DIWA Report

Sub-Activity 4.4: Data Quality

Version: v1.0 final version, October 2023

Main author: Project team Masterplan DIWA

Contributing: viadonau – Österreichische Wasserstraßen-Gesellschaft mbH
Generaldirektion Wasserstraßen und Schifffahrt
Rijkswaterstaat
De Vlaamse Waterweg
Voies navigables de France

Co-funded by
the European Union

Main authors: Gert Morlion, De Vlaamse Waterweg
Eric Duchesne, BDO
Jürgen Helmer, BDO
Laurien Van den Heuvel, BDO
Marie-Claire Schug, viadonau

Contributing: Christoph Plasil, viadonau
Thomas Zwicklhuber, viadonau
Robert Schwarz, viadonau
Quirine de Kloet, Rijkswaterstaat
Martijn van Hengstum, Rijkswaterstaat
Jeffrey van Gils, Rijkswaterstaat
Therry van der Burgt, Rijkswaterstaat

Co-funded by
the European Union

Table of contents

Co-funded by the European Union

1 Executive summary

Activity 4 of the Masterplan Digitalisation of Inland Waterways (DIWA) project is focusing on 4 topics. Standardisation, legal and regulatory framework, cybersecurity and privacy and last but not least data quality.

Sub-Activity 4.4 of the Masterplan DIWA project identifies existing frameworks, tries to link data quality to existing standards, projects and guidelines and will provide pre-conditions and requirements on data quality management related to IWT services, systems, information and data.

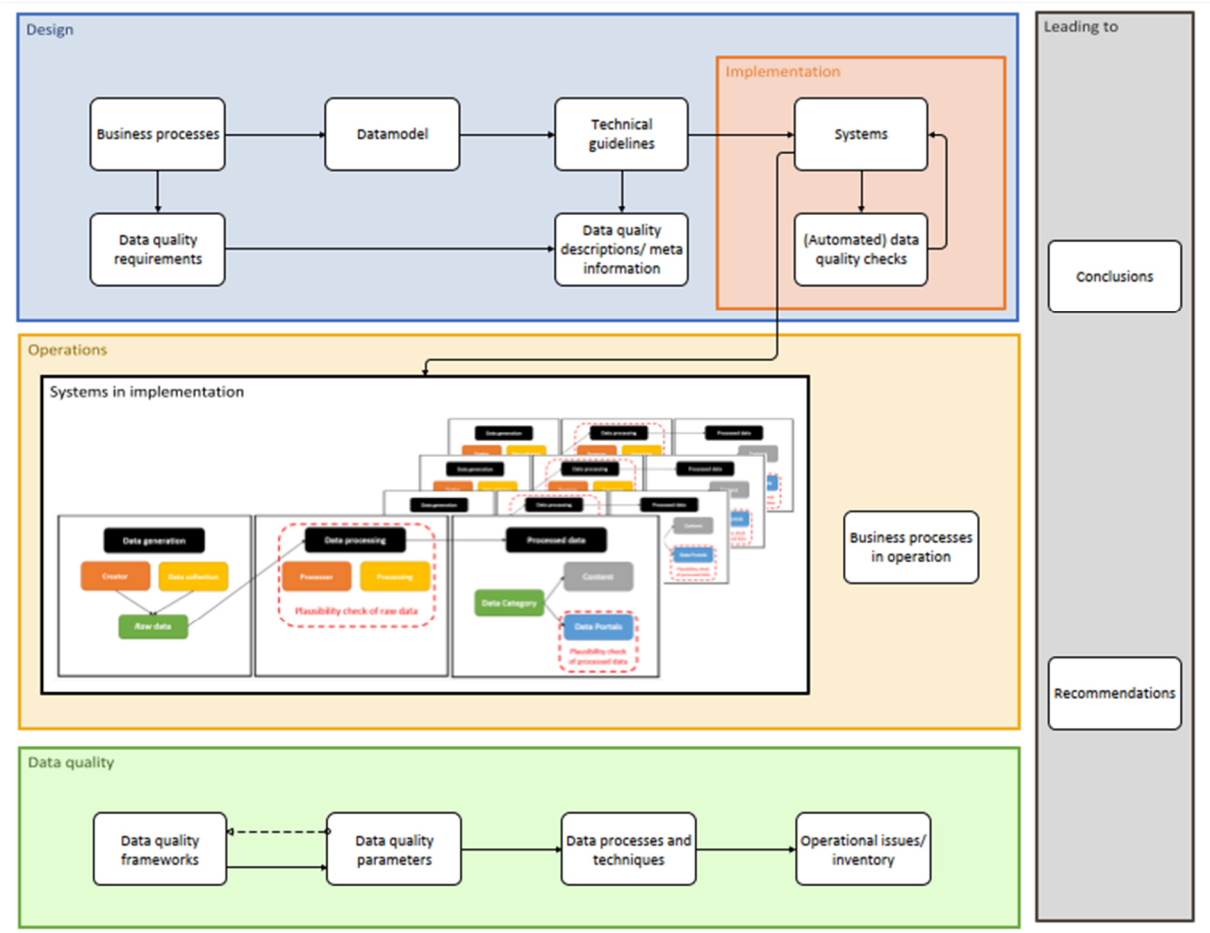

Figure 1 Reading guide for this report (see also Figure 2)

Definitions

Before explaining what data quality is all about, it is important to know what is meant by terms like 'data', 'information', 'quality', etc. There is a very thin line between 'data' and 'information'. While 'data' is defined as facts/figures without any meaning, 'information' is giving meaning to the data so it can be interpreted by e.g., humans.

Data quality is the extent to which data is suitable for the purpose for which it is used. Therefore, data should pursue all or some of the data quality parameters. Data quality will have an impact on the level of quality of services. Building a valuable framework fitted for Fairway Authorities and the data services is essential for improving the quality of the services.

Data quality management is about checking whether data is correct, complete and compatible with data provided by other systems while information quality is described as the quality of the information that is produced by systems and therefore the quality of the content of information systems.

Information management is the process of acquiring, organising, storing, and using information. The goal is to provide the right information based on high quality data. People and systems cannot make effective business decisions with faulty, incomplete or misleading information because it is based on incorrect data.

Misinformation can make the difference between a profitable voyage or a delayed trip. In some cases, incorrect data can make the difference between life and death. That is why effective data quality management is essential to any consistent data analysis, as the quality of data is crucial to derive actionable and - more important - accurate insights. Poor data quality causes problems for organisations. They must adopt good practices to improve data quality and reduce errors to prevent loss of business and produce satisfied customers. Every organisation depends on data to support its operations and ultimately its customers.

Data source types and processing concepts

When analysing which types of data sources exist, it is important to consider the (usage) purpose of the data. Depending on the purpose of the data for the user, other data sources can be used.

Data generation summarises the generation of new raw data generated by a creator such as the public administration, skippers or even vessels. Data collection describes the tools used to generate the raw data, e.g. an echo sounder to generate depth information of a waterway.
The newly generated raw data is then processed in a further step, for example to remove errors or outliers or to convert it into a suitable format. Here the term processor refers to the organisation that processes the data. Data processing refers to the tools or software used to process the data. During the processing of the raw data, an initial check or plausibility check of the data can take place.

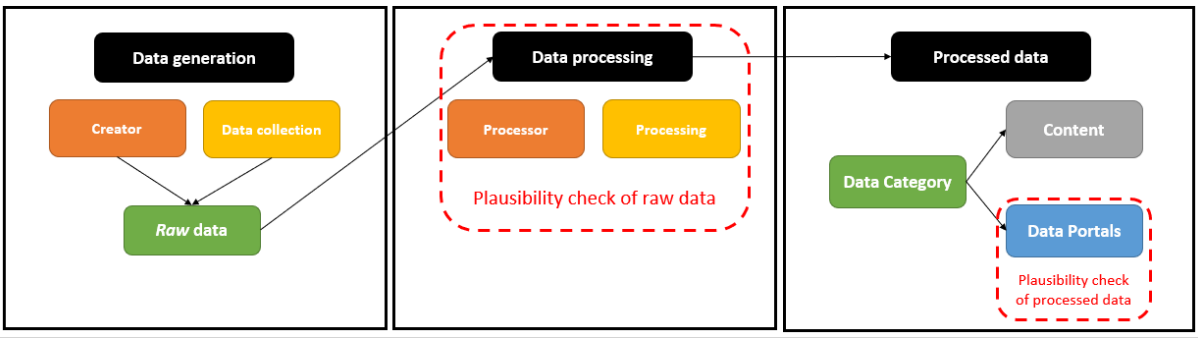

Figure 2 From Data generation towards processed data

The data source is usually included in the metadata of the data and can be found there. Metadata contains important information about the data itself. It plays a vital role in data quality, as it can be used to pass on information about various quality parameters to the data user.

Quality parameters

There are several parameters (or dimensions) to describe the level of quality of data. A subset of these parameters was identified and defined:

- Accessibility
- Accuracy
- Availability
- Completeness
- Comprehensiveness
- Consistency
- Currency
- Legitimacy
- Reliability
- Timeliness
- Unambiguous
- Uniqueness
- Validity

This list is not exhaustive and too long to be used in an easy manner for monitoring the data quality in all IWT related matters. Therefore, a selection was made of the most important parameters which resulted in accuracy, completeness, consistency, currency, timeliness, uniqueness and validity.

Quality frameworks

Describing the quality of data with these parameters is a first indication but how are you going to interpret the values? A data quality framework is a tool that you can use to not only measure the data quality within your organisation but also to define data quality goals and standards as well as the activities that must be taken to meet those goals. Existing frameworks are:

- Data quality framework of Rijkswaterstaat (2021)
- Total Data Quality Management (TDQM) (1998)
- Total Information Quality Management (TIQM) (1999)
- Cost-effect Of Low Data Quality (COLDQ) (2001)
- A Methodology for Information Quality Assessment (AIMQ) (2002)
- Data Quality Assessment (DQA) (2002)
- Hybrid Information Quality Management (HIQM) (2006)
- Comprehensive Methodology for Data Quality Management (CDQ) (2006)
- A Data Quality Practical Approach (DQPA) (2009)
- A Data Quality Methodology for Heterogeneous Data (HDQM) (2011)
- Data Quality Assessment Framework (DQAF) (2013)
- Task-Based Data Quality Method (TBDQ) (2016)
- The Observe-Orient-Decide-Act Methodology for Data Quality (OODA DQ) (2017)

All frameworks have their own way of dealing with parameters and processes. After comparing the different frameworks on the used parameters with the selected set, two frameworks were identified:

- **Cost-effect Of Low Data Quality**: In this framework, the following parameters were consistent with those selected for IWT: accuracy, completeness, consistency, currency, timeliness.
- **A Data Quality Practical Approach**: The data quality parameters that occur in this data quality framework are: accuracy, completeness, consistency, currency, timeliness, uniqueness. Only validity is missing this framework.

An important consideration has to be made! The two frameworks were selected based on the number of parameters they reflect compared with the selected subset above. However, this doesn't mean that other frameworks would not be applicable for IWT related data topics. Getting more information about the frameworks was a tough, nearly impossible job, and therefore the available parameters were the only criteria that could be used. Other criteria such as the relations between the data or the way it was processed were not taken into consideration.

Desk research

New techniques and data processes which could support the monitoring of data quality were examined through desk research.

1. Aggregation and anonymization:
   Aggregation and anonymization do not lead directly to increasing data quality but is closer linked to privacy instead. There is a trade-off between the protection of privacy and not losing too much information when anonymizing (or aggregating) the dataset. In case of poor quality, back tracing to the root cause is even more difficult than with pure data.

2. Management of big data:
   Overlooking errors within the data is much easier when working with big data. It is no longer possible to check data by hand, so you need to have the metadata of a dataset.

3. Process mining:
   Process mining can support in the improving of the quality of data by detecting flaws and outliers in the data. On the other hand, the principle "Garbage in, garbage out" is also applicable to process mining and can lead to misleading decisions.

4. Artificial intelligence:
   AI can be used to monitor and correct data in a reliable way, but a machine learning algorithm that uses irrelevant or faulty data as input, will not be able to solve tasks that become more and more complex. Therefore, it is critical to pre-process datasets before using them to train a machine learning model.

5. Semantic modelling[1]:
   Especially when data has to be shared amongst different transport modes (road, rail, inland waterways, maritime transport, air, hyperloop, …) it is difficult to establish dedicated syntactical mappings from one format to another. It is there that semantics can be of use (e.g., already proven in the elaboration of "smart cities"). Although semantic modelling can bring different worlds together, care should be taken on the influence on the data quality. Namely the quality of the resulting data and the derived information is strongly dependent on the mapping algorithms between the different domains. The governance on the definitions on an atomic data level used within the different domains, where the automated mappings are based on, is of utmost importance to get satisfactory and trustworthy results.

6. Data sharing versus data exchange:
   Instead of sharing data by copying it, is also possible to share the link to the source of the data. Since the data is maintained at the source, one could expect that the quality is better when using this method (up-to-date data and with less data loss than via the traditional data exchange chain). However, availability of the different parts of the data puzzle that can be scattered over multiple data bases and networks becomes more crucial than ever.

[1] According to Klas and Schrefl (1995), the "overall goal of semantic data models is to capture more meaning of data by integrating relational concepts with more powerful abstraction concepts known from the Artificial Intelligence field. The idea is to provide high level modeling primitives as an integral part of a data model in order to facilitate the representation of real world situations" source : Wolfgang Klas, Michael Schrefl (1995). "Semantic data modeling" In: *Metaclasses and Their Application*. Book Series Lecture Notes in Computer Science. Publisher Springer Berlin / Heidelberg. Volume Volume 943/1995.

IWT related topics

IWT is working with a set of different topics, platforms, standards, … which are all using and depending on data. During the desk research these commonly known topics are examined on the quality aspect. *Are the requirements on the level of data quality described and if so, how is this described?*
The topics that are examined are:

- RIS COMEX (EuRIS, CEERIS)
- eRIBA
- Inland ECDIS
- RIS Guidelines 2019

There are many differences between the IWT related topics. Quality of data is described in different wordings and in most cases on a very general level. Requirements on e.g., accuracy are not described in units and/or values, so further investigation is needed for each operational process.

Results and conclusions

The most important conclusion of this research is that data quality is and remains most important for inland navigation and data exchange / data sharing. If the data quality is poor, analyses based on the data are unusable. For further digitalisation in inland navigation, data quality will play a key role for the necessary further technological developments and e.g., Smart Shipping. Therefore, to check the data quality, it is important to make use of the data quality parameters in IWT as researched to ensure that the used data, meet the associated parameters.

The quality framework includes the definition of the overall set of parameters and their values, mechanisms and guidelines aligned to the implementation of new business and technical services and their intended quality.

Because of the wide range of IWT related applications, a broad range of data quality frameworks can be used. It is impossible to assign one particular framework as 'the data quality framework for IWT'. However, using one is needed for good data quality in business processes.

