Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS

Otmane Azeroual, Joachim Schöpfel, Dragan Ivanovic, Anastasija Nikiforova

15th International Conference on Current Research Information Systems

Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS

Otmane Azeroual[a]*, Joachim Schöpfel[b], Dragan Ivanovic[c], Anastasija Nikiforova[d,e]

[a]*German Centre for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin, Germany*
[b]*GERiiCO-Labor, University of Lille, 59650 Villeneuve-d'Ascq, France*
[c]*University of Novi Sad, Novi Sad 21000, Serbia*
[d]*Institute of Computer Science, University of Tartu, Narva mnt 18, 51009 Tartu, Estonia*
[e]*European Open Science Cloud (EOSC) Task Force "FAIR Metrics and Data Quality", 1050 Brussels, Belgium*

Abstract

Consolidation of the research information improves the quality of data integration, reducing duplicates between systems and enabling the required flexibility and scalability when processing various data sources. We assume that the combination of a data lake as a data repository and a data wrangling process should allow low-quality or "bad" data to be identified and eliminated, leaving only high-quality data, referred to as "research information" in the Research Information System (RIS) domain, allowing for the most accurate insights gained on their basis. This, however, would lead to increased value of both the data themselves and data-driven actions contributing to more accurate and aware decision-making. This cleansed research information is then entered into the appropriate target Current Research Information System (CRIS) so that it can be used for further data processing steps. In order to minimize the effort for the analysis, the proliferation and enrichment of large amounts of data and metadata, as well as to achieve far-reaching added value in information retrieval for CRIS employees, developers and end users, this paper outlines the concept of a curated data lake with the data wrangling process, showing how it can be used in CRIS to clean up data from heterogeneous data sources during their collection and integration.

Keywords: CRIS; research information; research information system; heterogeneous data sources; data quality; data wrangling; data lifecycle; data consolidation; data lake; data cleaning; data warehouse; data lakehouse.

* Corresponding author. Tel.: +49 30 2064177–38.
E-mail address: azeroual@dzhw.eu

1. Introduction

Organizations and employees representing them, i.e. researchers in research institutions, must be able to integrate increasing volumes of data into their institutional database such as Current Research Information Systems (CRIS), regardless of the source, format or amount of the research information. The processing of data plays a central role in modern society, where the data is an integral part of various operational processes.

Given that CRIS are designed to store and manage data about research conducted at institution or organization providing an opportunity to extract from them knowledge useful for research management (Jeffery, 2004), (Schöpfel et al. 2017), it is important to wisely use the increasing amount and sometimes even variety of data to derive/ create value from them faster. More precisely, CRIS typically operates with the data on projects, persons, organizational units, funding programs, research outputs such as publications, patents, or related products, facilities and equipment, and events (Jeffery & Asserson, 2009). These data are the basis for decision-making for including but not limited to procedures in regards to hiring, promotions, preparation of annual reports, and submission of portfolios for accreditation and assessment (Yair, 2021). Poor quality of data and research information in particular can adversely affect the results of data-driven activities or decisions. In other words, the quality of the research information or trustworthiness of data is of paramount importance. This requires the selection and use of an appropriate data storage / repository and intelligent data processing to maintain the data and prepare them for use.

We suggest that this can be ensured by combining a data lake as a data repository and a data wrangling process, which allow data to be stored, managed and enriched in a central location serving as a single entry point (Mathis, 2017), (Sharma, 2018). In other words, the data can be stored in a storage different from the system when the data are collected, i.e. in a separate system such as CRIS or a repository. A data lake stores the data in a flat and raw / unprocessed format (Hai et al. 2016) and are only converted if formatting is required for their further use. Due to the diversity of data and their sources, connections between the data can be quickly recognized and used. The data lake should be integrated into the organization's IT landscape and can be connected to other data lakes (Miloslavskaya & Tolstoy, 2016). The integration of the data lake means that the research information is extracted from operational applications (e.g. HR, CRM and SAP systems as well as publications databases, etc.) and stored in the data lake, where public data can also be integrated and used for the enrichment of the above. However, data management can be seen as part of data governance and can be done using the data wrangling process (Endel & Piringer, 2015).

