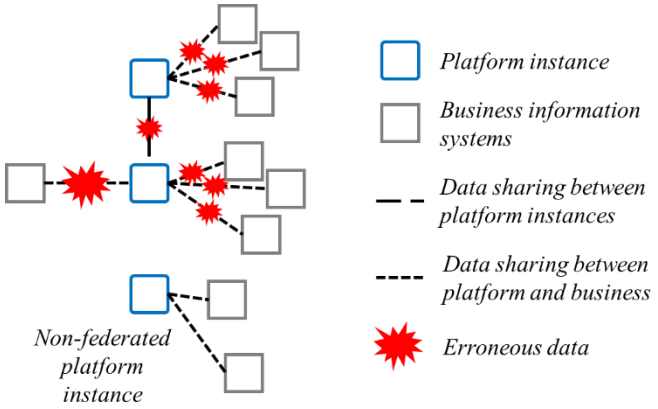

Figure 2: B2B platforms and the potential impact of information quality problems

The *impacts of information quality problems* are difficult to quantify because they depend on the specific cases where users apply the information. For B2B software platforms, no specific information about these quality problems is available to our knowledge. Eppler (2006) summarized generic symptoms and causes of eight quality problems (Table 1).

Table 1: Information quality problems, symptoms and causes based on Eppler (2006)

Problems	Symptoms	Causes
Limited usefulness	Information overload	Lack of cleansing, maintenance, analysis or synthesis
Ambiguity	Different or wrong interpretations	Lack of precision or accuracy; use of abbreviations or jargon; different viewpoints
Incompleteness	Inadequate decisions	Fragmentation of work; infrequent communication and exchange of information, incompatible IT systems; lacking alignment between It strategy and business strategy
Inconsistency	Confusion; contradictory statements	Lacking coordination between information authors and distributors; unclear responsibilities; use of multiple, inconsistent information sources
Inadequate presentation format	Expensive conversion tasks; order, format, style that does not allow direct use	Insufficient dialogue between information producer and consumers; constant time pressure;

Not reliable or trustworthy	Great risk of errors; information's background missing	Mistakes in the information production and distribution; unidentified sources
Not accessible	Lost over time; demotivated staff; wrong decisions	Unclear responsibilities; technological changes
Distortion of information	Original message not the same when received	Too many intermediaries, specialization and jargon; misinformation; modifying, delaying, and blocking information

A second aspect of information quality problems is its *impact on trust and user acceptance*. If the platform cannot maintain a high quality of the information it distributes, it could lose its users' trust. This could motivate companies to leave the platform, which reduces the benefits for the remaining companies. The beneficial side of the network effect of platforms is turning against the platform as its user base diminishes. This will potentially translate into less revenue up to the point where the platform cannot provide sufficient benefits to justify its operation.

The situation above explains why B2B platforms must minimize information quality problems. An effective management process for this domain is a suitable instrument for platform operators. Chapter 3 presents NIMBLE's approach to it.

3 NIMBLE information quality management

The NIMBLE IQM adopts concepts of the ISO 9000 standards family and ISO/IEC 25012:2008. It uses the PDSA cycle as the basis for activities. This report suggests activities that help the platform operators and users to maintain high information quality. The suggested activities cover awareness, programmatic, and organizational measures.

3.1 Basic structure

The PDSA cycle provides an efficient general-purpose management structure and is an important component in the ISO 9001:2015 standard for quality management systems. Chapter 3.1.1 introduces the four steps of the NIMBLE IQM. The general structure is simple and flexible to allow each platform instance the case-specific detailing. Chapter 3.1.2 outlines important information quality characteristics in the context of B2B software platforms.

3.1.1 Plan-Do-Study-Act

Plan. The first step in this cycle identifies goals, formulates a theory, and defines success metrics for information quality on the NIMBLE platform. It also plans the activities to realize the goals, such as new or revised platform functions and organizational procedures, and the collection of data needed to assess the progress against the goals.

- *Goals* clarify how the platform instance operator wants information to be. Reaching the goal means to achieve a change in information quality (improvement).
- *Theory* outlines, for instance, how quality problems relate to interface components or storage procedures.

- *Success metrics* specify with numbers under which conditions the information quality fulfils the goals.

Do. In this step, the platform operator performs the planned activities to improve information quality. This includes changes to the platform functions (e.g. improved user interfaces) and non-functional procedures (e.g. business identify verification, and user training). The staff also collects the data needed to assess the success metrics. In NIMBLE, the data collection can use the platform instance's logs for example.

Study. The third step analyses the collected data and calculates the success metrics. It identifies issues in the plan and removes obstacles that hamper achieving the goals. Potential issues include human and computational resource bottlenecks, and interference from technical platform changes. Cybersecurity threats and attacks targeting data quality, e.g. ransomware attacks resulting in hackers encrypting data, can cause another set of potential issues. Cyber Threat Intelligence (CTI) or relevant security threat evidence are methods that can explore such issues.

Act. This step concludes the study results and identifies further actions to reach the goals. It can also change goals. A new cycle starts with new goals or adapted ones. In the light of cybersecurity, this step includes a variety of actions to prevent and detect attacks that affect the data integrity. They range from implementing audit trails, to establishing management security qualification and maintenance programs.

Figure 3: Summary of NIMBLE's basic IQM structure

Figure 3 summarizes the basic structure of the IQM approach in NIMBLE.