Not knowing whether the used data is correct, accurate and complete leads to specific high risk issues that IWT is facing today. This is a high risk because incorrect data can lead to human, material and/or infrastructure damage. Therefore, it is important to always take a look at the data quality parameters that are expanded in this report and to implement them.

Smart shipping and autonomous navigation will require a higher or different data quality than is currently possible. To complete the above list, synchro modality and digital twins will also require a more robust data quality (framework).

Depending on the purpose of the data for the user or for other systems, other quality requirements can be in place. By consequence, when the purpose of a certain data element changes, also the requirements on that data element must be checked whether they have to change too.

Besides these main conclusions, a lot of recommendations were identified in chapter 8. Recommendations regarding further investigations, meta data, governance, … and all concerning steps towards a better and higher quality of data.

2 Introduction

Data quality is an important precondition for the digital transformation in IWT. In the past a process of digitising, digitalisation and harmonisation of data has been executed to provide IWT stakeholders with reliable data services that can be used on the European waterways. For more advanced data services like route and travel planning and in the future Smart Shipping operations, data quality becomes even more important.

The requirements with respect to data quality have to ensure an agreed level of quality of the services. Additionally, a quality mechanism has to provide a means to monitor and evaluate the quality level and, if needed, take measures for the improvement of the data quality leading to a potential increased quality of the services. A study is needed to define the quality framework for data quality in IWT.

This quality framework includes the definition of the overall set of parameters and their values, mechanisms and guidelines aligned to the implementation of new business and technical services and their intended quality.

The study in SuAc4.4 will include the definition of a quality framework and the specification of quality parameters such as:

- Quality parameters that affect the functional suitability.
- Quality parameters that affect the performance.
- Quality parameters that affect the reliability.

This document will address the importance of using a data quality framework and parameters to pursue high data quality. This report provides recommendations for the Masterplan DIWA.

3 Objectives of SuAc 4.4 Data Quality

The objectives, tasks and expected results for this Sub-Activity are outlined in the following subchapters.

3.1 Objective

The objective of SuAc 4.4 is to provide pre-conditions and requirements on data quality management related to IWT services, systems, information and data to be taken in account in the context of the Digitalisation of Inland Waterways.

3.2 Tasks

Following tasks were identified in order to meet the objective of SuAc 4.4:

- Make a study on the Quality Framework needed for, and based on, the business developments and technological developments as specified in activities 2 and 3 and define the effects on the digital transition in the period 2022-2032.
  The party responsible for this is the SuAc leader DVW.

- Draft the report (study) on the Quality Framework with pre-conditions and requirements on data quality management measures to be taken in relation to the Masterplan Digitalisation of Inland Waterways.
  The party responsible for this is the SuAc leader DVW.

3.3 Expected Results

This SuAc will make the distinction between data and information. Both terms are often used at the same time while they mean something differently. A clear definition is important certainly when the identification of the different types of data sources and processes will be one of the results of this activity.
What are the parameters that describe the quality of data and which of these parameters are important when talking about IWT? The SuAc will recommend a limited set of parameters.
Taking this into account, existing platform, guidelines, regulations, standards and projects will be checked on the requirements regarding data quality.

1. Are there big differences between these IWT related topics?
2. Which existing technologies and/or processes can be used for describing the quality of data?
3. Can they be used to monitor the status of the quality?
4. Can they improve the data?

In the end, an intermediate report (study) on data quality will be written.

This SuAc has several aims:

- Desktop research on existing references
- Study on the most important parameters
- Identify existing data quality frameworks
- Examine data quality references in existing topics of IWT?
- Make recommendations for the future

4 Work approach

A work approach is implemented to give an overview of the different meetings and brainstorm sessions during the past months.

4.1 Timeline

4.2 Work approach

During the first meetings of the Sub-Activity, we discussed the scope of our research. The conclusion was that the focus of Sub-Activity 4.4 is on Data Quality Management and not on service quality (IWT services and systems). Once our scope was clear to all members, we began to further elaborate on the various data quality parameters, how to monitor data quality, and why it is so important. Meanwhile, we also distributed the components of the desk research among members. In later meetings, we discussed the available desk research done by members of the Sub-Activity.

In further meetings, we organised two interactive brainstorming sessions. The first session focused on the current situation, more specifically on the current data quality gaps in IWT, their impact and solutions to resolve these deficiencies. The second session dealt with developments for the future requiring higher data quality than is currently available. These two sessions were designed to get a better picture of the data quality problems in IWT and the future developments that will require higher/different data quality. Based on the required input, we set up an impact matrix of the current situation.

After all these meetings, we started writing on this report using all relevant input of the previous meetings.

4.3 Interdependencies with other sub-activities

SuAc 2.1: Smart shipping
Sub-Activity 2.1 identifies the following needs for Smart Shipping to support increasing automation levels of vessels:

- Increase the quality of the data by investing in quality of existing data instead of a focus on sharing new types of data. A solution might be to build a digital twin of the waterway with the possibility for users to add or suggested changes.
- Need for more clarity on the quality (meta data) of existing data. This allows users to verify on critical functional parameters.

SuAc 2.1 also refers to PIANC WG 210 report where data quality indicators: Availability, Completeness and Accuracy are investigated with respect to Smart Shipping.

SuAc 2.2: Synchromodality
SuAc 2.2 does not explicitly mention data quality as a topic.

SuAc 2.3: Port & terminal information service
SuAc 2.3 considers data quality improvement as an important effect of increasing digital information exchange between skippers, terminals and authorities.

SuAc 2.4: RIS enabled corridormanagement
SuAc 2.4 identifies poor data quality as a risk for the success and adoption of RIS enabled corridormanagement and subsequently data quality improvement in the areas of data consistency/quality checks and improvement of national data acquisition as a prerequisite for future actions. EHDB data (outdated, incomplete) is specifically mentioned as data with issues.

SuAc 3.1: New technologies
The New technologies draft report stresses the importance of data quality in the context of Big Data (veracity). Poor data means a high risk of biased or incorrect analyses.

SuAc 3.2: IWT connectivity platform
SuAc 3.2 Identifies connectivity platforms such as EuRIS and European Mobility Data Spaces Initiatives as a means to improve the availability, quality and interoperability of data on multinational level – both in domain-specific settings and across sectors.

SuAc 3.3: Smart sensoring & PNT
SuAc 3.3 calls specific attention to data quality of vessel position related data.

SuAc 3.4: Information model & data registry
SuAc 3.4 reiterates the observation from SuAc 2.4 about poor data quality of the EHDB reference data and extends this to ERDMS (not up to date due to synchronisation issues). EuRIS is found to exhibit a high(er) level of data quality.

SuAc 3.5: Technology in other transport domains
The SuAc 3.5 report states that data quality is a critical characteristic when trying to achieve higher digitalisation levels.

5 Data Quality: definitions and frameworks

5.1 Introduction

To explain data quality and why it is so important, the difference between data and information must be made clear first. The terms "data" and "information" are often used interchangeably, but they are not the same. There are subtle differences between these components and their purpose. "Data" is often defined as facts/figures, while "information" is the organisation and interpretation given to those facts. Data consists of facts/figures without meaning attached to those facts/figures, while information consists of data placed in a certain context, it is the organisation and interpretation given to those facts/figures.

As an example, the data as water levels and bridge height can be combined with the height of his vessel (Figure 3) and other considerations which leads to the information and a decision for a skipper if he is able to pass a bridge.

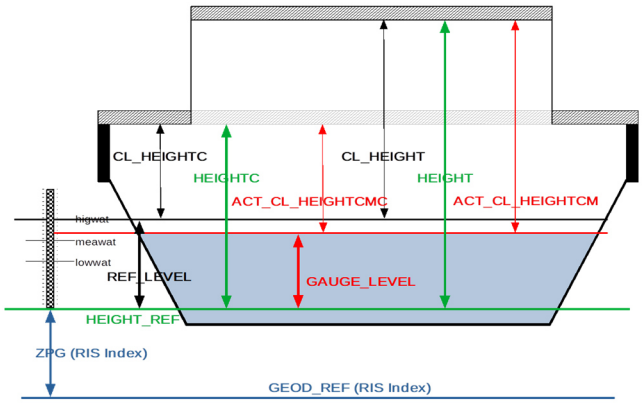

Figure 3 Several data elements that can lead to information

Another example which makes the difference between data and information clear is the route planning advice given by platforms as EuRIS. The platform is fed with all kinds of data: sizes of a vessel, the size and properties of locks, bridges and operation times. This data is used to calculate a route for the vessel taking all these different data elements into account. Once done the platform delivers the information as a result to the user.

5.1.1 What is data quality, data quality management and information quality?

Data quality is the extent to which data is suitable for the purpose for which it is used. Therefore data should pursue all or some of the data quality parameters (see further).
The quality of data will have an impact on the quality of services (see results in Activity 2 report). An adequate framework and analysis will increase the quality of data which will lead to an improvement of the level of services. The reverse - the expected quality of services will determine the needed quality of data and measures - is also correct because data quality is important to quantify the quality of a service. Building a valuable framework fitted for Fairway Authorities and the data services is essential for improving the quality of these services.

Data quality management is about checking whether data is correct, complete and compatible with data provided by other systems as well as taking care of correct processing and complying with data standards, data transfers and avoiding data loss in sub-systems. To summarize, it is having the right data at the right time and place, for the right people or systems to provide correct and complete services and optimal performance.

Information quality is "the quality of the information that the systems produces" (DeLone & McLean, 1992). It is the quality of the content of information systems. Reflecting this on the customer of information, Gustavsson and Wänström (2009) define information quality as the "ability to satisfy stated and implied needs of the information consumer". Here customer and consumer of information refer to the user, so the user influences information quality (Naumann & Rolker, 2000).

5.1.2 Why is data and information quality management important?

Information management is the process of acquiring, organising, storing, and using information. The goal is to provide the right information based on high quality data. People cannot make effective business decisions with faulty, incomplete or misleading information because it is based on incorrect data. People need information they can trust to do the job most effectively.

The right information derived from data can help people make long-term and short-term plans based on accurate and reliable data. It can help shippers make timely decisions to improve performance and operating efficiency and even prevent accidents and calamities.

Misinformation can make the difference between a profitable voyage or a delayed trip. In some cases, incorrect data can make the difference between life and death. That is why effective **data quality management** is essential to any consistent data analysis, as the quality of data is crucial to derive actionable and - more importantly - accurate insights. Poor data quality causes problems for organisations. They must adopt good practices to improve data quality and reduce errors to prevent loss of business and produce satisfied customers. Every organisation depends on data to support its operations and ultimately its customers.

A poor data quality can be seen as unreliable. Affected stakeholders will no longer trust the digital services provided by the fairway authorities and will stop using the offered services. This constitutes loss of a substantial investment of taxpayer's money and obstructs efficiency/effectiveness goals fairway authorities are attempting to achieve through digitalisation initiatives (such as, but not limited to: safety of navigation, reduction of waiting times, optimal sailing speed to reduce GHG emissions, attractiveness of IWT as a modality, etc.).

5.2 Data source types – Data processing concept

Before using data in IWT it is important to know how the data is collected or retrieved. Is it raw data coming from a sensor, or is it copied from a digital platform? In the first case, the data possibly must be processed before using, in the latter case, the user should be able to rely on the data. Knowing where the data is coming from (source) and what happened with it (processing) can be important indications of the quality of the data.

When analysing which types of data sources exist, it is important to consider the (usage) purpose of the data. Depending on the purpose of the data for the user, different data sources can be used for the same data.
So, for example, if the user wants to display Inland ENCs on his website, the user will use the various national Inland ENCs as a data source and not the raw data required to generate the ENCs, as the ENC data are based on Inland ECDIS guidelines. However, if the user wants to create his/her own user-defined service, such as a route planner, the user will probably request access to the raw data, in this case the national reference data or separate files like Aids to Navigation.
Depending on the purpose of the data use, the data source changes, which is dependent on the data processing process, e.g., from data generation and data processing to the final processed data. An overview of this process is shown schematically in **Fehler! Verweisquelle konnte nicht gefunden werden.**.
In addition, the figure also indicates at which point in the data processing plausibility checks are carried out on the data in order to ensure data quality.

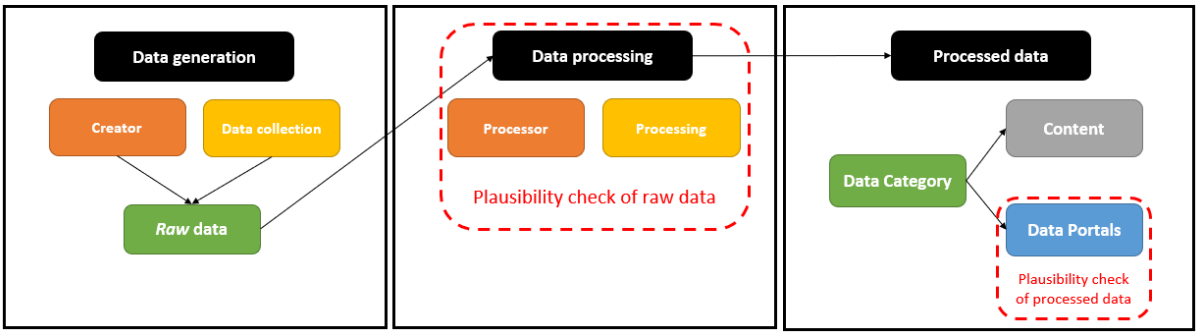

Figure 4: Data source types - Data processing concept

Data generation summarises the generation of new raw data generated by a **creator** such as the public administration, skippers or even vessels. **Data collection** describes the tools used to generate the raw data, e.g. an echo sounder to generate depth information of a waterway.
The newly generated raw data is then **processed** in a further step, for example to remove errors or outliers or to convert it into a suitable format. Here the term **processor** refers to the organisation that processes the data. Data **processing** refers to the tools or software used to process the data. During the processing of the raw data, an initial check or **plausibility check** of the data can take place.
The resulting **processed data** is finally fed into different **data or information categories** depending on the **content** to create the final products such as Inland ENCs, Notices to Skippers or similar. In most cases, the data published on **data portals** on the web are subject to further **data checks**.
Consequently, the verification of the data and thus of the data quality does not take place in the context of raw data generation/collection, but only in the first or second processing step.

"Annex 3: Types of data sources" contains the most important data sources for the operation and maintenance of the waterway and navigation on it. The structure of the overview presented in that Annex will be explained here using the Inland ENC charts as an example.

Processed data
Inland ENCs contain **data** concerning the fairways (e.g. permitted dimensions), bathymetric data as well as other information, such as navigation marks or berths, etc. Many European Inland ENCs are made available online on the **data portals** D4D Portal or EuRIS, for example.

Data Processing
Inland ENCs are produced by the **national authorities** of the respective countries using different **data processing tools** such as GIS software and surveying.

Data generation
The **raw data**, such as the echo-sounder point clouds, are collected by different **measuring instruments**, e.g. by an echo-sounder. In the case of Inland ENCs, the client or **data creator** of the data collection is the national authority.