Data wrangling (also referred to as data mungling) is a process of iterative data exploration and transformation that enables their further analysis by making them (1) usable, (2) credible and (3) useful (Kandel et al. 2011). Kandel et al. (2011) suggest to "*determine usability in relation to the tools used to process the data, which can include spreadsheets, statistical packages, and visualization tools*". This makes the process of making data useful, where the preferable result of wrangling is an editable and auditable transcript of transformations coupled with a nuanced understanding of data organization and data quality issues rather than simply data. This means that many errors or anomalies can be corrected by data wrangling, e.g. structuring attributes into rows and columns, changing the layout of a dataset, deriving new attributes, filtering observations, aggregating values, grouping data, splitting a set of attributes and merging combining with other records (Azeroual, 2020). In addition, the data lake and data wrangling provide a scalable platform for storing and processing large amounts of research information. The data lake and data wrangling as a concept allows the storage of different data structures (internal and external, structured, semi-structured and unstructured). This makes it necessary to among other things enrich the data being not sufficiently complete and clean the data determined as dirty with data wrangling to be able to serve current and future analytical questions. The aim is to convert complex data types and data formats into structured data without programming effort. This means that users can prepare and transform their research information without being able or required to program with an ETL (Extract-Transform-Load) tool or other programming languages (e.g. Java, Python or SQL) (Azeroual, 2022). Once the data are read, these transformations are automatically suggested based on machine learning algorithms, greatly speeding up this process.

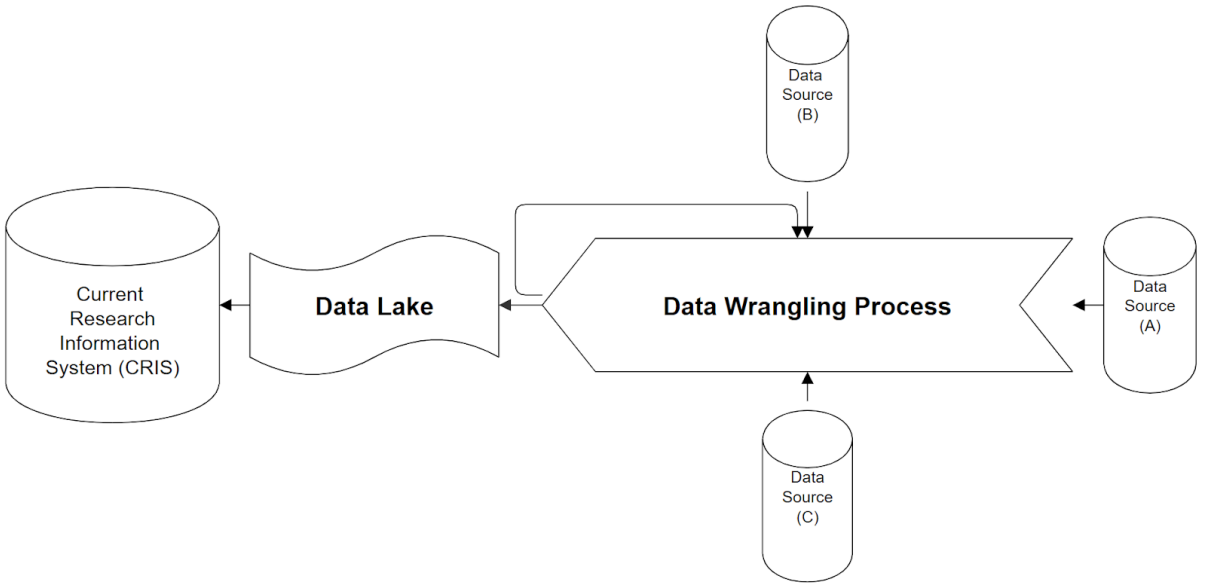

Fig. 1: Architectural model overview.

While CRIS intends to modernize routine managerial tasks involved in academic administration, in some cases they fail to do so (Yair, 2021). Thus, new advances and improvements are needed.