3.1.2 Information quality characteristics

The *Plan* step includes the definition of success metrics for the information quality goals. The ISO/IEC 25012:2008 standard provides a data quality model with 15 quality characteristics. It helps organizations to define and evaluate data quality requirements, identify quality assurance criteria, and evaluate the compliance of data in terms of privacy and security. Platform operators can extend the model where needed in their specific IQM. Table 2 summarizes the ISO/IEC 25012:2008 characteristics in the light of the NIMBLE platform.

Table 2: Overview of NIMBLE information quality characteristics

Characteristics	Descriptions
Accuracy	Syntactic and semantic closeness of information in relation to the information defined as correct in the targeted domain.
Completeness	To what extent the data or entity has values for all expected attributes.
Consistency	To what extent information is free from contradiction and coherent with other information.
Credibility	How true and believable the information is. Believability is a surrogate characteristic taking a bundle of quality characteristics into account.
Currentness	Adequacy of the age of the information for a specific context of use.
Accessibility	How well users can get the needed information also considering capabilities of the individual user.
Compliance	How well data adheres to standards, conventions and regulations.
Confidentiality	How accessible and interpretable data is by authorized users in a specific context of use.
Efficiency	To which degree the data allows processing without wasting resources.
Precision	How well data is exact and allows users to differentiate in a specific context of use.
Traceability	How well users can understand data changes and access of data.
Understandability	How well users can read and interpret the data.
Availability	How well authorized users and applications can retrieve information in a specific context of use.
Portability	How well the platform providers and users can install, replace and move data from one platform to other systems.
Recoverability	How well the data contributes to maintenance and preservation of platform operations and quality of service.

The descriptions of these characteristics can vary among the NIMBLE platform instances. Their specific meaning depends on the supported business processes and use cases, the platform infrastructure, and the information shared via the instance (e.g. Internet of Things data). Platform operators can use the quality characteristics to define **measures** that support the *Study* step of the NIMBLE IQM. Table 3 summarizes examples for information quality measures.

Table 3: Examples of information quality measures for B2B platforms

Characteristics	Measures

Accuracy	Number of available calibration protocols for sensor data sets in relation to all sensor data sets shared via the platform.
Completeness	Number of filled product catalogue fields in relation to all product catalogue fields for an item.
Precision	Number of fields with the requested decimal places in relation to all database fields.
Understandability	Number of translated texts in relation to all texts in default language.
Portability	Number of exportable fields with user information in relation to all available fields with user information.

Even more than the quality characteristics, the quality measures depend on the specifics of each platform instance. Measures are mathematical fractions that have, for instance, the desired attribute values (e.g. calibration information) in the denominator and the totality of entities (e.g. all sensor data) in the numerator:

(# of available calibration protocols for sensor data sets) / (# of all sensor data sets)

The platform operator can calculate the percentage of coverage and use it to measure how well functional or organizational changes affect accuracy in this specific context. The goal could be to ensure that all sensor data have a calibration protocol. Potential activities are:

- Making the form field for uploading the calibration protocol mandatory.[5]
- Raise an alert when the protocol is missing but the user wants to share a data set.
- Penalise users by warning sensor data consumers when using the inaccurate data.
- Incentivise users by introducing a data value that depends on accuracy.
- Inform users and create awareness that the protocols are important for sensor data.

3.1.3 IQM activity framework

Not all IQM activities in the Do step must be technical. NIMBLE uses a simple framework to structure the IQM activities meant to avoid quality problems. Figure 4 illustrates it along with the expected cost to develop and maintain the activities.

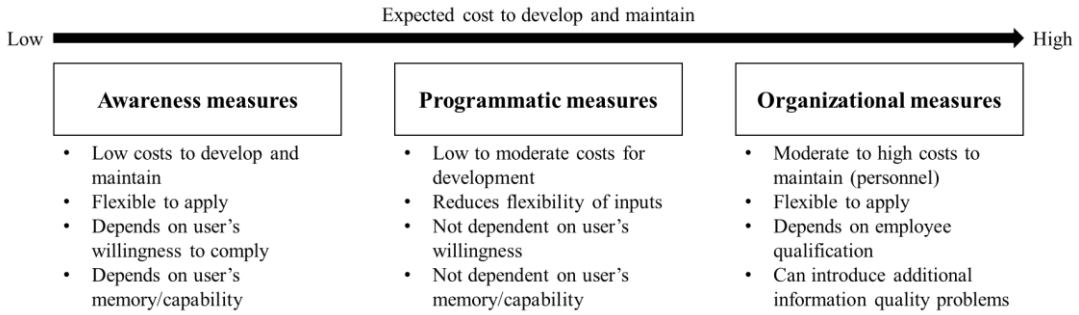

Figure 4: IQM activity framework for the NIMBLE platform

- *Awareness measures* inform platform users about information quality and the factors that affect it. These measures will be cheap to develop and maintain because they do

[5] A downside of this measure is the decrease in user experience – the user might not have the protocol.

not require a deep integration in the platform software – static websites with information could be sufficient to raise awareness. Awareness measures are flexible because one solution can make users aware of various topics (refer to Chapter 4). The downside of this measure is that it depends on each user's willingness and capability to behave in a way that avoids the quality problems. Awareness measures alone would not lead to an effective IQM for A NIMBLE platform instance.