As already mentioned, the data sources can therefore vary depending on the purpose of use:

- Data portals (for already processed data)
  - D4D Portal
  - Danube FIS Portal
  - EuRIS
  - CEERIS (reporting requirements), NaMIB
  - ECDIS, DoRIS, ELWIS, Vaarweginformatie.nl, VisuRIS.be
  - National Weather Institute sites
  - Etc.
- Data collection: / generation:
  - Provided by sensors or surveys
    - Bathymetry by single or multibeam echo sounders
    - Bathymetry/topographic by LiDAR, ...
    - Satellite imagery
    - Position information provided via Global Navigation Satellite System (GNSS) such as GPS or Galileo

        - Water level/depth measurements
        - Meteorological data
        - Radar data
        - Camera images
    - Mainly provided by human input, often supported by automated processes
        - Electronic Reporting International (ERI) messages
        - Notices to Skippers (NtS) - fairway related messages
        - Facility files
        - Reference network
        - RIS Index
    - Etc.

The data source is usually included in the metadata of the data and can be found there.
Metadata contains important information about the data itself. It plays a vital role in data quality, as it can be used to pass on information about various quality parameters to the data user.

5.3 Data quality parameters and frameworks

There are countless standards that can be used for data quality, e.g. ISO 9126 (describes quality model for software products), ISO2500 "Software Product Requirements and evaluation (SQuaRE)", ISO 8000 "Data Quality" and ISO 19100 "Geographic Information".

Likewise, there are countless quality parameters, not all of which can be interpreted in the context of DIWA for individual data services related to inland navigation. Therefore, a selection has to be made which quality parameters are relevant for the future and in particular for digitalisation and should therefore be considered and investigated in more detail within the framework of this Sub-Activity.

Similar work has already been carried out within the framework of IRIS Europe II and III. It is to be investigated whether the requirements for the quality parameters are shifting with regard to digitisation, and thus, building on already existing work, to supplement these if necessary. This could be done e.g. in a future European project dealing with the implementation of EuRIS/CEERIS.

5.3.1 Parameters

To have good **data quality,** the use of data quality parameters - also called dimensions - can foster the data quality.
In what follows, a brief overview of a subset of data quality parameters identified during this study is expanded with their definitions:

Accessibility: Extent to which information is available, or easily and quickly retrievable

Accuracy[2]: Does the data fit the defined range? Accuracy refers to the exactness of the values in the various fields of a data element, dataset or database. It must be within a certain ranges.

- Data are accurate when data values stored in the database correspond to real-world values
- The extent which data is correct, reliable and certified
- Accuracy is a measure of the proximity of a data value, v, to some other value, v', that is considered correct
- A measure of the correction of the data (which requires an authoritative source of reference to be identified and accessible

Availability: Extent to which information is physically accessible.

[2] Positional accuracy: "The accuracy of the position of features within a spatial reference system."[ISO 19157:2013(E) 7.3.4 Positional accuracy].
Thematic accuracy: "the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships."
[ISO 19157:2013(E) Clause 7.3.5 Thematic accuracy]

Completeness[3]**:** How complete is the data(set)? Completeness is a criterion for determining whether all required data is currently available in a given data element, data set or database.

- The ability of an information system to represent every meaningful state of the represented real-world system.
- The extent to which data are of sufficient breadth, depth and scope for the task at hand.
- The degree to which values are present in a data collection.
- Percentage of the real-world information entered in the sources and/or the data warehouse.
- Information having all required parts of an entity's information present.
- Ratio between the number of non-null values in a source and the size of the universal relation.
- All values that are supposed to be collected as per a collection theory.

Comprehensiveness: the state or condition of including all or nearly all elements or aspects of something.

Consistency[4]**:** Is the data synchronised between different systems? It Indicates whether the same information stored and used in different instances matches. Consistency is a criterion for observing whether values in different databases and data sets match. The data must be consistent with specific, applicable standards and regulations.

- The extent to which data is presented in the same format and compatible with previous data.
- Refer to the violation of semantic rules defined over the set of data.

Currency:

- Currency is the degree to which a datum is up to date. A datum value is up to date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value.
- Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real-world.

Legitimacy: Conformity to the law or to rules.

Reliability: The degree to which the result of a measurement, calculation, or specification can be trusted to be accurate.

- Extent to which information is correct and reliable.
- It is the capability of the function to maintain a specified level of performance when used on specified condition.

Timeliness: A criterion for estimating the exactness of the data over time. The data must be true with respect to the needs of the activities, this especially in real time environments, where timely refresh of current data is an important aspect:

- The extent to which age of the data is appropriated for the task at hand.
- Timeliness refers only to the delay between a change of a real-world state and the resulting modification of the information system state.
- Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long ago it was recorded. Volatility is a measure of information instability the frequency of change of the value for an entity attribute.

Unambiguous: Not open to more than one interpretation; having one obvious meaning.

---

[3] The presence and absence of features, their attributes and relationships." [ISO 19157:2013(E) Clause 7.3.2 Completeness]

[4] Logical consistency: "the degree of adherence to logical rules of data structure, attribution and relationships." [ISO 19157:2013(E) 7.3.3 Logical consistency]
Format consistency: "the degree to which data are stored in accordance with the physical structure of the data set." [ISO 19157:2013(E) Annex I.4.2.4 Format consistency]
Topological consistency: "the correctness of the explicitly encoded topological characteristics of a data set." [ISO 19157:2013(E) 7.3.3 Logical consistency]
Conceptual consistency: "the measurement of how well the data set conforms to rules of the conceptual schema, itself. If the conceptual schema explicitly or implicitly describes rules, these rules shall be followed." [ISO 19157:2013(E) Annex D.3.1 Conceptual consistency]
Temporal consistency: "a measure of the correctness of the order of events within data values."
[ISO 19157:2013(E) Annex D.5.2 Temporal consistency], [I.4.4.3 Temporal consistency – correctness of the order of events], [ISO 8000-8:2015(E) Annex B Syntactic quality rules]

Uniqueness: There are no duplicates within your data. It indicates whether it is a single registered instance in the data set or database.

Validity[5]**:** Is the data a correct representation of the element it describes? The value attributes are available for alignment with the specific requirement. Any invalid data will degrade the completeness of the data.

Defining all these parameters doesn't necessarily make it easier to check and monitor the data quality. Combining and classifying the parameters could help the users deciding which parameters to use and measure. Looking at the functionality, performance level and reliability of the data, a first classification can be made:

Functional suitability	Performance	Reliability
Consistency Unambiguous Uniqueness	Accuracy Currency Timeliness	Accessibility Availability Completeness Legitimacy Reliability

5.3.2 Selected data quality parameters for IWT

We filtered the above data quality parameters, because some of them will be important for the digitalisation of the IWT more specifically to improve IWT data quality and prevent future data errors. We have identified these parameters based on the most recurrent parameters that we faced in our study. This was also a question towards SuAc 4.4 from SuAc 3.3: Smart sensoring and PNT.
These data quality parameters are:

- Accuracy
- Completeness
- Consistency
- Currency
- Timeliness
- Uniqueness
- Validity

5.3.3 Data quality frameworks

A data quality framework is a tool that you can use to measure data quality within your organisation. With a data quality framework, your business can define its data quality goals and standards as well as the activities you are going to take to meet those goals[6]. In what follows, a broad overview of several existing data quality frameworks is given with short explanation:

Data quality framework of Rijkswaterstaat (2021)
This framework has been developed according to international standards (ISO 8000, 25012, 19157) only available in Dutch, see Annex 1: "Datakwaliteitsraamwerk hét naslagwerk" for more details.

Total Data Quality Management (TDQM) (1998)[7]
The TDQM methodology has been shown to be effective for improving Information Product (IP), particularly when top management has a strong commitment, as expressed in the organisation's Information Quality (IQ) policy. Organisations must harness the full potential of their data in order to

[5] Temporal validity is a measure of conformance of date and time values to formats specified in the conceptual schema, as well as the validity of values when compared to natural time (correct number of days in a month, up to 24 hours in a day, etc).
[ISO 19157:2013(E) Annex D.5.3 Temporal validity], [I.4.4.4 Temporal validity – validity of data with respect to time], [ISO 8000-8:2015(E) Annex B Syntactic quality rules]

[6] acceldata.io

[7] Richard Y. Wang. 1998. A product perspective on Total Data Quality Management

gain competitive advantage and attain strategic goals. The TDQM methodology has been developed as a step to meeting this challenge.

Total Information Quality Management (TIQM) (1999)[8]
The TIQM methodology focuses on the management activities that are responsible for the integration of operational data sources, by discussing the strategy that has to be followed by the organisations in order to make effective technical choices. Cost-benefit analyses are supported from a managerial perspective. The methodology provides a detailed classification of costs and benefits.

Cost-effect Of Low Data Quality (COLDQ) (2001)[9]
The fundamental objective of the COLDQ methodology is to provide a data quality scorecard supporting the evaluation of cost-effect of low data quality. Similar to TIQM, the methodology provides a detailed classification of costs and benefits. Direct benefits are obtainable from the avoidance of poor quality costs due to the adoption of improvement techniques. The goal is to obtain a quantitative assessment of the extent to which business processes are affected by bad information.

A Methodology for Information Quality Assessment (AIMQ) (2002)[10]
The AIMQ methodology as a whole provides a practical IQ (Information Quality) tool to organisations. It has been applied in various organisational settings, such as financial, healthcare and manufacturing industries. The methodology is useful in identifying IQ problems, prioritizing areas for IQ improvement, and monitoring IQ improvements over time.

Data Quality Assessment (DQA) (2002)[11]
The DQA methodology has been designed to provide the general principles guiding the definition of data quality metrics. The DQA methodology is aimed at identifying the general quality measurement principles.

Hybrid Information Quality Management (HIQM) (2006)[12]
Hybrid Information Quality Management (HIQM) methodology is conceived to be a support to solve run-time data quality problems. The analysis of the business processes and context in the design phase allows identifying critical points in the business tasks in which information quality might be improved. In these points, information quality blocks have to be inserted in order to continuously monitor the information flows. Through suitable checks, failures due to information quality problems can be detected. Furthermore, failures and warnings in service execution may depend on a wide variety of causes. Along the causes, the methodology also produces a list of the suitable recovery actions for a timely intervention and quality improvement.

Comprehensive Methodology for Data Quality Management (CDQ) (2006)[13]
The main aim of the CDQ methodology is the integration and enhancement of the phases, techniques and tools proposed by previous approaches. In particular, the CDQ methodology is conceived to be at the same time complete, flexible and simple to apply. Completeness is achieved by considering existing techniques and tools and integrating them in a framework that can work in any organisation. The methodology is flexible, since it supports the user in the selection of the most suitable techniques and tools within each phase and in any context.

A Data Quality Practical Approach (DQPA) (2009)[14]
The DQPA provides seven different steps for applying a data quality assessment. In the first step, useful data quality properties are identified for the assessment. Then, existing metrics are analysed about their suitability to provide unbiased, user-independent evaluations of data quality aspects. In the third step, methods to represent, interpret and assess data quality indicators are described. The notion of

[8] Carlo Batini, Cinizia Cappiello, Chiara Francalanci, Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement

[9] Carlo Batini, Cinizia Cappiello, Chiara Francalanci, Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement

[10] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y.Wang. 2002. AIMQ: a methodology for information quality assessment

[11] Carlo Batini, Cinizia Cappiello, Chiara Francalanci, Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement

[12] Cinzia Cappiello, Paolo Ficiaro, Barbara Pernici. 2006. HIQM: A Methodology for Information Quality Monitoring, Measurement, and Improvement

[13] Carlo Batini, Federico Cabitza, Cinizia Cappiello. 2008. A comprehensive data quality methodology for web and structured data

[14] Corinna Cichy, Stefan Rass. 2019. An Overview of Data Quality Frameworks

data lineage is regarded as an important aspect of this model and crucial to the process. In the fourth step, quality scores of primary data sources are estimated and stored as metadata. Then, the derived data is assessed in the fifth step. In step six, the data quality is analysed either by selecting the best data sources before the query execution based on its quality scores or by comparing data quality aggregated scores that correspond to different query plans for the same business question. Finally, in the seventh step, data sources are ranked according to the data quality stores and priorities provided by the user. The DQPA further makes use of a data lineage algorithm with a conflict resolution function for tracing back towards providing more information on the data quality.
The DQPA underlines the importance of data quality prevention, correction costs as well as cost effectiveness.

A Data Quality Methodology for Heterogeneous Data (HDQM) (2011)[15]
The main idea underpinning HDQM is to map the information resources used in an organisation to a common conceptual representation and then to assess the quality of data considering such homogeneous conceptual representation.

Data Quality Assessment Framework (DQAF) (2013)[16]
The DQAF provides a structure for assessing data quality by comparing country statistical practices with best practices, including internationally accepted methodologies. Rooted in the United Nations Fundamental Principles of Official Statistics,3 it is the product of an intensive consultation with national and international statistical authorities and data users inside and outside the Fund. It focuses on the quality-related features of governance of statistical systems, core statistical processes, and statistical products. Under the DQAF, assessments have a six-part structure starting with a review of the legal and institutional environment (prerequisites of quality) and followed by an analysis of five dimensions of quality.

Task-Based Data Quality Method (TBDQ) (2016)[17]
TBDQ is mainly a process-driven DQ method which specially assists organisations in which people play a significant role in the creation and manipulation of data directly or indirectly.
The task-based DQ method (TBDQ) is most appropriate for small and medium organisations, and simplicity in implementation is one of its most prominent features.

The Observe-Orient-Decide-Act Methodology for Data Quality (OODA DQ) (2017)[18]
The OODA DQ methodology refers to the use of existing data quality metrics and tools for measurement.
The OODA DQ methodology proposes a rather different approach to structuring the data quality improvement process. This phase of the framework comprises the remaining steps of its iterative cycle, i.e., Orient, Decide and Act. The Orient phase includes a root cause analysis that should be performed by a data governance team as well as the assessment of the severity of the previously identified data quality issues. Decisions ranging from data cleansing to modifications in application systems are the main concern in the Decide phase of the process. The decisions can be on a tactical as well as on an operational level and also include decisions regarding the number of people needed for fixing the issues in an appropriate manner. Finally, the Act phase is where identified actions are performed, implemented and validated.

5.4 An overview with all parameters and components of the above data quality frameworks, can be found in Annex 2: Data quality frameworks

.

[15] Carlo Batini, Federico Cabitza, Simone Grega, Daniele Barone. 2011. A Data Quality Methodology for Heterogeneous Data

[16] International Monetary Fund. 2003. Data Quality Assessment Framework and Data Quality Program

[17] Reza Vaziri, Mehran Mohsenzadeh, Jafar Habibi. 2016. TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement

[18] Corinna Cichy, Stefan Rass. 2019. An Overview of Data Quality Frameworks

5.4.1 Selected data quality frameworks for IWT

To decide which data quality frameworks are applicable for IWT, we relied on the selected parameters for IWT that are reflected in the data quality framework.