In this paper, we first design and specify an architectural model that analyses research information, cleans it, and transfers it into CRIS as shown in Fig.1. Data lake used as a data repository makes structured and unstructured data available in a single location and accessible in a more trusted, secure and controlled manner being in line with the data lake paradigm (Ravat & Zhao, 2019), (Zhao et al. 2021).The data wrangling process is used to check and improve the quality of data, which also prevents data from misuse, increasing the value to be derived from it as a result of its consequent use. This ensures that data are properly updated, retained, and eventually deleted / removed according to the phase of their lifecycle. The data wrangling process consists of subsequent successive steps. Depending on the information system (IS) and the desired target quality that may differ from one use-case / task / application to another, same as be dependent on the stakeholder involved, these certain steps have to be run several times. In many cases, data wrangling is a continuous process that is repeated over and over again at regular time intervals.

The paper is divided into five sections, where the Section 2 explains the typical challenges and implications associated with data quality issues that organizations face in a real-world using their database management systems (DBMS) such as CRIS, which can be improved through the use of data lake and data wrangling. Section 3 presents the conceptual design for storing, processing and improving the research information and describes the process elaborating on the central concepts and alternatives and the appropriateness of the selected components constituting the architecture of our proposal. Section 4 discusses the concept and the future perspective of its application, while Section 5 summarizes the main findings of the paper and outlines future work on this issue.

2. Data Quality in Practice – Challenges & Implications

Data quality describes how well the dataset fits the intended application (Wang & Strong, 1996). In this context, one also speaks of the suitability of the data for a particular purpose making the concept of “data quality” application- or

context-dependent (Strong et al. 1997). In addition, it is important to keep in mind that the quality of data is relative and dynamic in nature, i.e., the context determined by the use and the requirements that depend on it and may change over time, sometimes being determined by data gradual accumulation, and changing data quality requirements (Nikiforova, 2020). While data quality may be adequate and sufficient for one use case, it may not be insufficient for another (Wang, 1998). True to the motto "*garbage in and garbage out*", even a sophisticated complex algorithm is useless if the quality of the data is poor. Even though a project may fail for various reasons, the success of a project often depends on the quality of the available data (Redman, 1998).

All research information – tables, text or image files, has one thing in common. It has direct and indirect costs of resources required to create, collect or generate these data, and resources to maintain the quality of these data, be it through continuous maintenance of data and/or research information by entering it into systems or automated collection and processing of research information from HR systems or CRM and publication databases. Understanding high-quality data itself as a relevant resource that creates added value for the entire organization is the first step to more successful high-quality data-driven action. Data owners, including the level of the organization, need to establish at least a basic awareness of the data quality, thereby developing "data quality literacy". This also includes the need to establish an awareness of the processing of research information.

Organizations have a significant impact on the creation, collection, generation and control of data. Even the selection of data sources is the result of decisions. Questions that are especially important to answer in times of digitization include – *what systems do we use now or in the future? What data are and will be needed for what task / job today, tomorrow, in a month / year? Who is responsible for creating, generating and maintaining the data?* etc. Answers to these questions must be found. Scientific or research organizations should be strategically concerned with the topic of data quality and related aspects, in particular data storage and management, data integration, data availability and data security constituting an extended / more advanced understanding of data quality term. This also includes the FAIRness of data (findability, accessibility, interoperability and reusability) gaining an increased popularity and importance in the context of open science. It is a role model for all staff and system users when it comes to research information management and handling. A uniform strategy within the framework of data governance could identify an important need for action in terms of data cleansing and continuous data maintenance at an early stage and be taken into account when planning a data management life cycle (Otto et al. 2016).

Poor data quality has many consequences, e.g. direct financial consequences due to the apparent losses in returns and, above all, additional work for employees (McCallum, 2012). But there are also consequences for the success of the entire organization if qualitatively inferior research information is used to assess organizational control. All reporting structures, dashboard landscapes, and scorecards are based on processed and enriched raw data. The information content and meaningfulness of the estimates deteriorate / suffer significantly from the data of poor quality. In the worst case, the manager makes inaccurate decisions with far-reaching consequences. What helps is raising employees' awareness of continuous data maintenance and building expertise in organizations for consistent data quality management, i.e. data quality literacy.