- *Programmatic measures* enforce user behaviour via technical functions. They are more costly to develop and maintain because developers must integrate them in the platform software. These measures typically restrict user inputs, which reduces the flexibility of forms and may lead to bad user experiences. The main advantage of programmatic measures is that they are not dependent on a user's willingness or capability to comply with a policy, practice, or instruction. They provide good complementary solutions for awareness measures.
- *Organizational measures*. Programmatic measures can be too costly or too restrictive for some complex use cases. The platform provider or the user can apply measures that rely on instructions for human employees in these cases. The measures aim to provide, organize or validate data in a way that increases or maintains the information quality. Organizational measures can introduce new information quality problems because the involvement of employees and work instructions creates new error causes. A demotivated employee, for instance, could perform a company validation less carefully. The result could be an illegitimate duplicate of an existing company.

Chapter 3.2 introduces a lightweight "awareness" approach supported by a platform function.

3.2 Create awareness with cause-effect diagrams

Creating awareness for information quality and the factors that influence information quality benefit IQM. Users may become more careful when they create information on the platform or when they connect machine generated data sources to it.

An important instrument in quality management that visualizes how factors influence quality is the Cause-Effect diagram. Practitioners also refer to it as Ishikawa or Fishbone diagram.[6] Liu and Chi (2002) applied this diagram type to IQM. The NIMBLE platform adopts this approach to IQM through the identification of the relevant *information characteristics and quality factors* that can help platform users to create awareness for information quality. The factors cover human generated information and machine generated information on the highest abstraction layer, as described in the remaining parts of this chapter. Below this layer, platform-related functions and concepts define categories of factors.

3.2.1 Human-generated information

Information generated by humans enters the platform through online forms and file attachments mainly. The following NIMBLE platform components contain this information.

3.2.1.1 User and company profiles

[6] https://asq.org/quality-resources/fishbone

User and company profiles describe the actors of a platform instance. They provide basic business information for the interactions between two actors, such as addresses and default terms and conditions.

Every user can create a *user profile* but to edit them they need to contact the platform provider. The user profile is important for the authentication and the identification of the users' by their names. Only users with specific user roles can create or edit *company profile* information, e.g. the legal representative. This restricted access is a measure to avoid fraudulent information changes (c.f. data integrity). It also helps to avoid quality problems because the users with these roles are typically qualified to provide the correct information, e.g. company legal data, delivery addresses, trade details, and terms and conditions. They are also familiar with the company's business processes and know when information received an update.

Potential quality problems in this area are:

- Typos and other *syntactic errors* that produce unwanted or unexpected results in the company search function. The impact is mainly bad user experience e.g. not being able to find a company. Frequent typos in a company description could reduce the credibility of other information this company provides (e.g. catalogue information). They could also reduce the trust score making future business difficult for a company.
- *False information* (e.g. delivery address and company certificates) that result in erroneous business processes and decisions that require costly corrections. Depending on the degree of inaccuracy, the consequences can be more or less harmful. A false address where only the building number is inaccurate is less costly compared to an address where the city is wrong. A variant of this quality problem is *misinformation*. Malicious users can create false information with the intention to produce harmful business processes and decisions.
- *Outdated information* (e.g. terms and conditions) is similar to false information. It is false but only in combination with a specific time or timeframe. The potential consequences are mainly the same.
- Inaccurate translations and other *semantic errors* can make profile information ambiguous or decrease its understandability. Potential consequences range from bad user experience to wrong business decisions that can affect both security and safety features related to users.

In parallel, potential data integrity problems could be related to either data stored in databases, or data in Microservices architecture, as in the NIMBLE platform:

- For databases, there are four types of data integrity (Brook, 2019):
    - **Entity Integrity:** The columns, rows, and tables are the main elements that contain the data in a database. None of these elements should be the same and none of these elements should be null (e.g. primary key).
    - **Referential Integrity:** It refers to the accuracy and consistency of data within a relationship. In relationship, data is linked between two or more tables, using foreign keys that relate data that could be shared or null. For instance, employees could share the same role or work in the same department. In other words, referential integrity requires that, whenever a foreign key value is used it must reference a valid primary key in the parent table.

    - **Domain Integrity:** All categories and values in a database are set, including nulls. It refers to the common ways to input and read data. For instance, if a database uses monetary values to include dollars and cents (with two decimal places), including three decimal places will not be allowed.
    - **User-Defined Integrity:** There are sets of data, created by users, outside of entity, referential and domain integrity.
- Data integrity for Microservices: When it comes to Microservices, the problem of **data consistency** between two users becomes evident. In Microservices, one atomic operation usually spans multiple Microservices, each implementing independent storage solutions for the data. Some solutions to this issue suggest using the Saga Pattern that forces rollback of the individual transactions (e.g. by introducing a "Cancel" operation). Another solution is to perform **data reconciliation** on a scheduled basis, by running a record-by-record comparison (e.g. by comparing aggregated values for each record). In some other cases, **event logs** can provide an insight into the status of a transaction, or on the transaction state (e.g. in cases of complex, multistep orders with booking flights, hotels, and money transfers).

3.2.1.2 Product and service catalogues

The description of product and service offers in catalogues is important information that provides the basis for many platform-supported business processes (e.g. negotiation and ordering).

Users provide catalogue information through forms, spreadsheet templates, or the NIMBLE API either directly in NIMBLE or via third party tools (e.g. Excel or catalogue management software). The access is restricted to specific roles (e.g. Publisher) typically assigned to users with background knowledge about the related business activities. They also know when catalogue information requires an update.