- **For Cost-effect Of Low Data Quality**: In this framework, the following parameters were consistent with those selected for IWT: accuracy, completeness, consistency, currency, timeliness
- **A Data Quality Practical Approach:** The data quality parameters that occur in this data quality framework are: accuracy, completeness, consistency, currency, timeliness, uniqueness. Only validity is missing in this framework

An important consideration has to be made. The two frameworks were selected based on the number and type of parameters they reflect. However, this doesn't mean that other frameworks would not be applicable for IWT related data topics. Getting more information about the frameworks was a though, nearly impossible job, and therefore the available parameters were the only criteria that could be used. Other criteria as there are the relations between the data, the way it was processed, … were not taken into consideration.

6 Results from desk research

6.1 Data processes and techniques

The evolution and development of new techniques is going on rapidly. This also means that with these new developments come new opportunities to monitor and control data quality. Some of these techniques can support the user of data in monitoring its quality and quality parameters.
The following paragraphs will examine some of the new techniques and data processes through desk research.

6.1.1 Aggregation and anonymization

Anonymization of data has become more important as it makes it possible to still use data that is protected by the GDPR. Sometimes there is a sufficient reason to be able use data in its original form, in other situations you are required to use an anonymisation technique or you won't be allowed to use the dataset.

There are different ways to execute the anonymisation, for instance you can simply remove the fields that contain the data specifically protected by the GDPR. Sometimes this is already enough, but in other situations it is then still possible to trace the data back to the original record. Removal of identifying characteristics is not enough to anonymise information, when there are external sources available that make it possible to restore these characteristics.

Aggregation is another method that can be used in order to create an anonymised dataset. When aggregating data you cluster individual records into groups, for each group you can look at the average/minimum/maximum/count or other values. As an example, it is possible to show the number of vessels within an area, without revealing exact location and details of the individual vessels.
It is important that clusters are meaningful and not too big or too small, big clusters will mean loss of a lot of information and small groups give outliers a lot of influence.

Aggregation and anonymization do not lead directly to increasing data quality but are closer linked to privacy instead. There is a trade-off between the protection of privacy and not losing too much information when anonymizing (or aggregating) the dataset.

After aggregation or anonymization, it is possible that data can no longer be combined with other datasets. When combined data from multiple sources is needed, it is sometimes necessary to combine the data before aggregating or anonymizing it. Once a dataset is aggregated or anonymized, it is no longer possible to go back to the original dataset. This also makes it hard to trace back the effect that

errors in the original dataset have on the resulting aggregated or anonymized dataset and by consequence the effect on the data quality.

6.1.2 The management of big data

Aggregation is also a great tool to make the management of big data easier, as it reduces the amount of data. It can be used to increase the data quality, averaging will smoothen out the outliers. However depending on the severity of outliers and the size of a group these outliers can still have a huge influence. When looking at the minimum or maximum outliers can result in an outlier in the aggregated data.

Whenever data is stored, it also needs documentation, for instance:

- What is the source of the data?
- Is it a direct (untouched) copy or are there processing steps taken?
- What is the explanation of all the features?

This makes it possible to trace back the origin of the data and when necessary recreate the data.

When looking at big data this becomes even more important as it is easier to overlook errors within the data. It is no longer possible to check data by hand, so you need to have the metadata of a dataset. This way it is possible to keep in mind the data format and range used in the dataset without having to analyse all the data.

6.1.3 Process mining

When trying to improve processes, people tend to immediately start drawing process models using "boxes and arrows", but these are rarely connected to the underlying data. Process mining aims to discover, monitor and improve real processes by extracting knowledge from event logs readily available in today's information systems. The four basic types of process mining are "process discovery", "conformance checking", "process reengineering" and "operational support" (see Figure 5).

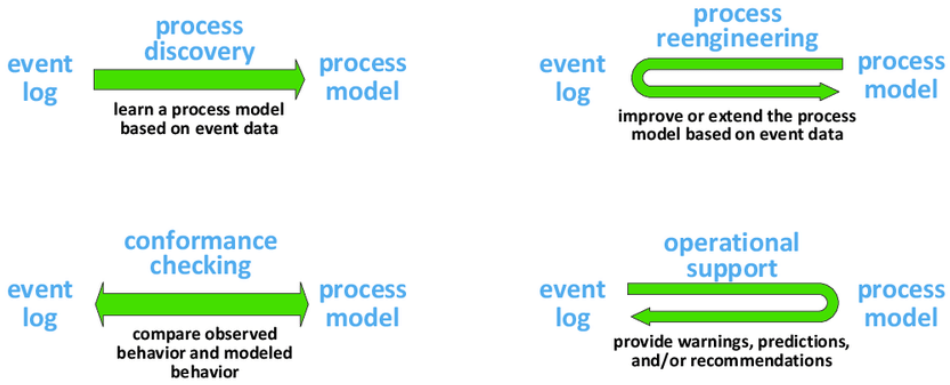

Figure 5: The four basic types of process mining[19]

For this SuAc 4.4 on Data Quality, the "conformance checking" is the most important type, as it will enable an audit on the observed behaviour by comparing it with the modelled behaviour, but we will explain briefly also the 3 other types. This "conformance checking" relates strongly to the "Plausibility check of raw data" that was described in paragraph 5.2 Data source types – Data processing concept. If the number of different raw data sources is limited and if their interrelation is clear, then this check can be done by "built-in rules". However, when there are multiple raw data sources and the interaction is not always known in advance, then Process Mining can come to the rescue.

[19] Aalst, Wil. (2018). Process discovery from event data: Relating models and logs through abstractions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 8. e1244. 10.1002/widm.1244.

6.1.3.1 Process discovery

Most organisations use information systems to support the execution of their activities and business processes. These information systems typically support logging capabilities that register the activities that have been executed by the organisation. Connected to the activities, the event log file may contain also the associated data - attributes. As explained above, also sensors are used to generate the raw data and store them in an event log, e.g. an echo sounder to generate depth information of a waterway at a certain moment in time.

case id	activity name	timestamp
12785	receive loading order	23-1-2018 09:30
42873	receive loading order	23-1-2018 12:30
12785	receive phyisical location on board	27-1-2018 14:32
72354	receive loading order	27-1-2018 15:57
12785	physical loading	28-1-2018 17:24
12785	send loading confirmation	28-1-2018 17:35
72354	receive phyisical location on board	29-1-2018 15:55
42873	receive phyisical location on board	30-1-2018 10:07
42873	physical loading	2-2-2018 10:02
42873	send loading confirmation	2-2-2018 11:00
72354	physical loading	4-2-2018 12:30
72354	send loading confirmation	4-2-2018 14:05

Figure 6: Event log from terminal management system

If a log provides the activities (and associated data) that are executed in the process and if it is possible to infer their order of execution and link these activities to individual cases. Then the process flow can be mined by process discovery algorithms. The generated model can then be used as a feedback tool that helps in auditing, analysing and improving existing business processes and data flows.

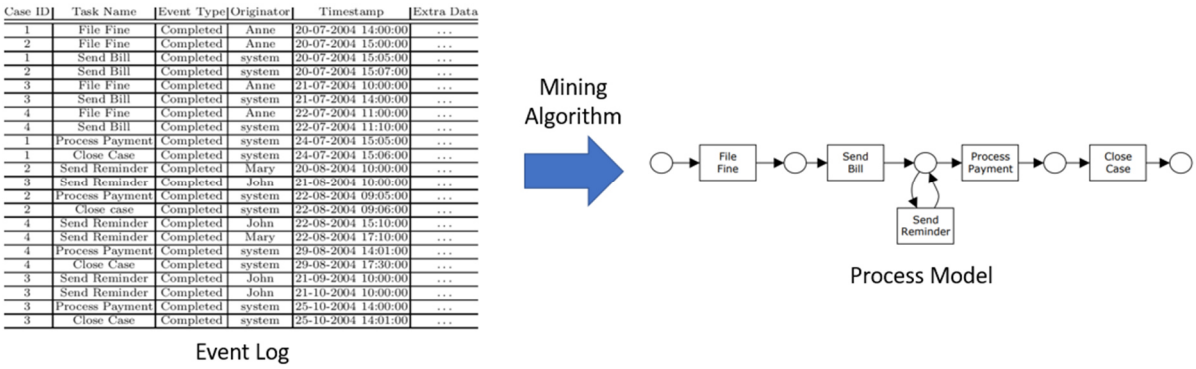

Figure 7: Petri net illustrating the control-flow perspective that can be mined from the event log[20]

6.1.3.2 Conformance checking

Conformance checking is a technique used to compare event logs with the existing reference model, or target model, for that process. It can be used to check if reality, as recorded in the log, conforms to the model and vice versa.

The goal of conformance checking is to identify two types of discrepancies:

- **Unfitting log behavior:** Behavior observed in the log that is not allowed by the model
- **Additional model behavior:** Behavior allowed in the model but not observed in the log

The interpretation of non-conformance depends on the purpose of the model. If the model is **descriptive**, discrepancies between model and log indicate that the model needs to be improved to capture reality better. For **normative models**, discrepancies may reveal undesirable deviations.

[20] Medeiros, A. & Aalst, Wil. (1970). Process Mining towards Semantics. 10.1007/978-3-540-89784-2_3.

Furthermore, by discovering these undesirable deviations, their root-causes can also be identified. Conformance checking can therefore also help to find inaccurate and missing data in a database.

Take, for example, the terminal loading process described in the table below. Not sending a loading confirmation to the carrier would be an undesirable deviation in this process. If the loading confirmation cannot be sent to a carrier, because there is no contact information available in the database, the missing data can be identified as the root-cause of this deviation.

Event data			Attributes			
Case id	Activity	Time stamp	Ordernumber	Weight	Size	contact
12785	Receive loading order	21-1-2018 09:30	SO14789	0,85t	1,1m³	name1@carrier.com
42873	Receive loading order	23-1-2018 12:30	SO14896	0,35t	0,55m³	
12785	Receive physical location	27-1-2018 14:32	SO14789	0,85t	1,1m³	name3@carrier.com
72354	Receive loading order	27-1-2018 15:57	SO15893	1,25t	1,85m³	name3@carrier.com
12785	Physical loading	28-1-2018 17:24	SO14789	0,85t	1,1m³	name1@carrier.com
12785	Send loading confirmation	28-1-2018 17:35	SO14789	0,85t	1,1m³	name1@carrier.com
72354	Receive loading location	29-1-2018 15:55	SO15893	1,25t	1,85m³	name3@carrier.com
42873	Receive loading location	30-1-2018 10:07	SO14896	0,35t	0,55m³	
42873	Physical loading	02-2-2018 10:02	SO14896	0,35t	0,55m³	
42873	Send loading confirmation	02-2-2018 11:00	SO14896	0,35t	0,55m³	
72354	Physical loading	4-2-2018 12:30	SO15893	1,25t	1,85m³	name3@carrier.com
72354	Send loading confirmation	4-2-2018 14:05	SO15893	1,25t	1,85m³	name3@carrier.com

Figure 8: terminal loading case with missing contact information

Inaccurate data can also be the cause of an undesirable deviation. When the goods are heavier than recorded in the database and their real weight is too high for the goods to be loaded on the vessel, the "physical loading" activity cannot take place. If this data is inaccurate for a significant part of the cases, certain measures should be taken to improve the accuracy of this data.

6.1.3.3 Process reengineering

Improving or extending a model based on event data is called "process reengineering". Like for conformance checking, both an event log and a process model are used as input, but the goal of process reengineering is to change the model instead of diagnosing the differences. With process reengineering techniques, it is possible to repair the model to better reflect reality. Additional perspectives can also be added to enrich an existing process model. Process reengineering yields updated models that can be used to improve the actual processes.

An example in IWT could be the sequence and the geographical position in which the skipper passes information to the fairway authorities If this process is always different than expected, but there are no arguments why the actual process is bad, then the expected process model could be changed to the actual one.

6.1.3.4 Operational support

Most process-mining techniques work on "post mortem" event data, i.e. they analyze events of cases that are already completed. However, nowadays many data sources are being updated in real-time and sufficient computing power is available to analyze events when they occur. By using "pre-mortem" event data, process mining can be used for operational support, directly influencing the process by providing warnings, predictions and/or recommendations. Based on the model and the event data of running process instances the remaining processing time, the associated costs, etc. can be predicted. Conformance checking is done "on-the-fly" allowing people to act the moment things deviate.

An example for IWT could be to give real time advise to the fairway authorities concerning traffic management based on process mining techniques.

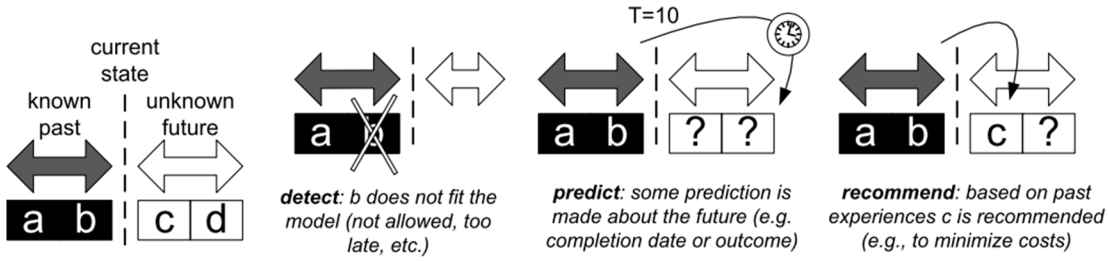

Figure 9: Online process mining using "pre mortem" event data[21]

6.1.3.5 Data quality in process mining

We started this chapter by stating that process mining can assist in improving that quality of data by detecting flaws and outliers in the data. On the other hand, we must be aware that the quality of event data has been recognized as a major challenge for applying process mining in practice. Despite the wide range of algorithms that have been developed over the past decade, the reliability of process mining outcomes depends on the quality of the input data. Consistent with the notion of "Garbage In, Garbage Out", applying process mining algorithms to low quality data can lead to counter-intuitive or even misleading decisions. Therefore it shows that data quality measures are most important for advanced data analysis.

6.1.4 Artificial intelligence

In DIWA SuAc 3.1 "New Technologies", Artificial Intelligence (AI) has already been assessed and elaborated on as a very promising technology,that could expand the potential of many other technologies and replace humans in many administrative and repetitive tasks. In terms of improving data quality, there are multiple ways how AI could be applied. AI systems are able to identify and ingest data without manual intervention. AI can automate the matching of third-party data, which results in a more complete data set. Furthermore, AI algorithms can be used to automatically clean-up and remove anomalies and duplicates within a data stream or database. However, as also stated in DIWA SuAc 3.1, the performance of the AI model depends on the quality of the data that is used to train the model.

Data capturing involves using technologies that allow machines to collect data and then transform it into meaningful insights ("information"). Technologies such as Intelligent Document Processing (IDP) can be used to translate elements from a bill of lading into structured data. The machine learning models are trained to extract specific information, including names, dates and figures from the document. Besides enhancing the speed at which data is captured, these technologies also reduce the risk of human errors while entering data, thus improving data quality.