Nowadays, especially in the era of digitization, it is essential to train employees on the systems they need to use to make the best use of in their daily activities with the aim of reducing or preventing errors, increasing possibility of their identification and elimination and thereby achieving consistently high data quality. Multiple systems with interfaces that are not clearly defined are breeding grounds for double data storage (duplicates, non-uniqueness), unwanted self-existence in irrelevant systems, employee frustration, and poor data quality with consequent negative effects on data-based activities. Existing systems need to check, test and use full integration capabilities. The links or associations created between used systems reduce the additional effort in data maintenance for the employees and simplify the control of individual data flows in IS or organizations. In the future, it is expected that organizational success will increasingly depend on how data and research information in the context of RIS are handled, processed

and evaluated. Research information system with its content is a valuable resource for the entire organization. Clear guidelines keep employees safe, secure and ultimately reduce data quality losses.

The real-world use-cases suggest that many CRIS users spend a lot of time pre-processing and preparing research information in their institutions (Schöpfel et al. 2020). The reasons for this are that the research information usually has to be collected / brought together from different source systems, where data quality has to be checked and ensured. In this phase, the sources are integrated in their original format into a data lake as a storage location, and exploratory data analysis is performed to identify anomalies in the research information (if any). This can be done by using different methods, where the methods from the data wrangling process are the most suited and popular because data wrangling methods can provide data-driven insights that help organizations detect, report, and prevent threats. It is important not only to analyse afterwards, but also to constantly monitor the process in order to foresee possible quality problems during execution and be able to intervene preventively.

Let's now provide more detailed insight on these concepts.

3. Data Lake and Data Wrangling as the key to success for data quality in CRIS

3.1. Data Lake vs Data Warehouse

The data lake in the era of big data focuses on the importance of a pool of raw data from different sources (Fang, 2015). The term data lake being a data repository represents also a methodology for using proprietary or native data formats for collection, archiving and analysis (Giebler et al. 2019) that makes it different from data warehouse – the most well-known alternative.

Despite the popularity of data lakes, data warehouses are often favored as a more "traditional" approach to deal with. Let us briefly elaborate on the key points of both concepts, emphasizing the difference between them that can be decisive when choosing a data repository.

Data warehouse is probably the most traditional and conventional data repository used to deal with highly structured, cleansed data that are pre-processed and refined organizing them into a single predefined data schema before they are put into the data warehouse and made available for further use for end-users. This cleaning and maintenance of data cleanliness is crucial for data warehouses, which is done before data ingestion with periodic data purges. The volumes of data that data warehouse typically deal with are measured in terabytes, which are relatively small amounts. In other words, they deal with relational data transformed for further processing for specific use-cases, mostly analyses, reporting, batch processing and business intelligence (BI) applications, where these tasks, for which the data will be used, are typically predefined, while the end-users are usually business analysts.

The key benefits of data warehouses are the standardization, quality, and consistency of data, and the ability to be used for providing business intelligence, increasing the power and speed of data analytics and business intelligence workloads, resulting in improved overall decision-making. At the same time, data warehouses lack data flexibility, requiring both high implementation and maintenance costs (Kutay, 2022).

The **data lake**, however, is considered the next step to replace Data Warehouses being a more modern concept of raw analytical data storage (Oreščanin et al. 2021) also referred to as second generation of data analytics (Armbrust et al. 2021). It deals with raw unprocessed data as received from an external source dealing with both unstructured, semi-structured and structured data with very little processing compared to data warehouses. Data lakes are suitable for dealing with large data volumes measured in petabytes with increased popularity and suitability for working with "unconventional" data, such as real-time and sensor data retrieved from sensors, IoT devices, social media (used for social network analysis (SNA) among others), cloud sources, cloud databases and on-premise databases. This makes data lakes suitable for use without a strict, predefined use-case or application, where data are prepared for a specific application on demand, processing them and transforming accordingly. As a result, data lakes are widely used for

stream processing, machine learning, AI and real-time applications, where data analysts and data scientists are considered the main audience. This definition, however, sounds that a data lake can cause "garbage out" since it takes "garbage in" as an input. To avoid this and take advantage of the amount and structure of the data to deal with, additional steps are taken, such as creation and maintenance of metadata, data governance, data cleaning or data wrangling, thereby preserving "garbage out" and the so-called "data swamp".