Potential quality problems in this area are:

- Typos and other *syntactic errors* that produce unwanted or unexpected results in the product/service search function. The impact is mainly bad user experience e.g. not being able to find a product/service. Frequent typos in a company description could reduce the credibility of other information this company provides (e.g. company profile information). They could also reduce the trust score making future business difficult for a company.
- *False information* (e.g. property values) that result in erroneous business processes and decisions that require costly corrections. Depending on the degree of inaccuracy, the consequences can be more or less harmful. A false colour value could lead to an order that the customer does not want and will return. Erroneous dimensions can affect transport capabilities and may result in an order that the transport company cannot pick up at the supplier's site. A variant of this quality problem is *misinformation*. Malicious users can create false information with the intention to produce harmful business processes and decisions.
- *Outdated information* (e.g. prices and discounts) is similar to false information. It is false but only in combination with a specific time or timeframe. The potential consequences are mainly the same.

- Inaccurate translations and other *semantic errors* can make catalogue information ambiguous or decrease its understandability. Potential consequences range from bad user experience to wrong business decisions.

3.2.1.3 Business processes

Business processes are the main value-adding interaction on a NIMBLE platform instance. Each process may use existing information from user and company profiles, and a catalogue. Negotiation and order information has the highest quality requirements because this business process builds contracts with the supplied information. Not fulfilling a contract, even an unintended or false one, may have costly consequences (e.g. fines and a loss of trust score).

Users with the related roles (e.g. Purchaser) can create the information necessary to order a product or service. Erroneous information provided in the order stage will typically result in erroneous contracts, which harms the collaboration.

Potential quality problems in this area are:

- Typos and other *syntactic errors*. The impact is mainly bad user experience e.g. not being able to negotiate further, prolonging the negotiation stage to correct the error, and difficulties in processing an order. Frequent typos could reduce the credibility of other information this company provides (e.g. company profile information). They could also reduce the trust score making future business difficult for a company.
- *False information* (e.g. quantities and prices) that result in an erroneous order that requires costly corrections. Changes are typically costly because they affect processes outside of the NIMBLE platform instance, such as rescheduling work, reclamation, and delayed delivery. A variant of this quality problem is *misinformation*. Malicious users can create false information with the intention to produce harmful business processes and decisions. The business processes are an attractive target because of their comparably high costs to correct wrong decisions.
- *Outdated information* (e.g. prices and addresses) is similar to false information. It is false but only in combination with a specific time or timeframe. The potential consequences are mainly the same.
- Inaccurate translations and other *semantic errors* can make catalogue information ambiguous or decrease its understandability. Potential consequences range from bad user experience to wrong business decisions with a tendency of being more expensive to correct than errors in a catalogue or profile.

3.2.1.4 Third-party tools

The NIMBLE API allows third party tools to interact with a platform instance. An example for a tool is the Balance.LCPA tool that calculates and provides product lifecycle information for catalogues. Third party tools can depend on human generated information as an input. Depending on how the NIMBLE instance uses the information of external tools, the consequences can range from bad user experience to costly corrections. The latter is the case if the information is part of a contract.

Third party tools are outside the NIMBLE platform instance authentication and authorization frameworks. It is not clear which users provide information in a third party tool nor how they relate to a corresponding company profile in the platform. Typically, service providers

maintain third party tools based on a contract with the platform operator or a company present on the platform instance.

- Typos and other *syntactic errors* that produce unwanted or unexpected results in the platform instance component that uses information from third party tools. Frequent typos could reduce the trust score making future business difficult for a company. This is especially problematic, if a third party provides the information without being on the platform. In this case, the company on the platform must pay close attention to its service provider.
- *False information* that result in erroneous business processes and decisions that require costly corrections. Depending on the degree of inaccuracy and the area where the platform uses the information, the consequences can be more or less harmful. A variant of this quality problem is *misinformation*. Malicious users can create false information with the intention to produce harmful business processes and decisions. For third party software, this problem is more important because the software is outside the platforms security framework and could be the ideal access point to feed misinformation into the platform.
- *Outdated information* (e.g. prices and discounts) is similar to false information. It is false but only in combination with a specific time or timeframe. The potential consequences are mainly the same.
- Inaccurate translations and other *semantic errors* can make catalogue information ambiguous or decrease its understandability. Potential consequences range from bad user experience to wrong business decisions.

3.2.2 Machine-generated information

Businesses create information through machines. In this context, machines mean systems that produces information automatically based on a programme. Examples include measurement systems, software that maintains log files, and software that calculates values. The following paragraphs use the NIMBLE's open tracking and tracing (T&T) service to clarify quality problems with machine-generated information.

T&T is a service that typically relies on information created by automated measurement systems. These systems ground on technologies, such as Radio Frequency Identification (RFID) and environmental sensors. They use a software to capture events or environmental parameters, such as temperature and humidity, automatically. Organizations store their monitoring information in specific databases. The structure of these storages must consider standards, such as Electronic Product Code Information Services (EPCIS). Common information quality problems include incomplete, inconsistent, and unreliable information.

- Many employees (or software) in manufacturing companies must deal with *incomplete historic information* about products. Gaps in historic data limit the capability to understand what has happened. A common area where this can be a problem is root cause analysis. The impact of incomplete historic information is difficult to generalize. In many situations, humans or software can compensate gaps by interpolation or making assumptions. If a temperature dataset always stays within the range of -25°C

to +35°C, the probability that missing values are within this range is high. Table 4 lists the identified quality factors that can cause incomplete historic data.

Table 4: Factors causing incomplete historic information

Factors	Descriptions
Broken hardware	Damaged environmental sensors and RFID antennas. Software cannot reach the hardware and does not create data.
Faulty software	The software does not trigger the measurement as intended. Software does not create data for these cases.
Connectivity problems	The network connection is weak or breaks down. Data does not reach the data storage.
Wrong data storage format	The software stores data in the wrong format e.g. following the EPCIS standard.
Faulty data reading	Reading does not capture all relevant events or data points due to e.g. faulty filter conditions.