Another way to improve data quality is to add more relevant data to the data set. By using matching algorithms and machine learning, data can be extracted from third-party sources. AI suggests what to fetch from a particular set of data and builds connections within the data. Third-party data inclusion results in a more complete data set, which adds value to the quality of the data.

e.g. all kinds of data (like weather predictions, traffic information on the fairways but also on the roads crossing the moveable bridges & locks, planned public events with many people that have to pass etc …) can be used to predict the operational impact on the bridges & locks. This can also challenge the Expected Time of Arrival from vessels. Using and combining more types of data can lead to better data quality.

AI can also be used to eliminate duplicate records and detect anomalies in a database. Duplicate entries of data can lead to outdated records that result in bad data quality. As a human being, it is often difficult to identify recurring data entries in a big repository. However, AI algorithms can automatically detect these duplicates. These AI algorithms can be very helpful to check datasets e.g. hull reference database where the hulls are registered by humans.

[21] van der Aalst, W.M.P. (2011). Operational Support. In: Process Mining. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19345-3_9

Anomaly detection is the task of identifying rare occurrences and filtering or modulating them (cf above: "process mining"). Anomalous events can be connected to some fault in the data source, such as equipment fault or irregularities in time series analysis. A machine learning-based anomaly detection model can be trained to detect and report such anomalies retrospectively or in real-time. These anomalous data entries can later either be flagged to analyse or immediately removed to maintain the cleanliness of the data before other processing is done ("data cleansing").

For example receiving position reports through AIS could result in relation to other sources like ship characteristics/data that the reported position could not come from this ship as it sails over land or on a fairway it never could sail. So the quality of this data is questionable. But if there are multiple occurrences with the same issue this could mean that the fairway data is questionable. AI could help detect and act on it.

As we also explained in the paragraph on process mining, we must be careful not to "walk in circles" when we say that AI can improve the data quality. Namely, the ability of a machine learning model to detect anomalies and duplicates within a database, highly depends on the quality of the available data. A machine learning algorithm that uses irrelevant or faulty data as input will not be able to find the right solution. On the other hand, if high quality data is used for training the model, it will be able to solve more and more complex tasks. Therefore, it is critical to pre-process datasets before applying them to train a machine learning model.

6.1.5 Semantic modelling, smart cities

In this paragraph, we will briefly explain how semantic modelling is already in use within "smart cities", then witness that this concept is also being adapted in transport & logistics and ending in making aware of the potential influence on data quality in all transport modes and specifically in IWT.

The European Commission defines a smart city as “a place where traditional networks and services are made more efficient with the use of digital solutions for the benefit of its inhabitants”. In a smart city, information & communication technologies (ICT) and various devices connected to the Internet of Things (IoT) are integrated to efficiently manage and govern the city. This does not only concern a more interactive and responsive city administration, but also smarter facilities including libraries, hospitals, utilities and the transportation system.

To manage assets, resources and services more efficiently within a city (or a terminal in IWT), many types of data are needed, such as public transport schedules, waste collection data, cultural city events data, air quality data, etc. Smart city data sources offer this information, but these different types of data come from heterogeneous sources and are often non standardised, which results in varying data quality depending on the data source. Take for example sensor devices that measure different types of observations such as light, temperature, or sound. The different sensors will provide data of different and even changing quality (e.g. the sensor device wears out over time). Furthermore, devices of the same type will deliver data in various formats (e.g. different units of measure, different standards, …) and even the periodicity at which data is captured might differ. It can be stated that smart city data are very heterogeneous in nature.

This heterogeneity issue can be partially solved by semantic modelling. Semantic modelling can help map data between different schema models at a higher level. When semantically annotating data streams, expressivity and complexity must be carefully balanced, as well as the sheer amount of generated data to be processed. Additionally, the entire data processing pipeline must be designed with scalability in mind. Several models have already been developed, including W3C SSN and oneM2M.

- W3C Semantic Sensor Network (SSN) ontology is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators. So, it does not only describe sensor device capabilities but also organises the sensors into systems and describes the processes of sensor operations.

- OneM2M base ontology aims to provide a high-level ontology for the IoT market in order to provide a minimal set of common knowledge that enables the cross-domain syntactic and semantic interoperability. The oneM2M ontology is very abstract and general, thereby oneM2M expects external ontologies that describe a specific domain of interest in a more detailed way to be mapped to the oneM2M base ontology.

The data provided by these semantically annotated streams eventually have to be interpreted and combined with other data sources. The usual data integration problem comes in different forms: data has to be integrated with metadata and different types of data from other sources such as static databases, semantic web knowledge bases or social network APIs. To solve this issue, semantic models can help create interoperable representations of data provided by different heterogeneous resources.

This concept of semantic modelling is also picked up by the sector of transport & logistics (e.g., by DTLF, Digital Transport and Logistics Forum, an initiative from the European Commission, DG MOVE). Especially when data has to be shared amongst different transport modes (road, rail, inland waterways, air, hyperloop, …) it is difficult to establish dedicated syntactical mappings from one format to another. It is there that semantics can be of use.

Although this semantic modelling can bring different worlds together, care should be taken on the influence on the data quality. Namely the quality of the resulting data and the derived information is strongly dependent on the mapping algorithms between the different domains. Also, the governance on the definitions on an atomic data level used within the different domains, where the automated mappings are based on, is of utmost importance to get satisfactory and trustworthy results.

6.1.6 Data Sharing versus Data Exchange

In the 90'ties, the dematerialization of paper document flows was focusing on EDI, Electronic Data Interchange (ANSI X12, UN/EDIFACT, EDI/XML, …). The target was to digitize data as early as possible in the Supply Chain and then pass it on from one computer system to another with as little human intervention as possible. This **Data Exchange** had clearly advantages compared to the paper flow. One disadvantage was that when the data was changed at the source, the other chain pieces not always were notified of it and by consequence they were relying on old, out-of-date data.
With the expansion of the internet, where companies and their IT systems are connected 24/7, other possibilities arose. Rather than sending copies of data to the rest of the Supply Chain, the source could share its data with other chainpieces (clearly based on Identification - Authentication - Authorization principles). This principle of **Data Sharing** is combined with **Data at the source**: instead of passing actual data from one chain piece to another, the link to the relevant data for a certain player in the chain is passed. Such a link can be referred to as a Unified Information Link (UIL) - e.g. in the eFTI directive of 2020. Other names are Unique Resource Identification (URI).

In IWT the details on the cargo could be provided by the consignor or by the importer, while the skipper only has to pass the Unified Information Link (UIL) to this data without reprocessing this data. In many cases, this principle needs a modernisation of the responsibilities on the provided data and the associated enforcement rules.

Since the data is maintained at the source, one could expect that the quality is better (up-to-date data and with less data loss than via the traditional data exchange chain). However, availability of the different parts of the data puzzle that can be scattered over multiple data bases and networks becomes more crucial than ever. This principle is also used in the Data Spaces e.g. the European Mobility Data Space of the European Commission - DG MOVE.

6.2 IWT related topics

The quality of data is important in a lot of current IWT related topics. Well known are the European reference databases (ERDMS, ECDB, EHDB) which are managed by the European Commission. The European Reference Data Management System (ERDMS), European Crew Database (ECDB) and

European Hull Database (EHDB) are important tools which are used by all fairway authorities and which needs to be complete, correct and up-to-date! But also other applications are using data for tactical and strategic decisions.
In the next sections, the usefulness of data quality is explained for the following projects and standards. Where and how is quality of data described?

6.2.1 RIS COMEX (EuRIS, CEERIS)

RIS COMEX is a CEF funded multi-Beneficiary project aiming at the definition, specification, implementation and sustainable operation of Corridor RIS Services following the results of the CoRISMa study. RIS COMEX started in the course of 2016 and lasted until mid of 2022. During that time there was the successful launch of two portals:

- EuRIS (www.eurisportal.eu): This portal implements the majority of the foreseen Corridor RIS Services:
  - Level 1: Static and dynamic Fairway- and infrastructure information (water levels, lock status, predictions, NtS, etc.)
  - Level 2a: Actual traffic situation (traffic density, passage durations, anonymized vessel positions, etc.)
  - Level 2b: Predicted traffic situation based on traffic planning and forecasts
  - Level 3: Information about specific vessels (positions, ETAs, ETA delays) for authorized users.
  - Cargo and/or voyage information of specific transports for authorized users.
  - Supporting services such as user management, data management (RIS index and Network Route Model), position service and route/voyage calculation.
- CEERIS (www.ceeris.eu): This web application provides specific Level 3 services such as reporting requirements, electronic reports provision and distribution. CEERIS depends on EuRIS for authentication, authorization, position information, reference data and route/voyage calculation. CEERIS provides published electronic reports to announce voyages towards EuRIS.

During the realisation of these services a lot of supporting/reference data needed to be generated by 14 COMEX Partners covering 13 countries. The scope of the project also expanded from delivering data for corridors towards a coverage of the whole inland navigable waterways. Many partners delivered data covering the whole network or are still doing so.

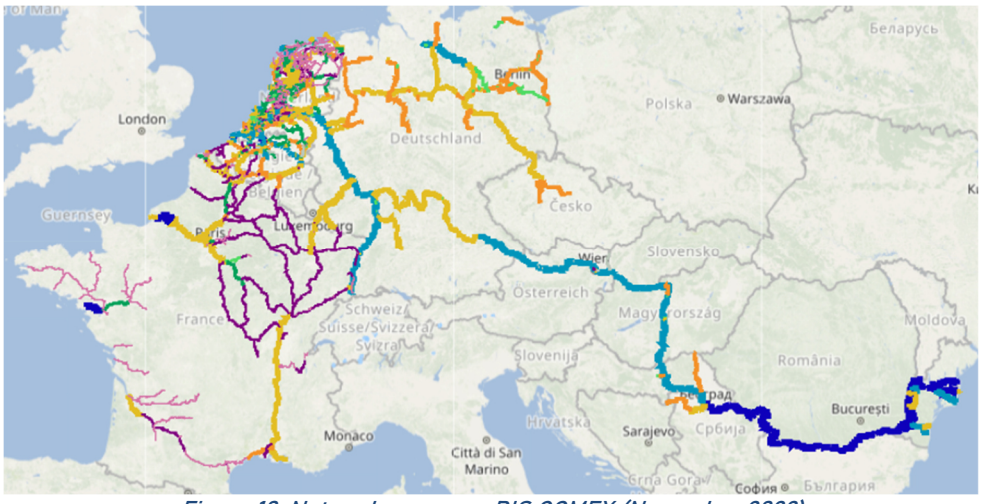

Figure 10: Network coverage RIS COMEX (November 2022)

During the project (realisation of EuRIS and CEERIS and data delivery) several challenges or problems occurred due to the complexity of the project, high ambitions and many stakeholders. From these challenges and problems several lessons were learned.
These lessons were also mentioned in the DIWA SubAc 2.4 report "RIS enabled corridor management". Chapter 5.3 of this report contains a number of risks and challenges which can occur during the further proposed actions (short, medium and long term).

Based upon this input several risks and challenges can also be marked as potential data quality issues. Lessons learned and the potential data quality issues are listed below:

1) National/regional differences in data availability and usefulness:
    a) The level of implementation and operation of RIS Corridor Services and related issues (data availability, completeness, management, validation, etc.) are **considerably different in the individual regions/countries**
    b) Standardisation leaves room for interpretation:
    Standards were differently interpreted within national implementation which lead to the problem, that the related data could not **easily be exchanged and consolidated** (e.g. webservice validation, import of data etc.), in some cases **standards are not followed properly** (even though there are clear requirements). Sometimes a standard lacks detailed information about how to encode the information or a shortcoming is present which can't be solved in an agile manner.
    c) Update interval:
    In each country there are **different update intervals** (e.g. water level: BE: 5 min, Eastern Danube: once/day). The same applies to the update interval of IENC cells.
    d) NtS:
    Organisational **differences when an NtS is published** (e.g. within **minute** in case of an incident up to **working days** because of involved signature process prior to publication in the respective ministry, WERMs are published occasionally, but no important NtS are published). The text content fields is used too often instead of using the structured coding. Missing end date information when publishing NtS limitations like obstructions. This leads to problems with the voyage calculation where an user thinks that a voyage is not possible in the future potentially losing interest in IWT transport.
    e) Hydrometeo:
    Water level/flow forecast is **very limited geographically. In a smaller extent this also applies to measurements.**
    f) Bottleneck:
    least depth/bottlenecks (commonly agreed proprietary web service) **does not work** for all Danube countries properly.
    In addition, there are **differences in the timeliness** of the data based on the survey.
    g) Reference data (network model and RIS Index) is not as mature everywhere or completely available with the same level of detail. The national networks aren't always properly connected. Difficulties while encoding common border areas. Missing agreements on hectometre and waterway/fairway name encoding.
2) Technical
    a) Position (AIS) data could be missing due to a lack of sufficient AIS basestation coverage or missing basestation redundancy (availability issue)
    b) Users are not always online due to lacking internet coverage (no WIFI or GSM coverage or too expensive). In this case they sometimes don't provide the necessary data like vessel/voyage intentions.
    c) Missing information concerning reference levels, units and quality of measurement. E.g. Does a water level measurement reported with cm unit really have 1 cm accuracy?
    d) Faulty information due to missing time zone information.
3) Interaction with the private sector
    a) Competition with commercial parties concerning data provision (e.g. ETAs - who is right)
    b) Lack of additional information like the intentions of the skippers (8h/12h or 24hour non-stop sailing and where to rest) for correct ETA calculations
4) Sceptical users / organisations
    a) Data quality not good enough to be trustful
    b) Digitalisation in a rather conservative environments
    c) Some vessel operators (or other data owners) do not want to share information which can lead to a lack of data availability
5) Organisation related
    a) Different status and progress in digitalization per country/region
    b) The cooperation between partners/Member States can be further improved

c) Slow update cycles for standards & legislation hindering technical/functional improvements
d) Lack of commitment or lack of resources at authorities in order to maintain ref data up to date
e) Dependency on other initiatives and organisations outside the influence of the data provider
f) Missing flexible update mechanisms/procedures to correct data when reported
g) Lack of metadata management concerning the data quality of the systems
h) Missing data (e.g. not every fairway user has an AIS transponder on board)
i) Poor data quality (e.g. wrong vessel-IDs within AIS data, non-accurate or outdated ETA information)

As CEERIS is strongly intertwined with EuRIS, the lessons learned from EuRIS naturally also play an important role in CEERIS. CEERIS-specific lessons learned in terms of data quality can only be applied after the system has been officially launched and is being used by users.

More experience on the subject of data quality related to electronic reporting is expected to be gained from the live operation of EuRIS.

To improve the sustainability of EuRIS and CEERIS further work on the accessibility, accuracy, availability, completeness, consistency, reliability, timeliness, unambiguity of provided data is recommended. There is a need for data quality framework together with the commitment of all involved partners to support this.