To sum up, the key advantages of a data lake are data consolidation, data flexibility, cost savings, and support for a wide range of data science and machine learning use cases. At the same time, data lakes has several disadvantages, such as poor performance for business intelligence caused by poor data organization, as well as the lack of a consistent data structure and ACID (atomicity, consistency, isolation, and durability) transactional support, which can result in sub-optimal query performance when required for reporting and analytics use cases. There is also a lack of data reliability and security due to the difficulty of implementing proper data security and governance policies to serve sensitive data types (Kutay, 2022).

All in all, the difference between a data warehouse and a data lake exists at both – the data, user, and use-case or application levels, where data warehouses deal with small, refined, relational data suitable for predefined tasks, while data lakes deal with large amounts of raw data suitable for undefined tasks.

Considering the benefits / strengths and weaknesses of both concepts, attempts are being made to eliminate their weaknesses and achieve the best possible result by combining them together seeking for a repository called "**data lakehouse**" (data lake + warehouse). A data lakehouse is defined in (Armbrust et al. 2021) as a data management system based on low cost and direct access storage that also provides traditional DBMS management and performance features such as ACID transactions, data versioning, auditing, indexing, caching, and query optimization. Data lakehouses are considered particularly suitable for cloud environments with separate compute and storage, where different computing applications can run on-demand on completely separate computing nodes, e.g., a GPU cluster for ML, while directly accessing the same storage data, while in some cases data lakehouses can be implemented in a local / on-premise storage system (e.g. HDFS). The term is becoming increasingly popular in the current reality as it is seen as a "silver bullet" and the third generation of data analytics that takes full advantage of both concepts, thereby reducing the number of cornerstones defined as a new paradigm in data architectures that embodies and integrates already established concepts for systematic management of disparate large-scale data – a data lake for managing heterogeneous data, using open standards for high-performance queries, and systematically keeping data "fresh" (Begoli et al. 2021). However, being a relatively new concept, it is currently considered an immature and rather a conceptual construct.

As a result, the data lakehouse is considered to be characterized by reduced data redundancy, cost efficiency, support for a wider range of workloads, as well as ease of data versioning, governance, and security (Kutay, 2022). While the key disadvantage of the data lakehouse is its immaturity.

The approach presented in this study is based on the concept of a data lake enriched with data wrangling, although some similarities with the concept of the data lakehouse are observed such as improved data security, reduced data redundancy, and data reliability achieved by means of data governance and data wrangling in general. Therefore, let us refer to the key points to be considered about the data lake in the context of this study.

Although there is a number of definitions of data lakes in the literature, in this study we use the term "data lake", which definition is a combination of the above mentioned definitions. In other words, we define a data lake as a concept that deals with the storage of raw data from various internal and external data sources (Gorelik, 2016). The boundaries of data silos are removed and a central data management organized by metadata is created. Criteria for ensuring data quality and consistency are stored in data governance storage. Storage management is based on this data governance. Processing of the data records in the schema and evaluation and combination occurs at access time.

Various aspects affect the construction of a data lake. The following list summarizes the key points of the typical data lake (Ravat & Zhao, 2019):

- metadata that describes a dataset in containing information about the origin, structure and content of the data. In a data lake, metadata is used not only to enrich the data with additional information, but also as sorting, filtering or categorizing properties. In addition, metadata are used for system management and system administration;

- data mapping that describes the context of the data. The so-called integration map is a detailed specification of which application data from which data sources are linked / associated with which characteristics (mostly metadata);

- data lake context that describes the higher-level use case on which the data lake is based. Therefore, the selection of the required data sources is more targeted. This avoids the misuse of the data lake as a data swamp;

- data context – the individual datasets also have context so that they can be better classified for analysis purposes. The context for records can be data origin / lineage, categorization, or some other contextual feature in the metadata;

- processing logging that refers to the raw data processing that takes place in the data lake. The data record and its metadata are manipulated in the process. This data processing is of particular interest to data analysts to analyse data lake usage, data set and use case.

All in all, data lakes can store different data of different structures.