The medical sciences developed a comprehensive view on missing data (Papageorgiou et al., 2018). They use categories of "missing" and suggest measures to address this problem.

- An important component of T&T is the master data management. The T&T service uses these data to describe locations and business activities. Monitoring data contains a reference to the master data. *Incomplete master data* limits the capability to understand events and environmental parameters. If the location name of an event is missing, humans will not be able to understand it. An important factor that influences incomplete master information is oversight.
- T&T informs users about product locations, environmental conditions, and business steps. The combination of this information allows the user to comprehend product-related events and states. *Inconsistencies* (e.g. contradictions) between the information or between the information and the user's knowledge limit this capability. One factor that resulted in inconsistencies is the flawed installation of measurement equipment. The measurement generates accurate information but about the wrong phenomenon or product. A second factor is the installation of a flawed measurement that generates inaccurate information not aligned with user knowledge.
- *Lack of credibility* is a common reason why employees will not use information for their tasks. Factors that influence the credibility of machine-generated information are, for instance, errors in historic data, and missing background about the source (e.g. calibration protocols) or the information processing (e.g. intransparent algorithm).

3.3 Programmatic and organizational measures

Web development practices propose efficient **programmatic** solutions to avoid information quality problems. This technical enforcement comes at the cost of flexibility, which can cause bad user experience (e.g. user being unable to extend a list of options). Table 5 summarizes those measures relevant for the NIMBLE platform.

Table 5: Programmatic measures to avoid information quality problems

Measures	**Realization in NIMBLE**
Form field value limitations	Checkboxes, radio buttons, selections; regular expressions for text fields; number of allowed characters; file extension limitations
Form field value suggestions	Auto complete, auto correction
Form field value validations	Regular expression checks, value presence checks
File content validations	JSON document validation (D3.8)

In NIMBLE we differentiate between limitations, suggestions and validations as follows:

- *Input limitations*. These measures focus on restricting which data the user can enter in a form field. The platform's frontend (client-side) is mainly responsible for handling these limitations. It can restrict inputs on the level of symbols and it can enforce that inputs follow pre-defined patterns, such as an email address.
- *Input suggestions*. These measures focus on suggesting form field values to the user while they do not restrict the actual input. The auto completion during the product category selection in product/service publishing is an example where the NIMBLE platform applies this measure.
- *Input validations*. These measures focus on testing input values after the user submits a form. The backend (server-side) mainly handles these validations. A common validation measure is to check if a "required" form field value is present. Presence validation is an effective measure to ensure a minimum completeness of a dataset (e.g. a product description).

Some programmatic solutions are too restrictive or would be too complex and costly to create and maintain. In these cases, **organizational** measures can help platform operators and users to maintain high information quality on a NIMBLE platform instance. Their downside is that users can deviate from them, which can result in information quality problems.

Table 6: Organizational measures to avoid information quality problems

Measures	**Realization in NIMBLE**
Human in the loop validations	Company validity check
Complementary quality information	Attach sensor calibration protocol to an order

The *validation of company profiles* is the most critical organizational measure in a NIMBLE instance. The accuracy and credibility of company information has far-reaching consequences for legal matters (e.g. contracts) and trustworthiness between the platform users. The platform provider handles the validation off-platform. Sometimes platform users utilize the flexibility in forms to cover recurring use cases that the platform instance does not support. An example is the sharing of *quality information* for sensor data. Users that require measurements need confidence in the measurement accuracy. The product specification sheet contains information about the precision and the tolerances of measurement (e.g. ± 5%). Regular sensor calibration can be necessary to address tolerance changes caused by degradation and damage.

Users that provide a calibration protocol along with an order allow the information user to understand and interpret the measurements.

Complex IQM methods integrate **programmatic and organizational** measures. This report provides two examples as an outlook of such measures. The first method is about *data curation* service on the platform, while the second method deploys an incentive model that encourages users to maintain a high data quality.

Data curation is *"[…] the act of discovering data sources of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite*" (Stonebraker et al., 2013). A NIMBLE platform instance can store or forward a large amount of business-related data. If the platform provider operates a curation service, it can pre-process and analyse this data to create value for platform users. In this process, maintaining the quality of data is critical and the service provider must employ human resources to, for instance, clean data from duplicates, missing entries, inaccuracy, and imprecision. A data curation service integrates with the platform instance's business model and requires a transparent legal basis with clear ownership and use conditions.

The second method is to deploy an *incentive model* on a platform instance that encourages users to maintain a high information quality. This method could integrate with the current trust rating functionality, introduce a new rating type, or go as far as calculating a value for datasets. The DataBroker DAO platform developed an IoT data marketplace that implements the latter concept.[7] They use an Ethereum-based token to calculate prices for IoT datasets. Incentive models deeply integrate into a platform instance's business model and require new NIMBLE platform functionality not present yet.

4 QualiExplore

QualiExplore is a software service for the NIMBLE platform that means to increase the users' awareness for information quality. It visualizes information quality characteristics and quality factors relevant for the NIMBLE platform. The software adopts the Evolutional Data Quality Concept of Liu and Chi (2002). QualiExplore is fully customizable for a platform instance.

Chapter 4.1 presents the Evolutional Data Quality Concept. It structures the quality characteristics and factors by stages in the information lifecycle. Chapter 4.2 outlines the implemented concept for QualiExplore.