6.2.2 eRIBa – Functional and operational requirements

eRIBa (electronic Reporting for Inland Barges) is a smart communication platform for the exchange of digital reporting information between the inland shipping operator and the waterway authorities in Flanders and on the Western Scheldt.

In principle eRIBa is nothing more than a smart data hub for which skippers enter all obliged and required information necessary for the waterway authority in their inhouse application (e.g. BICS, RiverGuide, Autena, …) before starting a voyage (trip). This notification containing hull data, information about the cargo and voyage information, is received by eRIBa. The latter will distribute all data further towards the fairway authorities in the region based on the predicted route and in accordance with the ERI standard that they support.

Before passing the data to the different fairway authorities, the data quality is checked to comply with the following eRIBa – Functional and operational requirements:

- Check correctness of listed ERI IDs (blocking) - Uniqueness
- Control filling persons (blocking) - Consistency
- Control filling in date and time (blocking) - Accuracy, Reliability, Timeliness, Validity
- Check reference data ship - part 1 identification data (blocking) - Validity
- Check reference data vessel - part 2 reference database (non-blocking) - Validity, Consistency
- Checking completion of commodity codes and dangerous cargo codes (non-blocking) - Validity, Uniqueness, Consistency
- Control of ADN signal completion (non-blocking) - Validity, Uniqueness, Consistency
- Check filling in location codes (blocking) - Validity, Uniqueness, Consistency

6.2.3 Inland ECDIS

Inland ECDIS (Electronic Chart Display and Information System) is a visualisation system for electronic charts which are mostly used on board of a vessel. The Inland ECDIS viewer will display the charts according the international Inland ECDIS Standard. The standard defines which features are allowed to use and how the features and attributes are linked together. This is described in the feature catalogue. All allowed features, attributes and enumerations are defined and the relationships between the different objects and attributes are set.
Initially, the ENC (Electronic Navigational Chart) was developed solely for maritime navigation and was based on the S-57 data model. Because there was also a need for inland navigation, the S-57 data model of the ENC was used as base for the Inland ENC and amended with inland specific features and attributes.
The Inland ENCs were developed for navigation of manned vessels, the accuracy of the coded objects in the charts was not the biggest issue, rather the completeness and timeliness. However, all charts are coded with a certain intended use, called the **navigational purpose**. This usage already indicates a little bit the quality of the chart.

Nr.	Navigational purpose **(usage)**	Intended use
1 S57	Overview	For route planning and oceanic crossing.
2 S57	General	For navigating oceans, approaching coasts and route planning.
3 S57	Coastal	For navigating along the coastline, either inshore or offshore.
4 S57	Approach	Navigating the approaches to ports or mayor channels or through intricate or congested waters.
5 S57	Harbour	Navigating within ports, harbours, bays, rivers and canals, for anchorages.
6 S57	Berthing	Detailed data to aid berthing.
7 new	***River***	***Navigating the inland waterways (skin cell).***
8 new	***River harbour***	***Navigating within ports and harbours on inland waterways (skin cell).***
9 new	***River berthing***	***Detailed data to aid berthing manoeuvring in inland navigation (skin cell).***
A new	***Overlay***	***Overlay cell to be displayed in conjunction with skin cells***

The higher the intended use, the more detailed the data will become. This doesn't automatically meanthat the features are more accurate in a higher level

The Inland ECDIS standard distinguishes two types of accuracy:

- Accuracy of data;
- Data in relation to the display accuracy.

Accuracy of data (navigation mode):
Information regarding the position and orientation of other vessels, gathered by other communication links than the own radar, may be displayed only if they are up-to-date (nearly real-time) and meet the accuracy that is required for the support of tactical and operational navigation. Position information of the own vessel that is received does not clearly describe the required accuracy of the data but does describe the data in relation to the display accuracy:

Data in relation to the display accuracy:

- The accuracy of the calculated data that are presented shall be independent of the display characteristics and shall be consistent with the System Electronic Navigational Chart (SENC) accuracy.
- The Inland ECDIS in **navigation mode** shall provide an indication as to whether the display uses a smaller display range than the accuracy of the Inland ENC data offers (overscale indication).
- The accuracy of all calculations performed by Inland ECDIS shall be independent of the characteristics of the output device and shall be consistent with the SENC accuracy.

- Bearings and distances drawn on the display or those measured between features already drawn on the display shall have accuracy no less than that afforded by the resolution of the display.

Besides the accuracy, the quality of a chart is also depending on the completeness and timeliness of the data.

Because all relationships between features and attributes are described, the charts can be analysed on the correctness of these relationships. This can be done by software as for example the SevenCs Analyzer or dKart application. Not only the relationships are checked but also topology will be analysed in order to have a correct encoded chart.
For future developments as automated navigation, the accuracy of the features and attributes will become more important as the data will be used for navigational reasons. Currently, in the S-57 data model there is the opportunity to add accuracy information which describes the accuracy of specific features. This additional information could be or shall be an indicator for the calculations. In the future, the accuracy of the coded data should increase to make automated navigation safe and possible.

Within Europe there are many differences in the quality of the available IENCs. In some countries the detail is a lot higher than in other countries, there are many differences in the coded features, some charts are up-to-date, others are only updated after long periods.

6.2.4 RIS guidelines 2019

The concept of River Information Services was developed within the EU as a result of research projects like INDRIS and COMPRIS. Several regulatory organisations and river commissions recognized the potential of RIS to improve the position of inland navigation in the logistic chain. In 2002 PIANC (World Association for Waterborne Transport Infrastructure) developed the RIS guideline which led to the European RIS Directive in 2005. After several updates of the guidelines, PIANC published edition 4 in 2019 and replaced the term 'RIS Key Technologies' by 'Technical Services' and 'Services' were changed into 'Operational Services'. The RIS Directive itself, is under evaluation.

The updated RIS guidelines are defining RIS enabled Corridor Management and is covering different operational services.

> ***"Corridor Management is defined as information services among fairway authorities mutually and with waterway users and related logistic partners in order to optimise use of inland navigation corridors within a network of waterways"***

There are three levels of corridor management services:

Level 1: Operational services to enable reliable route planning by providing harmonised and standardised – dynamic and static – infrastructural information.

Level 2: Operational services to enable reliable travelling times for voyage planning and for traffic management, by providing traffic information:

a) considering the actual use of the waterway network (e.g. actual waiting times)
b) also, considering predictions during a voyage (e.g. predicted waiting times on the corridor) where considered reasonable

Level 3: Operational services to support transport management of the logistic partners (e.g. deviation management) and dealing with the information on vessels and the cargo.

The following information categories are covered by the related level:

Corridor Management Level	**Operational services on**
Level 1	• Static Infrastructural information • Dynamic Infrastructural information

	• Prediction water levels and ice
Level 2a	• Vessel related information • Traffic related information • Voyage related information
Level 2b	• Traffic planning/prediction
Level 3	• Tracking information of specific vessels and/or cargo • Prediction of delays for specific vessels both made available only through RBAC (role based access control)

Without mentioning how the quality has to be ensured, the guidelines are stating that Corridor Management is requiring a structured cooperation among the fairway authorities to provide a precisely defined set of harmonised operational services.

Although PIANC recognizes the importance of working with high quality data, the guidelines emphasize only once how the quality could be expressed. In the recommendations for the implementation of RIS Operational Services, fairway information services should be given with some kind of indication of the quality of the information. The quality can be defined by the parameters selected in 5.4.2, by the conformity to standards, and other parameters depending on the type of information. According to the guidelines the user should at least be informed about:

- The reliability of the information
- The accuracy and age of the information
- And the completeness of the information.

Furthermore, the guidelines indicate that for urgent information a high update frequency should be used.

For other services like Traffic Information Service the guidelines refer to the existing standards.

A last direct reference to the quality of information can be found in the chapter regarding reference data. Special attention is required for data quality and maintenance to guarantee a solid basis for the use of reference data and code tables (PIANC RIS Guidelines 5.6 Reference data).

The new PIANC RIS guidelines are also describing how a quality framework should look like to ensure the quality of the different services. In chapter 8.4 Considerations on Quality of River Information Services, the guidelines are describing a vision on Quality of Service without going into detail.

7 Inventory of data quality issues

7.1 Methodology

During the Sub-Activity 4.4 meetings, we felt the necessity to organise brainstorm sessions with all members to discuss not only the current data quality problems in IWT, but also the errors that will occur in the future of the digitalization based on new developments and if data quality does not improve. We decided to hold two brainstorm sessions: one on the current situation and one on the future situation.
We sent the questions in advance to all members so that everyone had time to think about possible answers in advance.

The goal was to use the input from the brainstorm sessions to create an impact matrix, in which we categorize the current data problems according to their impact on IWT (high, medium, low) and also include the data quality parameters that are not met and thereby cause the listed problems. Based on this, we identified the current critical problems so that we can make recommendations to avoid them during further digitalization of IWT. These are added in section "1 Executive summary".

7.2 Current situation

In the first brainstorm session[22], we discussed the following 4 questions about the current situation:

1. List the most common and annoying data errors you are currently experiencing
2. What is the impact of these data errors?
3. Rank these data errors according to importance (high – medium – low)
4. How do you think we can best solve these errors?

In order to rank the data errors by importance, some definitions for the impact classifications have been established. In this way, it is easier to classify the data errors according to importance:

- High: risk of human/material/infrastructure damage
- Medium: high business impact leading to business disruption
- Low: minor business impact leading to business inefficiencies

We ranked data errors according to the effect classification, and associated data quality parameters that are not met to the respective errors.
The top 3 high risk errors are:

1. Vessel dimensions and vessel types in AIS that are not correct, are leading to incorrect dimensions, incorrect berth occupation and make it impossible to correctly plan lock passages
2. Missing or confusing bridge height due to different definitions possibly leading to a vessel that is passing a bridge, but does not fit under the bridge
3. Not knowing whether data is correct and complete resulting in the data not being used. It is important to seek for different ways or sources to collect the data

7.3 Future situation

In the second brainstorm session, we discussed the following 3 questions about the future situation:

1. What developments require higher/different data quality than we (can) deliver now?
2. What can we (as waterway authorities) do about it?
3. What can we (as waterway authorities) not do about it?

In general, the most important developments that require higher or different data quality than we can deliver today, are smart shipping (SuAc 2.1), synchromodality (SuAc 2.2) and digital twins.

In the report of Sub-Activity 2.1 Smart Shipping, the importance of good data quality is described as follows: Automation of the sailing process on board of a vessel, requires information from different sources and systems. This is regardless of the level of automation. Although not all information is already availed, some is. With respect to the quality of that data the accuracy, completeness and availability play an important role. As the level of automation rises and the role of the human in the loop decreases, the importance of making sure that the data is correct increases. Besides high data quality, this also means more redundancy of different sources and systems as reliability becomes even more important.

The waterway operators can do certain things with respect to these future developments that require higher/different data quality measures than can currently be provided. These recommendations are included in Paragraph 8.3 Recommendations, along with the things that the waterway operators cannot do themselves, but where other stakeholders are needed.

One of the results of this brainstorm was the use of data quality key performance indicators. These KPIs give the authority the means to quantify the level of data quality and monitor the evolution of this level. A KPI concerning data quality can be implemented in several ways and monitor different quality parameters.

[22] See annex 9.4

8 Results and conclusions

8.1 Interactions with other Sub-Activities

In masterplan DIWA activity 2 (Business developments), several references are made to data quality.

SuAc 2.1: Smart shipping

The following needs for Smart Shipping have been identified to support increasing automation levels of vessels:

- Increase the quality of the data by investing in quality of existing data instead of a focus on sharing new types of data. A solution might be to build a digital twin of the waterway with the possibility for users to add or suggest changes.
- Need for more clarity on the quality (meta data) of existing data. This allows users to verify critical functional parameters.

SuAct 2.1 also refers to the PIANC WG 210 report where data quality indicators: Availability, Completeness and Accuracy are investigated with respect to Smart Shipping.

An aspect that should be taken into account when looking at the (needed) data quality is the investments that are needed to update the quality of the data versus the use of the data. When investments that are needed to reach the required quality exceed the expected business value towards the users, other solutions should be investigated. The distinction between different navigational tasks is an important factor when looking at the needed data quality.

When looking towards the desired future state SuAc 2.1 calls for the development of feedback loops which allows users to help contribute to the improvement of data quality. In 2032, it is expected that there will be an increased awareness in the whole sector that data quality is a combined responsibility of the whole sector, not only the waterway authorities. Especially with regard to AIS: an increased level of autonomy needs highly reliable and high-quality information. Consequently, control mechanisms for the correctness of AIS data are thus required. In addition, data availability and harmonisation should be improved:

1. All data that is shared by the authorities is available in a machine-readable way.
2. Authorities make sure that the way in which the data is shared, is truly harmonised.
3. Raw data can be shared as well.
4. Platforms for data sharing are used.
5. Data and information available on the EURIS portal can be retrieved by external systems.

It is acknowledged that despite the giant step forward in harmonisation as a result of the COMEX project, even in 10 years most likely not all data will be available on every stretch of the waterway. The (meta)data about the availability and quality of that data should however be available for every stretch of the waterway for Smart Shipping stakeholders.
It is furthermore stressed that data quality should be supported by a process that constantly does checks and never stops.

<u>Questions specifically directed to SuAc 4.4:</u>
1. Knowing that data quality is essential for use of that data. Which measures are possible to indicate the quality of the data in a harmonised way across entire corridors?
Answers on the first question can be found in paragraph 5.1.1 What is data quality, data quality management and information quality?, 5.1.2 Why is data and information quality management important?, 5.3.2 Selected data quality parameters for IWT and 5.4.1 Selected data quality frameworks for IWT.
2. Are there limitations on the use of data that is shared by an authority, for example in mission critical processes? What recommendations could be given?
There will be a limitation on the use of data between authorities but this is not due to the quality of the data but rather due to privacy requirements and cyber security reasons.

SuAc 2.2: Synchromodality

The topic of data quality has not been explicitly mentioned. However, the underlying research[23] does indicate that the choice of transport mode is partly determined by the reliability of the transport mode, which can be influenced by authorities via information exchange. High data quality or at least known data quality determines the reliability of this government instrument to influence mode choice.

SuAc 2.3: Port & terminal information service

The improvement of data quality improvement is seen as an important effect of increasing digital information exchange between skippers, terminals and authorities.

SuAc 2.4 RIS enabled corridormanagement

Poor data quality has been identified as a risk for the success and adoption of RIS enabled corridor management and subsequently data quality improvement in the areas of data consistency/quality checks and improvement of national data acquisition as a prerequisite for future actions. EHDB data (outdated, incomplete) is specifically mentioned as data with issues.
Just like SuAc 2.1, SuAc 2.4 also acknowledges feedback by skippers on fairway, RIS data, ENCs and any other RIS data used on-board to other skippers (increase of navigation safety) and to responsible organisations as a means to increase data quality.
Another suggestion to increase data quality is to implement and optimise automatic data consistency and quality checks. Also trust in the data could be enhanced by providing a data quality/reliability dashboard.