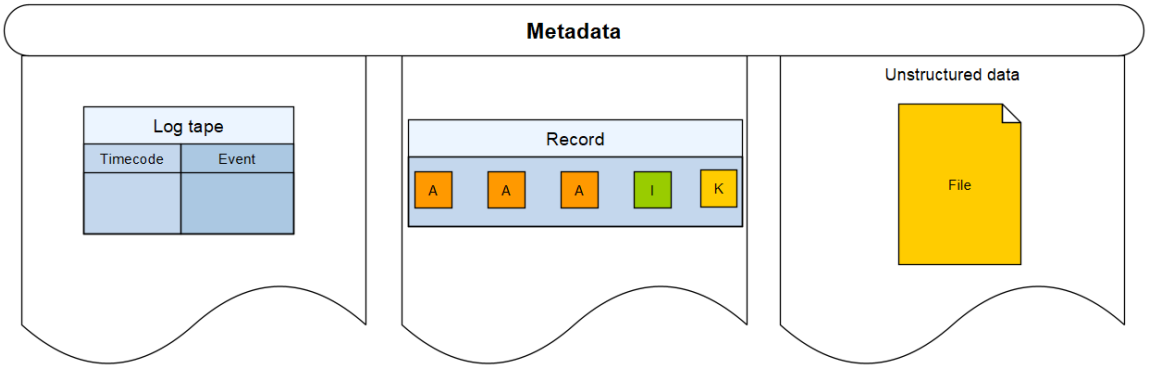

Fig. 2: Data sets in data lake.

Figure 2 provides an illustration, where these different types of data are:

- analog data: data sources automatically generate data in a specific predefined and therefore known data format. Due to the automatic generation of data, they accumulate in a very large amount and are mostly repeated / duplicated. For this reason, they are usually stored in tabular form in so-called “log tapes”;

- application data also have a known structure and are significantly different from analog data in their origin. While analog data typically represents physical measurement data, application data arises during the operation, transactions of an application. Application data includes, among other, transmitted system data or analysis data. So-called “records” are used as a common storage solution for these data. Typical for records is their uniform / homogeneous structure. A data record usually consists of a key attribute *K*, an index attribute

I, and other predefined attributes *A*. Depending on the data origin and data type of the application data, the predefined attributes may differ from each other. This application data structure is based on database management systems (DBMS);

- text-based data that are also closely related to the application, but the data of this category are stored as separate files with metadata. A transformation is required to be carried out for further processing of these data. The process of converting text-based / textual data into analytically processable data is called textual disambiguation.

3.2. Data Wrangling vs Data Cleaning

Similarly to the previous section considering different types of data repositories, let us first provide a brief overview of processes to be carried out dealing with the data and their processing. The most obvious and widely used process is data cleaning associated with data warehouses and being a part of Extract-Transform-Load (ETL) process, while another more advanced process sometimes considered to be an upper set of data cleaning is data wrangling. This discussion is particularly important given that these concepts tend to be used interchangeably as synonyms, even though they are not.

Data cleaning or data cleansing is the process or a series of processes dealing with the so-called "dirty" data thereby improving the overall quality of data. The simple data cleaning process refers to duplicate identification and elimination, while the advanced data cleaning also considers format standardization, data entry mistake identification and error correction, anomaly removal (dealing with missing values (incompleteness), spelling variations, unit differences, outdated codes, misuse of abbreviations), while deduplication also take into account inexact duplicate records, inferencing of missing values, and error correction (Low et al. 2001).

Data wrangling, also known as "data munging" or even "janitorial work", is the process of examining and transforming data into a usable form by mapping them into a required form that will enable further work with them, making them suitable and valuable for defined tasks. This requires a deep understanding of the content, structure, quality issues and necessary transformations, as well as the appropriate tools and technological resources, which makes this process relatively complex (Endel & Piringer, 2015). In practice, data wrangling is characterized by a series of steps to be taken, where a simplified approach is described as (1) gather – (2) assess – (3) clean, but a more detailed as (1) discover – (2) structure – (3) clean – (4) enrich – (5) validate – (6) publish for their further analysis and visualization. The goal, however, makes it appropriate to be used in data lakes, and therefore this study refers to this more advanced and appropriate concept.

3.3. *Proposed Solution: Data lake and Data Wrangling for increased data quality in CRIS*

In the context of research information, data wrangling refers to the process of identifying, extracting, preparing and integrating data into a database system such as CRIS. At the end of this process, research information can be used by analytical applications and protected from unauthorized access by access control. Figure 3 illustrates the concept of a data lake with a data wrangling process, and references to different data wrangling steps indicated by numbers. These steps have the advantage that the errors in recordings can be identified, eliminated and used for other scenarios (Endel & Piringer, 2015). The developed method is based on (Azeroual, 2020), where data wrangling in database systems

with a focus on the purging of dirty data has been presented. The data wrangling process serves to prepare research information and integrate it into CRIS for further analysis.