4.1 Evolutional Data Quality Concept (EDQC)

Liu and Chi (2002) developed a theory-based view on data quality that focuses on the evolution of data along a life cycle. Their data evolution life cycle contains four phases:

[7] https://databrokerdao.com/

- Data *collection* concerns data capturing through observation of real world processes, measurement, and perception.
- Data *organization* means structuring and storing of data in files, databases and other forms of data storage.
- Data *presentation* subsumes processing, interpretation, summarizing, formatting and presentation of data in views.
- Data *application* is the final phase where users utilize data to achieve a purpose, which can trigger further data collection.

An important aspect of the EDQC concept is that quality characteristics in a phase contribute to the characteristics of the following phases. Figure 5 illustrates this behaviour through a cause-effect diagram.

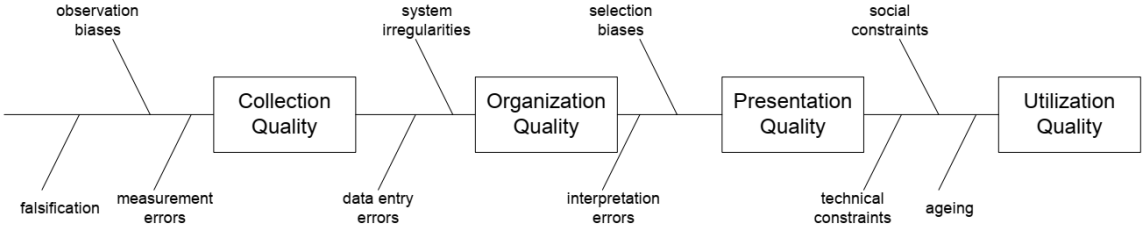

Figure 5: Cause-Effect diagram for EDQC (Liu and Chi, 2002)

4.2 Implemented concept

The QualiExplore implementation provides a *2-staged user interface* to support learning about platform-related information quality and the factors that influence it. The first stage serves as a filter because the high number of factors can cause information overload for platform users. Figure 6 illustrates the interface to select filters.

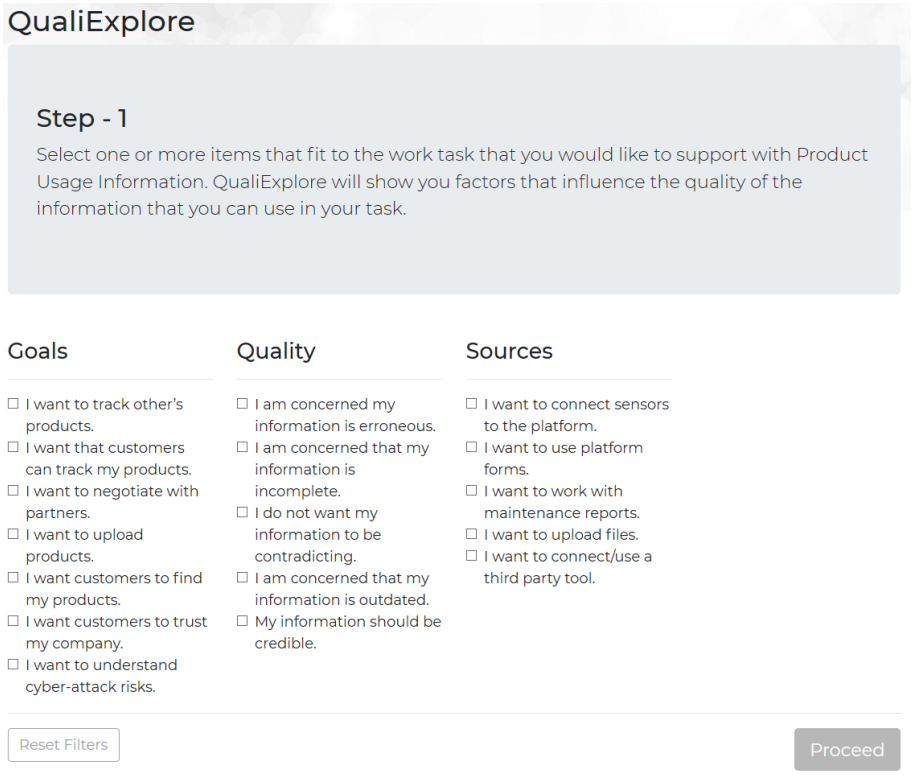

Figure 6: First step with filter functions of the QualiExplore tool

Relevant *filter categories* are the user's goals (platform services), quality (information characteristics), and channels/sources. The goals include the perspective of the information user and the information creator/author. This is useful because it emphasises that many measures to avoid quality problems require the involvement of both parties. *Statements* represent areas where the user should be or might want to be aware about information quality problems and its related factors. The indicated *factor categories* structure the factors and provide a link between statements and factors. Table 7 provides an overview about the QualiExplore filters.

Table 7: Overview on filter function contents and their relation to quality factors

Filter by	Statements	Factor categories
Goals	I want to track other's products.	Track and trace
	I want that customers can track my products.	Track and trace
	I want to negotiate with partners.	Business process
	I want to upload products.	Product catalogue
	I want customers to find my products.	Product catalogue
	I want customers to trust my company	Company profile
	I want to understand cyber-attack risks.	Platform security
Quality	I am concerned that information is erroneous.	Accuracy
	I am concerned that information is incomplete.	Completeness
	I do not want my information to be contradicting.	Consistency
	I am concerned that my information is outdated.	Timeliness
	My information should be credible.	Credibility
Sources	I want to connect sensors to the platform.	Machine-generated

	I want to use platform forms.	Human-generated
	I want to work with maintenance reports.	Human-/machine-generated
	I want to upload files.	Human-generated
	I want to connect/use a third party tool.	Human-/machine-generated

Figure 7 illustrates the second step of the QualiExplore with applied filters.