Questions specifically directed to SuAc 4.4:
It is important to define minimum quality requirements towards relevant IWT data and also to focus on procedures to increase and maintain a high quality level of data (data monitoring, consistency checks, update procedures, etc.)
The minimum quality requirements are defined as the selected data quality parameters for IWT and can be found in paragraph 5.4.1. Selected data quality frameworks for IWT. Information about the manners to increase and maintain a high quality of data, paragraph 5.2 Data source types – Data processing concept.

SuAc 2.5: ITS, ERTMS, E-navigation

Question specifically directed to SuAc 4.4:
1. Which frameworks will support the high quality quality in the offered services?

- It will be answered in Act. 4. However, this will applies in both directions. From ITS, ERTMS and e-navigation we will learn which data is crucial and what level of quality needs to be maintained. As soon as this is investigated and clear this information can be shared to Act. 4. In first instance, several existing data quality frameworks are addressed in paragraph 5.3.3 and we agreed that dedicated frameworks for IWT services need to be identified.

SuAc 3.1: New technologies

The New technologies draft report stresses the importance of data quality in the context of Big Data (veracity). Poor data means a high risk of biased or incorrect analyses.

SuAc 3.2: IWT connectivity platform

Connectivity platforms such as EuRIS and European Mobility Data Spaces Initiatives are identified as a means to improve the availability, quality and interoperability of data on multinational level – both in domain-specific settings and across sectors.

SuAc 3.3: Smart sensoring & PNT

[23] Mode Choice in Freight Transport research report 2022; International Transport Forum

This Sub-Activity calls specific attention to data quality of vessel position related data. This concerns not only basic accuracy of the position and assuredness of data integrity, but also completeness and accuracy of transmitted additional data. What is mentioned as important is not especially the absolute accuracy of a data value, but that data is accompanied by an indication of its reliability. The importance of this will increase when vessels are operated on higher automation levels.

Question specifically directed to SuAc 4.4:
1. There are a lot of parameters that can be used for describing the quality of data. Which of the parameters are important and applicable for IWT? Is there a recommendation?
- The selected data quality parameters for IWT can be found in paragraph 5.4.1. Data quality frameworks.

Recommendation specifically directed to SuAc 4.4:
Rec 16: Ensure high data quality of data generated to sensors. Investigate the data quality parameters to be met, in function of smart sensors (cfr. 7.4 Safety of navigation)

SuAc 3.4: Information model & data registry
This Sub-Activity reiterates the observation from SuAc 2.4 about poor data quality of the EHDB reference data and extends this to ERDMS (not up to date due to synchronisation issues). EuRIS is found to exhibit a high(er) level of data quality.

SuAc 3.5: Technology in other transport domains
It is stated that data quality is a critical characteristic when trying to achieve higher digitalisation levels. In addition, it is noted that technologies required for higher degree of automation via remote operation up to supporting vehicle autonomy are required to have a (much) higher degree of certain quality parameters (such as reliability and accuracy) as opposed to when the same technologies are employed for traditionally operated or partly automated vehicles/vessels.

SuAc 3.5 has specific recommendations/questions for investigation by 4.4:
Study-REC-IDL@Data-Quality-Requirements to Sub-Activity 4.4 and to whom it may concern: Study the impact of higher degrees of desired IDL (I and above) on data quality requirements for authorities when providing data. (IDL = IWT Digitalisation Level; described in the 3.5 report).
Study-REC-AV's+ROV's-Demand-of-High-Data-Quality to Activity 4.4 and to whom it may concern: Study the expected higher demand of data quality of Autonomous Vessels and/or Remotely Operated Vessels to be provided by Waterway Field Infrastructure and Inland Waterway Centres operated by IWT fairway & navigation authorities and ports, when entering into operational relationships with these vessels (see SuAc 2.1: Smart shipping) .
Study-REC-S100-Metadata-Registry-Impact: The S-100-Framework contains the S100-Metadata-Registry which is built in conformity to ISO 19135 Metadata standard and allows for, amongst many other things, the capture of data quality per data object. As motivated by Study-REC-S100-Framework-Application, the potential impact of the S100-Metadata-Registry is brought to the attention of Sub-Activity 4.4 for their study of the potential impact on IWT Fairway & Navigation. The IHO Registry is describing all features, attributes and partrayl items which can be used in the different domains (Hydro, Inland, IALA, …). The Feature Catalogue (FC), Portrayal Catalogue (PC) and Data Classification and Encoding Guide (DCEG) are build with the data from the Registry.

8.2 Conclusions

The most important conclusion of our research is that data quality is and remains most important for inland navigation and data exchange. If the data quality is poor, analyses based on the data are unusable. For further digitalization in inland navigation, data quality will play a key role for the necessary further technological developments. Therefore, to check the data quality, it is important to make use of the data quality parameters in IWT as researched to ensure that the used data, meets the associated parameters and objectives.

The quality framework includes the definition of the overall set of parameters and their values, mechanisms and guidelines aligned to the implementation of new business and technical services and their intended quality.

Because of the wide range of IWT related applications, a broad range of data quality frameworks can be used. It is impossible to assign a particular framework as 'the data quality framework for IWT'. However, using one is needed for good data quality in business processes.

Not knowing whether the used data is correct, accurate and complete leads to specific high risk issues that IWT is facing today. This is a high risk because incorrect data can lead to human, material and/or infrastructure damage. Therefore it is important to always take a look at the data quality parameters that are expanded in the report.

Smart shipping and autonomous sailing will require a higher or different data quality than is currently possible. To complete the above list synchro modality and digital twins will also require a robust data quality framework .

8.3 Recommendations

Listed below are the recommendations from this Sub-Activity. These can be directed to stakeholders as well as follow-up projects after the DIWA project.
We ordered the recommendations in different categories. Between the categories, the recommendations are listed based on how easy the recommendation is to implement.

- Basic: B
- Intermediate: I
- Advanced: A
  As a suggestion a stakeholder is mentioned to take action. In some case this might be a new task for this organisation, therefore an extension of tasks or a another/new body may be needed to take action..

REC	Recommendation	B-I-A	Suggestion Action for	In the
Additional study				
REC 1	Include a topic in future European projects to analyse and improve the data quality based on the conclusions of this study (cfr. 8.1 Interactions with other Sub-Activities)	B	EC	
REC 2	Develop a new quality standard with associated quality monitoring tools for RIS key services.	A	CESNI/TI	
REC 3	Investigate if bulk analyses of data could improve data quality. (Cfr. Annex 4: Inventory current situation data quality issues)	A	CESNI/TI	
REC 4	Investigate the needed level of accuracy of data which will be used for autonomous navigation.	A	CESNI/TI	
Data governance				
REC 5	Provide unambiguous reference data for common sections (jointly provided by competent authorities)	B	Each fairway authority	
REC 6	Appoint a single point of contact (SPoC) per national data provider for EuRIS. This SPoC can be contacted by the users of the data to report any data issues and the SPoC will respond and act accordingly. A feedback loop to the different levels of users is recommended.	B	Each fairway authority	
REC 7	Investigate how the awareness of the users and the providers of data can be increased regarding the quality. The responsibility of these key users in the results of derived products should be known.	B	Each fairway authority / CESNI/TI / EC	
REC 8	Closer cooperation with departments that are building and maintaining the infrastructure (GIS data).	I	Each fairway authority	
REC 9	Improve quality of Notice to Skippers (NtS): be faster (timeliness), be more complete (e.g. end date of limitations) and be more accurate. (Cfr.6.2.1. RIS COMEX (EuRIS, CEERIS))	I	Each fairway authority	

REC 10	Set up a European control body in force to monitor the quality of RIS related data according to the minimal requirements in the standards.	I	CESNI/TI	
REC 11	Develop a way so that databases such as ERDMS, EHDB and ECDB are updated at the right time. (Cfr. 6.2 IWT related topics)	I	DG MOVE	
REC 12	Define data quality KPIs per service (displayed in a data quality dashboard). (Cfr.7.3 Future situationFuture situation)	I	Future European project(s)	
REC 13	Maintaining data quality should be obtained by cooperation and by offering help by DG MOVE to the Member States.	I	DG MOVE	
REC 14	Agree on an official, short and realistic time frame for the update cycle of data. (Cfr.6.2.1. RIS COMEX (EuRIS, CEERIS))	A	CESNI/TI	
REC 15	Make use of automated data validation with feedback to data providers. (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	A	Each fairway authority (in the future a new EuRIS data maintenance organisation)	
REC 16	A data provider must include information/disclaimer on how to use the data and for which purposes it can be used. (Cfr.8.1 Interactions with other Sub-Activities)	B	Each fairway authority	
Data quality checks				
REC 17	Foster automated quality checks on data numbers/fields, especially on the input side(Cfr 9.4. Annex 4: Inventory current situation data quality issues)	I	Each fairway authority	
REC 18	Foster automated exchange of reference data. (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	I	DG MOVE	
REC 19	Make use of other (reference) data to check the quality of our own data and vice versa deliver data to other (non) waterway authorities. (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	I	Each fairway authority	
REC 20	Foster the use of Process Mining to detect irregularities (and possible errors) in the data. (Cfr. 6.1.3.5 Data quality in process mining)	A	Each fairway authority	
REC 21	Control mechanisms for the correctness of AIS data must be installed.	I	Each fairway authority	
Metadata				
REC 22	Add metadata about the level of accuracy , completeness, consistency, currency, timeliness, uniqueness, validity of delivered data. (Cfr.7.3 Future situationFuture situation)	B	CESNI/TI	
REC 23	Investigate the necessity of metadata concerning quality of data on different levels: level of complete database, dataset or on the level of individual records and fields. If necessary amend requirements.	I	CESNI/TI	
REC 24	Investigate how the source of data can be identified (e.g. GIS data).	A	Each fairway authority	
REC 25	Analyse the different S-100 domains (IHO Registry, FC, PC, DCEG) and the impact on the data quality for IWT.	I	CESNI/TI	
Requirements				
REC 26	Investigate the different data quality needs from the regions in Europe within the existing IWT standards. These needs can differ depending on the specific geographical situation. (Cfr 9.4. Annex 4: Inventory current situation data quality issues	B	Each fairway authority	
REC 27	Check with other modes the data quality requirements. (Cfr.7.3 Future situationFuture situation)	B	DG MOVE	
REC 28	Define a set of "business rules" to check data at the source. It can be offered as a EuRIS Service: pass XML data and receive a quotation on how consistent it is. (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	I	Future European project(s)	

REC 29	Make use of integration of information from 3rd parties (self-maintained or via focal point). (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	A	Each fairway authority	
REC 30	Define the same level of up-to-dateness, leading to a comparable level of quality of the available IENCs over different countries (Cfr. 6.2.3 Inland ECDIS)	B	Each fairway authority	
Standards				
REC 31	Define minimum requirements more detailed (e.g. less general). (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	B	CESNI/TI	
REC 32	Define detailed definitions because current definitions can often be interpreted in different ways (e.g. definition of 'vertical clearance'). (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	I	CESNI/TI	
REC 33	Organise a hands-on EU wide data quality team for harmonising IWT data. (Cfr 9.4. Annex 4: Inventory current situation data quality issues)	A	CESNI/TI	

9 Annexes

9.1 Annex 1: “Datakwaliteitsraamwerk hét naslagwerk”

Co-funded by
the European Union

9.2 Annex 2: Data quality frameworks

Co-funded by
the European Union

Acronym	Name of methodology	Year	Main Components	Data Quality Dimensions	Applicable for IWT?
AIMQ	A Methodology for Information Quality Assessment	2002	Data quality categorization model Information quality assessment instruments Bench-marking gap analysis Role gap analysis	Accessibility, Appropriate amount, Believability, Completeness, Concise representation, consistent representation, ease of operation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	2/7
CDQ	Comprehensive Methodology for Data Quality Management	2006	State reconstruction Assessment of data quality dimensions and setting targets Choice of optimal improvement process	Structured: accuracy, completeness, currency Unstructured: Currency, relevance, reliability Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	3/7
COLDQ	Cost-effect Of Low Data Quality	2001	Information Chain mapping Cost categorization Impact analysis cost determination Return of investment analysis	Data model: Clarity of definition, comprehensiveness, flexibility, robustness, essentialness, attribute granularity, precision of domaines, homogeneity, naturalness, identifiability, obtainablity, relevance, simplicity, semantic and structured consistency Data values: Accuracy, completeness, consistency, currency, null values, timeliness Information Policy: Accessibility, metadata, privacy, redundancy, security, unit cost Presentation: Appropzriatness, correct interpretation, flexibility, format precision, portability, consisten representation, representation of null values, use of storage Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	5/7
DQA	Data Quality Assessment	2002	Subjective and objective data quality assessments (metrics and surveys) Comparative analysis Root cause analysis Actions for improvement	Accessibility, appropriate amount of data, objectivity, believability, reputation, security, relevancy, value-added, timeliness, completeness, interpretability, ease of manipulation, understandability, concise representation, consistent representation, free-of-error Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	2/7
DQAF	Data Quality Assessment Framework	2013	Initial One-time assessment Automated process controls In-line measurement Periodic measurement	Completeness, timeliness, validity, consistency, integrity Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	4/7
DQPA	A Data Quality Practical Approach	2009	Identification of data quality problems Identification of relevant data Business rule development Data quality assessment Business impact determination Data cleansing Data quality monitoring	Accuracy, completeness, consistency, currency, timeliness, uniqueness, volatility Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	6/7
HDQM	A Data Quality Methodology for Heterogeneous Data	2011	State Reconstruction Quantitative evaluation of data quality problems Selection of appropriate improvement activities	Accuracy, currency Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	2/7
HIQM	Hybrid Information Quality Management	2006	Data quality definition Data quality evaluation Data quality monitoring Recovery support	Accuracy, completeness, consistency, timeliness Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	4/7
OODA DQ	The Observe-Orient-Decide-Act Methodology for Data Quality	2017	Iterative process: Observe Orient Decide Act	Speed, volume Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	0/7
TBDQ	Task-Based Data Quality Method	2016	Planning and evaluating assessment Evolution and execution of improvement	Accuracy, completeness, consistency, timeliness Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	4/7
TDQM	Total Data Quality Management	1998	Define, Measure, Analyze and Improve Focus: Information product	Accuracy, objectivity, believability, reputation, access, security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, consistent representation Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	3/7
TIQM	Total Information Quality Management	1999	Establish data quality environment Assess data definition and architecture quality Assess data quality Measure non-quality data costs Re-engineer and cleanse data Improve data process quality	Definition conformance, completeness, validity (business rule conformance), accuracy (to surrogate Source/to reality), precision, non-duplication, equivalence of redundant or distributed data, accessibility, timeliness, contextual clarity, derivation integrity, usability, rightness (fact completeness) Source: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8642813	4/7

Source : An Overview of Data Quality Frameworks[24]

[24] https://ieeexplore.ieee.org/document/8642813

9.3 Annex 3: Types of data sources

Cfr 4.1.3 What types of sources are there?