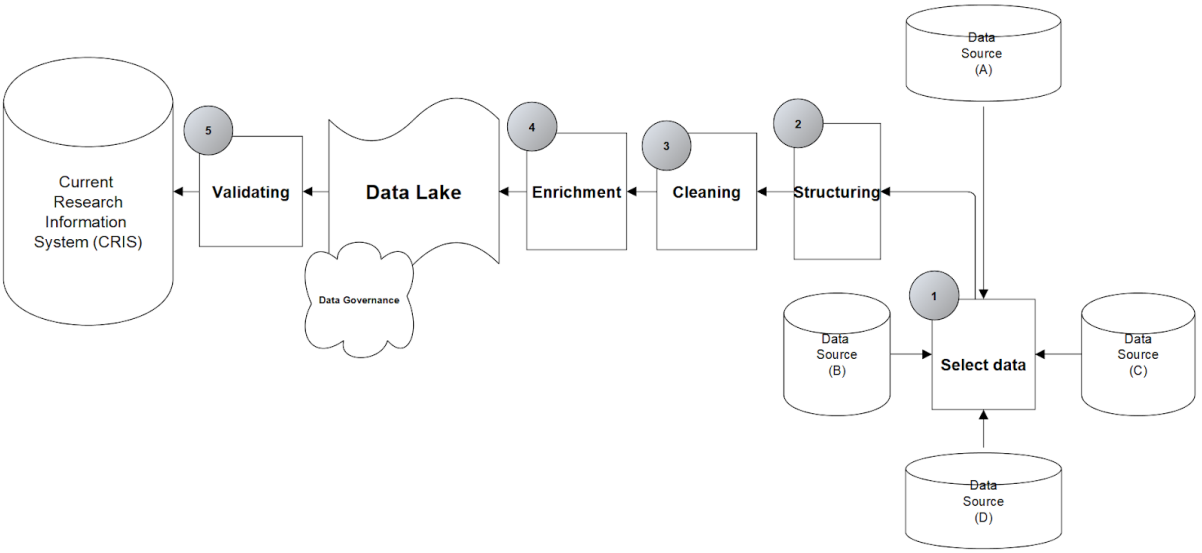

Fig. 3: Data lake and data wrangling process concept.

The very first step in the data wrangling process is data selection from different data sources, when the required data records are identified. This step can have a significant impact on the data lake. On the one hand, data must be triaged / filtered out to prevent the data lake from flooding with unused, unnecessary, and useless data. On the other hand, too much data should not be filtered out as well, so as not to lose the added value of the data. When selecting data, a record is evaluated by its value. If there is added value, the availability and terms of use of the data and subsequent data from this data source are checked. In other words, the data is an asset that can be legally protected by its owner. By issuing licenses, the owner sets the conditions that are linked to the use of the data. If a license is available, the next step is to check the terms of use (Terrizzano et al. 2015) that define the permitted use of the data. This step resolves legal issues.

In most cases, there is little or no structure in the data. Therefore, the second step is to change the structure of the data for easier accessibility. The change of the structure may mean splitting a column or row in two parts, or vice versa – whatever is needed for analysis.

Almost every dataset contains some outliers that can skew the analysis results. They need to clean up the data to get the best results. In the third step, the data are extensively cleaned for better analysis. This refers to the processing of null values, removing duplicates and special characters, and standardization of the formatting to improve data consistency.

After the third step, the data need to be enriched. That is, an inventory of the data set and a strategy for improving it by adding additional data should be carried out. This is done by using metadata, contributing to both organizing and structuring a data lake (Ravat & Zhao, 2019), and enriching it. In accordance with (Azeroual et al. 2019), these metadata can be described as:

- **schematic metadata** that provide basic information about the processing and ingestion of data. To do this, the data wrangler analyses / parses data records according to an existing schema, such as column names in

tables. Schema discovery does not always work automatically, so in some cases the data analyst intervenes manually;

- **conversation metadata** are exchanged between accessing instances (Terrizzano et al. 2015). The main idea is to document information obtained during the processing or analysis of these data for subsequent users. In this way, the recognized peculiarities / features of a data set can be saved so that the next user does not have to go through the same knowledge gain. This documentation is usually presented in text form.