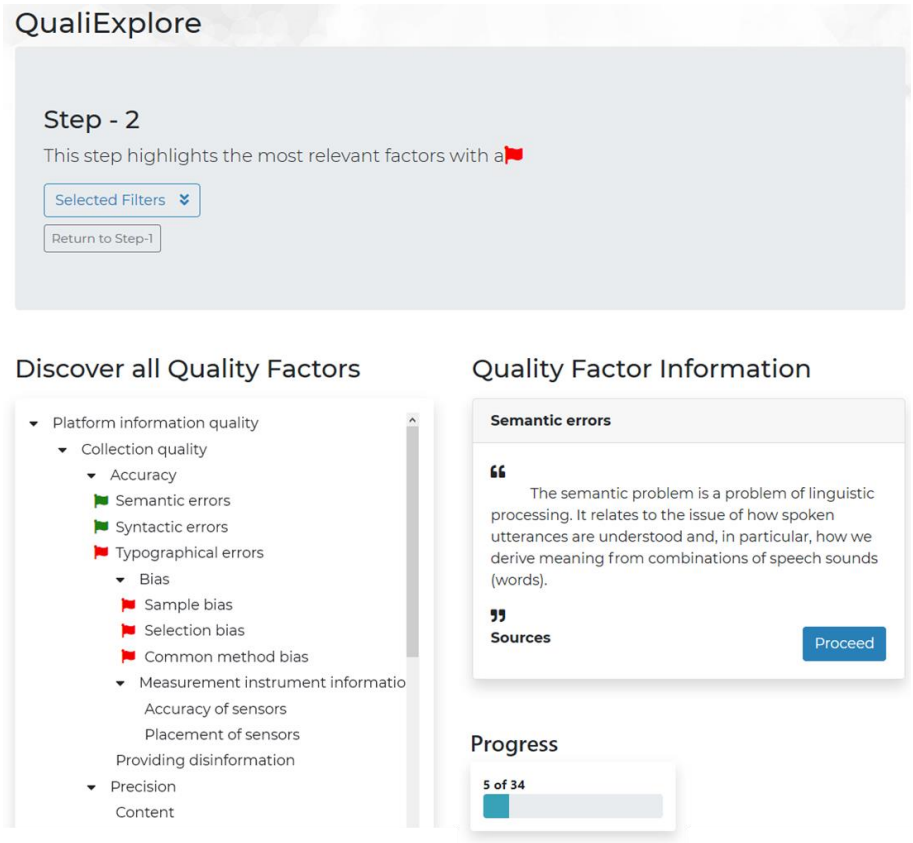

Figure 7: Second step with factor overview, factor details, and progress bar

The basis for this illustration is a tree of categories that has the quality factors as leafs. Leafs receive a flag if they are within the scope of the selected filters. The user can click on a factor to receive a description. A push on the "Proceed" button next to the description turns a factor's red flag into a green flag. This indicates the user is now aware of the factor. At the same time, a progress bar indicates the number of green flags in relation to all flags.

The NIMBLE platform integrates QualiExplore as a supportive component that serves as a "Guidebook" for IQM (D6.4). Every user can access it via the navigation bar. The progress bar provides a self-assessment to users that reflects how much information they accessed about information quality in relation to how much information is within the filters' scopes.

Two JSON files contain the description of the first and second step. The description of factors contains identifiers that reference the filters of the first step. Platform owners can adjust these files as needed for their platform instance. Figure 8 illustrates the structure of the files.

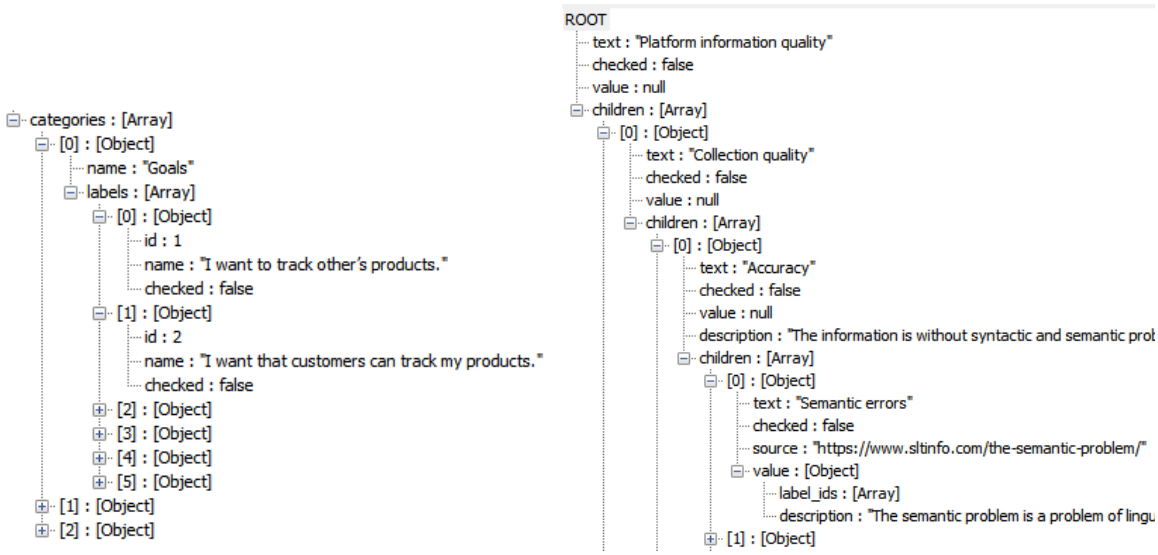

Figure 8: JSON structures for filter function (left side) and quality factors (right side)

The figure above illustrates that the statements below the filter categories act as labels. They have unique IDs. The quality factor description contains an array of label IDs that links them to statements.