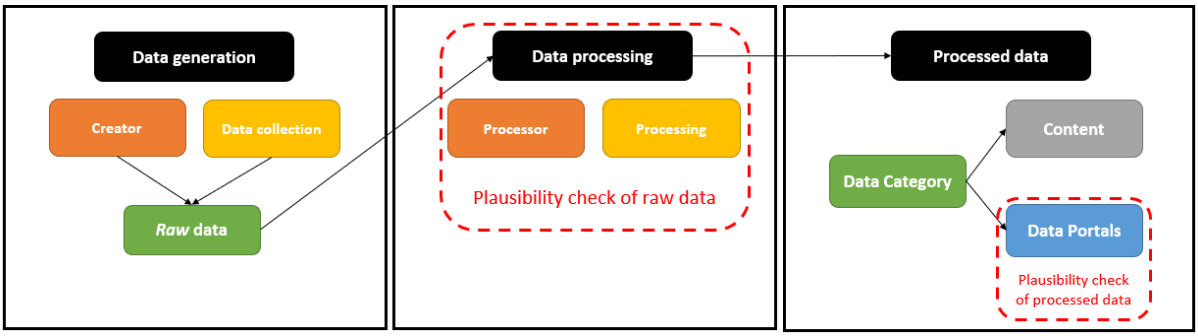

Figure 11- From Data generation towards processed data

The table below (also attached as Excel file1[25]) contains the most important data sources for the operation and maintenance of the waterway and navigation on it. For a better overview, the table was created in the opposite order (Processed data -> Data Processing -> Data generation) of the data generation process shown in Figure 11- From Data generation towards processed data**Fehler! Verweisquelle konnte nicht gefunden werden.**. This means that the table has been built up based on the data or information categories (Inland ENC, Bottlenecks, RIS Index, etc.).
The structure of the overview presented in Annex 1**Fehler! Textmarke nicht definiert.** will be explained here using the Inland ENC cards as an example.

[25] Excel „2022_04_22_DIWA_4.4_inventory_v0.1" – Sheet „inventory_data_sources"

Co-funded by
the European Union

PROCESSED DATA			DATA PROCESSING		DATA GENERATION		
Information Categories	**Data Content**	**Data Portals**	**Data Processor**	**Data processing**	**Data creator**	**Data collection** Measurement Tools (Sensors etc.), User Interfaces	**Raw Data**
Inland ENC	• fairway • bathymetric data • further information (navigation signs, berths etc.)	Danube: d4d portal	National Authorities	GIS Survey	National Authority (Survey, GIS)	• Echo sounder • LiDAR • Satellite images • Global Navigation Satellite System (GNSS)	• Measurement points/profile • Point cloud • Orthophoto • Coordinates • Bathemetry
Bottlenecks	• bathymetric data	Danube: Danube FIS portal		GIS			
RIS Index **Network data (RNM)** **RIS NET**	Geographical location of the objects of the waterway	EuRIS		Digitisation based on satellite images	National Authority	GIS	Geospatial data (normally the same information used to generate Inland ENC's) Rules and regulation
ERI	Content according to the message type • Vessel • Voyage • Cargo • Crew • Passengers • Inventory • Waste ...	• CEERIS • BICS • NaMIB • ...		graphical templates, etc.	Skippers, Vessel Operators	• analog paper forms • digital GUIs	• ERINOT (voyage, ship and (hazardous) cargo) • ERIVOY (voyage-plan or schedule) • PAXLST (details of passengers and crew) • BERMAN (purpose and service requirements for the call to the port) • INVRTP (Stores and especially bonded stores) • ERIMAN (CUSCAR) (details of the cargo carried including necessary permits, sent to other competent authorities such as Customs, Immigration, Police and Statistical offices) • WASDIS (waste on board and control including requests for the use of the Port reception facilities) • ERIRSP (confirmation: message recieved and processed) • MHDS (Minimum Hull Dataset) • APERAK (acknowledge receipt of the data)
NtS	• fairway and traffic • water level (see Hydrometeo) • ice • weather	• ECDIS • DoRIS • ELWIS • ...		graphical templates and conversion tools	National Authority	User interface and conversion of hydrometeo data	• FTM (fairway and traffic related message) • WRM (water level related message) • ICEM (Ice message) • WERM (weather related message)
Facility Files	• contact information • operating hours	• ENC • EuRIS • ...		graphical templates, etc.	National/regional Authority	information provided from facility operator	Static infrastrure information (photo, name and contact info) Rules and regulations Schedule information
Hydrometeo	values, time sequences (past, present, future)	Webservices of National authorities / weather institutes	National Authorities / institues	• Raw data is collected, processed and published by national Services (e.g. DoRIS)	National Authorities / institutes	Gauges weather stations	Waterlevels Discharge Weather information
Vessel and voyage related information	<u>dynamic data</u> • coordinates • speed information • course over ground / turning behaviour <u>voyage related data</u> • actual draught • port of destination • ETA • cargo information <u>static data</u> • vessel name • radio calling sign • type & dimensions of vessel	• EuRIS • National Services	National Authority	Triggering (e.g. ERI, ETA update), AIS Transponder	Vessels & Skipper	• GPS • Radar • vessel & cargo information • VHF * Sensors (heading, rotation, ...)	• coordinates • direction • speed • vessel & cargo information
Object Status	• waterlevels • bridge clearence • traffic light • status information: live data of lock chambers and gate) • traffic light • status information: live data of lock chambers and gate) • berth occupation	Bridges Locks Berths		• Raw data is collected, processed and published by national Services (e.g. DoRIS)	Objects	• Sensors (gauges, ...) • Info from SCADA/control system • Manual entry • AIS	• Signals • fault indications • Position • Dimensions

9.4 Annex 4: Inventory current situation data quality issues

Cfr 7.2 Current situation

1. List the most common and annoying data errors you are currently experiencing	Explanation (if needed)
Lack of (qualitative and/or verified) data	Which data is the official one? Maritime vs inland IMP, ITU, UNECE Split it up in data for waterway authorities and skippers
Lack of awareness for the importance of data	Need to make the people who upload data aware of the importance of a good quality of data
Primary data control is not used	Most of the time primary data control is missing E.g. boundary limits min and max values E.g. AIS coordinates are out of the working area
Missing or confusing bridge height due to different definition	
Use of - Generic ISRS codes, like NLXX.0000 - Unknown ISRS codes (ID mismatch)	Other version or synchronization issue, generic: somewhere in France to somewhere in the Netherlands, not everybody can pinpoint their exact destination, but should be able to pinpoint the place of departure ISRS codes in ERI messages that are unknown (different versions – no sync) ID Mismatch: source has a different reference data set than existing. Reference data is not persistent (e.g. RIS Index, water network, EuRIS data set used in NtS) - Issue within the waterway authorities - Issue which goes over the borders (over different waterway authorities) - Publication by one waterway authority, but not yet by another
(Non-recreational) AIS - Without ENI or IMO number - Without correct vessel dimensions	
Location codes are not aligned yet in the software (despite of ERDMS)	
Different sources of the same data with different Common Denominators	Different sources: receive data of the same vessel but from different sources Common denominator: same voyage but it is generated two times for the same voyage.
Inconsistency/change of the fairway network	
Discontinuity in data service (like berth occupation)	Skippers start to contact us (e.g.: berth occupation: report not possible with missing data)

Changed operating times of bridges/locks (also for recreational vessels)	
Limitations/fairway information on cross-border section is not provided in a uniform way. Overlapping/adjacent areas of competence of authorities. How to process for voyage planning?	Limitations: not always taken into account in EuRIS
Voyage planning: information on reduction of actual clearance sometimes is provided in relative values only, as it is depending on the water level, algorithm is not able to process such information	Voyage planning: difficult or impossible for the algorithm to take information into account. Not possible to have the clear consequences
Lock management: proper reference data of vessels, but with different sources (AIS, HULL, ERI) which lead to mismatches	
Different accuracies in different regions	

1. List the most common and annoying data errors you are currently experiencing	2. What is the impact of these data errors?	3. Rank these data errors according to importance (high – medium low)
Vessel dimensions and vessel types in AIS not correct	Incorrect dimension, incorrect berth occupation, not possible to correctly plan lock passages	High
Missing or confusing bridge height due to different definition	The vessel that is passing the bridge does not fit not under the bridge	High
Not knowing whether data is correct and complete (waterway authorities and skippers)	Seeking for different ways or sources to collect the data Data is not used because definition and/or quality is unknown	High
Location codes are not aligned yet in the software (despite of ERDMS) --> uploading data in time	Routes are calculated wrongly because of problems with the location codes	Medium
Inconsistency/change of the fairway network		Medium
Discontinuity in data service (like berth occupation)	Missing data in EuRIS and in statistics	Medium
No data available for certain areas (e.g. smaller ports)	Certain locations are not available for route/voyage planning	Medium
No data available for certain areas (e.g. private ports)	Certain locations are not available for route/voyage planning	Medium
Data from other fairway authorities provinces/ports etc	/	Medium
Use of - Generic ISRS codes, like NLXX.0000 - Unknown ISRS codes (ID mismatch)	Generic impact: intended users don't users trust our data, services and don't use them ISRS issues: unable to calculate voyage & ETA	Medium

ID mismatch: e.g. in NtS a specific object is referred to via ISRS: receiving system does not have this code	Ambiguous information in common (border) area	Medium
Changed operating times of bridges/locks (also for recreational vessels)	Skippers aren't able to pass a lock/bridge because operating times are incorrect	Medium
(Non) recreational AIS without ENI or IMO number	AIS issues: missing/wrong AIS data results in incorrect information presented to intended users e.g. berth occupation, ETA calculation	Medium
Limitation/fairway info on cross border is not provided in an uniform way. Overlapping/adjacent areas of competence of authorities. How to process for voyage planning?	Certain limitations cannot be considered by voyage planning	Medium
Lock management: proper reference data of vessels, but with different sources (AIS, HULL, ERI) which lead to mismatches	Real convoy dimensions are not known to lock managers prior to VHF announcement of a vessel	Low
Different sources of the same data with different Common Denominators	Linking data with different Common Denominators is sub optimal and in case data is inconsistent, we lost the truth	Low
Primary data control is not used, e.g. boundary limits min and max values, also e.g. AIS coordinates are out of the working area	Derive wrong results, make wrong decisions, unable to fulfil functionalities, distribute data errors and decisions	Low
Current issue: which tracks should be used in auto track pilots	/	Low
Voyage planning: information on reduction of actual clearance sometimes is provided in relative values only, as it is depending on the water level, algorithm is not able to process such information	Voyage calculation is not working properly based on provided data	Low
Different accuracies in different regions (one country uploads data more frequently than other countries)	Different data for the same areas in e.g. mixed zones. Example: one country upload data every hour, another country one time/day.	Low

4. How do you think we can best solve these errors?
Define - Data quality parameters per service (displayed in a data quality dashboard) - Detailed definitions because current definitions can often be interpreted in different ways - Minimum requirements more detailed (e.g. less general)

- A set of "business rules" to check data at the source (can be offered as a EuRIS Service : pass data and receive a quotation on how consistent it is) - Data in a human interpretable/readable way - Add source of the data
Foster - Automated quality checks on data numbers/fields, especially on the input side - Automated exchange of reference data - The use of Process Mining to detect irregularities (and possible errors) in the data
Make use of - A digital feedback loop from users - A hands-on EU wide data quality team for harmonising data - Other (reference) data where quality checks can be made on but also deliver data to other (non) waterway authorities - Data validation with feedback to data providers (e.g. cannot find a location/ID an NtS refers to in the published ref. data - Integration of information from 3rd parties (self-maintained or via focal point) - A combination of data from different sources and put into one common system where validation takes place, it will show errors and inconsistencies
Improve - Awareness for the people that input the data that good data quality is important - Lowering of the threshold for users to insert right data maybe on a less detailed level
- Penalties from DG MOVE to Member States that do not maintain the location codes according to ES-RIS standard - Provide unambiguous reference data for common sections (jointly provided by competent authorities) - Investigate if bulk analyses of data could improve data quality

9.5 Annex 5: Brainstorm future situation

Cfr7.3 Future situation

1. Which developments require higher/different data quality than we currently (can) deliver?
(Semi) autonomous vessels and smart shipping
EuRIS with more precise information on limitations with concrete impact on navigation and voyage planning
Synchromodality
Digital twins

2. What can we (as waterway authorities) do about it?
Add metadata about accuracy of level of delivered data
Increase awareness of importance of data quality
Find ways to check data and improve quality on a large scale
Adding better 'source control' of GIS data
Streamlined feedback loop between GIS department – RIS department – skippers
Closer cooperation with departments that are building and maintaining the infrastructure
Get in touch with other modes to see what data they would need
Be faster and more accurate in terms of communication and limitations
Improve internal processes (output oriented, not procedure oriented) and amend law where needed

3. What can we (as waterway authorities) not do about it?
We cannot become the 'owner' of all data problems. Clear separation between data we 'own' and are responsible for, and data of other parties that we provide via RIS
We can't force other departments, other modalities, … to cooperate and provide/validate/update the/their data
Resources are not endless, the level of provided data might differ depending on the 'importance' of the waterway (less information on waterways of lower priority)
Guarantee full completeness and complete accuracy

9.6 List of abbreviations

Abbreviation	Explanation
AIMQ	A Methodology for Information Quality Assessment
AIS	Automatic Identification System
CDQ	Comprehensive Methodology for Data Quality Management
CEERIS	Central & Eastern European Reporting Information System
COLDQ	Cost-effect Of Low Data Quality
D4D	Digital for Development
DoRIS	Danube River Information Services
DQA	Data Quality Assessment
DQAF	Data Quality Assessment Framework
DQPA	A Data Quality Practical Approach
ECDIS	Electronic Chart Display Information System
ELWIS	Elektronischer Wasserstraßen-Informations-Service
ENI	European Number of Identification
ERI	Electronic Reporting International
eRIBa	Electronic reporting for inland barges
EuRIS	European River Information Services
FEDeRATED	EU project for digital co-operation in logistics
FIS portal	Fairway Information Service Portal
GNSS	Global Navigation Satellite System
HDQM	A Data Quality Methodology for Heterogeneous Data
HIQM	Hybrid Information Quality Management
ICT	Information & Communication Technologies
IDP	Intelligent Document Processing
IENC	Inland Electronic Navigational Charts
IRIS Europe	Implementation of River Information Services in Europe
LiDAR	Light Detection And Ranging of Laser Imaging Detection And Ranging
NaMIB	Nachfolgeanwendung des bestehenden Melde- und Informationssystems für die Binnenschifffahrt (NaMIB) der WSV
NtS	Notices to skippers
OODA DQ	The Observe-Orient-Decide-Act Methodology for Data Quality
RIS index	River Information Services index
SENC	System Electronic Navigational Chart
SSN	Semantic Sensor Network
TBDQ	Task-Based Data Quality Method
TDQM	Total Data Quality Management
TIQM	Total Information Quality Management

9.7 List of figures

Co-funded by
the European Union