5 Cyber threats to information quality and data integrity

Information quality and data integrity are among top priorities for model enterprises and their digital platforms. Maintaining information quality and data integrity is important for several reasons. For one, the accuracy of data that is ensured through data integrity increases stability and performances of the system. Similarly, information quality ensures reusability and maintainability. Data integrity ensures recoverability (e.g. critical in ransomware attacks) and searchability, for instance, searchable encryption ensures the availability of secret (encrypted) data (Wenjun and Zerong, 2017).

Information quality and data integrity can be compromised in a variety of ways. For example, data integrity can be compromised through the following:

- Unintentional and/or malicious human errors.
- Data compromised during transfer from one device to another, e.g. by intercepting data in transfer among two or more Microservices.
- Various cyber threats, bugs, viruses, hacking, including insider attack with data manipulation, etc.
- Compromised hardware, and more.

To prevent data integrity related hazards, best practices include data backup and data duplications. Input data validation is another important step that can preclude the entering of invalid data, error detection/ data validation to identify errors in data transmission. Best practices related to data security measures include data loss prevention, access control, data

encryption, and more. In case of an attack, the best practices for data recovery suggest the following:

- Recovery from trusted backups and snapshots.
- Rollbacks to a known good state of the working system.
- Effective recovery based on activity logging and monitoring, versioning and journaling file system, quick service restoration after the attack, and effective alerting system when the data is corrupted.

Hence, information quality and data integrity practices constitute an essential component of effective security protocols of digital platforms. In NIMBLE, cyber supply chains risks can be associated to a lack of visibility and control over many of the Microservices involved in the delivery of platform services. Furthermore, the associated threats can be either *adversarial* (e.g. tampering) or *non-adversarial* (e.g. low information quality). Vulnerabilities may be *internal* (organizational procedures) or *external* (e.g. part of a platform's supply chain ecosystem).

6 Conclusion

Awareness raising is an effective and comparably cheap IQM method to avoid information quality problems on NIMBLE platform instances. It relies on the users' knowledge and willingness to provide high quality contents on the platform. Providing high quality information should benefit trustworthiness of companies and improve their attractiveness as business partners. Areas of future improvement for the proposed platform IQM and QualiExplore mainly concern:

- Development and integration of IQM on programmatic level
- Development and integration of a self-assessment for platform IQM maturity analysis
- Development and integration of programmatic measures for IQM of IoT information
- Extension of the QualiExplore tree structure with more factors
- Identification of the links between data security and IQM
- Connection of the QualiExplore progress bar with the platform's user management

7 References

[1] Brook, C. (2019): *What is Data Integrity? Definition, Best Practices and More.* Available online at: https://digitalguardian.com/blog/what-data-integrity-data-protection-101

[2] Choo, Chun Wei (2002): *Information management for the intelligent organization*. The art of scanning the environment. 3. ed. Medford, NJ: Information Today (ASIS monograph series).

[3] Eppler, Martin (2006): *Managing Information Quality*. Increasing the Value of Information in Knowledge-intensive Products and Processes. 2nd ed. Berlin, Heidelberg: Springer.

[4] Liu, Liping; Chi, Lauren (2002): *Evolutional Data Quality: A Theory-Specific View*. In Craig Fisher, Bruce N. Davidson (Eds.): Proceedings of the Seventh International Conference on Information Quality. International Conference on Information Quality. MIT Sloan School of Management, Cambridge, MA, USA, November 8-10, 2002, pp. 292–304.

[5] Luo, Wenjun; An, Zerong (2017): *Searchable Encryption with Data Integrity Verification*. In Proceedings of the 2017 Asia-Pacific Engineering and technology Conference (APETC 2017). Online available from: http://dpi-proceedings.com/index.php/dtetr/article/view/11382/10926

[6] Papageorgiou, Grigorios; Grant, Stuart; Takkenberg, Johanna; Mokhles, Mostafa (2018): *Statistical primer: how to deal with missing data in scientific research?*, Interactive CardioVascular and Thoracic Surgery, Volume 27, Issue 2, pp. 153–158, https://doi.org/10.1093/icvts/ivy102

[7] Rowley, Jennifer (2007): *The wisdom hierarchy. Representations of the DIKW hierarchy*. In Journal of Information Science 33 (2), pp. 163–180. DOI: 10.1177/0165551506070706.

[8] Standard ISO/IEC 25012:2008, 2008: *Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model*.

[9] Stonebraker, Michael; Bruckner, Daniel; Ilyas, Ihab; Beskales, George; Cherniack, Mitch; Zdonik, Stan; Pagan, Alexander; Xu, Shan. (2019). *Data Curation at Scale: The Data Tamer System*.

[10] Wang, Richard Y.; Strong, Diane M. (1996): *Beyond Accuracy: What Data Quality Means to Data Consumers*. In Journal of Management Information Systems 12 (4), pp. 5–33. Available online at http://www.jstor.org/stable/40398176.

[11] Wilson, T. D. (2002): *The nonsense of 'knowledge management'*. In Information Research 8 (1), paper no. 144. Available online at http://InformationR.net/ir/8-1/paper144.html.