At this stage of the data wrangling process, the physical transfer of data in the data lake take place. Data sources usually support what is known as "bulk download", in which a certain amount of data is downloaded from data servers in the form of files. Either the data server triggers the download, or it is triggered by a request. For the initial filling of the data lake, it is recommended to transmit data via media storage instead of the Internet, as this allows to transmit large amounts of data faster. Data can be downloaded over the Internet using standard protocols such as FTP and HTTP. But there are also specialized protocols that are interesting for downloading data lake files. The CKAN[†] or Socrata[‡] protocols often used by open government data portals enrich the data with additional metadata (Terrizzano et al. 2015). The data set is already enriched with some essential metadata.

Although data in the data lake are prepared using metadata, the record is not pre-processed. The main goal of a data lake here is to avoid a data swamp. This means estimating the value of the data and deciding on the lifespan of the dataset. This decision not only affects but also depends on the data quality. The value of data correlates with its quality and its interconnectedness with the rest of the database.

After enriching the data and integrating it into the data lake, the data (including but not limited to dataset – a set of records) are ready for use. Analyses are not performed directly in the data lake, but only on the relevant data. To be able to use the data, the accessing party / requester needs the appropriate access rights. To do this, Data Wrangler performs data extraction. However, general viewing and exploration of the data should be possible directly in the data lake so that data analysts can get an overview of the data lake. Storing data in a data lake should be governed by processes. Not only the contents of the data lake is subject to constant change, but also the technologies and hardware used. For this reason, an audit is required to take care of the current state and the maintenance of the data lake. The main principles / guidelines and measures are defined in "Data Governance". In the data wrangling process, data governance regulates data maintenance (Rattenbury et al. 2017). It coordinates all processes in the data lake and defines the responsibilities for these processes. These include maintenance processes, audit processes, decision-making processes in data management and access processes. Data governance consists, among other artefacts, of usage guidelines, the wrangling guidelines and other general data quality guidelines and process descriptions. The data lake focuses primarily on data security, lifecycle management, data quality and the use of metadata.

During the validation phase, the data are checked one more time before it is integrated into the target CRIS system. The goal is to identify problems with the data quality and consistency of the data, or to confirm that the corresponding transformation has been successfully carried out. In any case, it should be verified that the values of the attribute or field are correct and conform to the syntactic and distribution constraints. The validation rule checks the data for inconsistencies and thus ensures high data quality.

When validating data, it is important to document every change to a data set so that older versions can be restored or history of changes can be viewed, i.e. versioning should be ensured. This is especially relevant when editing content to be able to guarantee the integrity of the data entry / record. If new data are generated during data analysis in CRIS, it can be re-included in data lake. New data go through the data wrangling process, starting with the step 2 of data

---

[†] https://ckan.org/
[‡] https://dev.socrata.com/

validating and structuring the data. The data wrangling process completes with the cleaning and enrichment of the data lake with the data obtained as results of the analyses.

4. Discussion

The emergence of data wrangling solutions and advances in this field is driven by real-world necessity. While in the past, institutions and especially CRIS users, did not have the right tools to access and even more important understand, clean and format research information, much of the research information that institutions deal with today is increasingly available in a variety of formats and sizes. These data are collected from different data sources and are either too large or too complex to handle in traditional self-service tools such as Excel. Data wrangling solutions are designed to process any type of complex research information at any scale making the data ready for their further value-adding analyses.

Data management from data generation or collection to their analyses and data visualization become increasingly important and will remain as such in the coming years. It enables data integration within an organization, simplifies IT infrastructure, and forms a valuable foundation for data usage in organizations. As the volume of research information and data sources increases, the prerequisite for data to be complete, findable, comprehensively accessible, interoperable, reusable (compliant with FAIR principles), but also securely stored, structured, and networked (integrated and then exchanged between users or entities) in order to be useful remain critical but at the same time become more difficult to fulfill. Data wrangling can be seen as a valuable asset in ensuring this. The goal is to counteract the growing number of data silos that isolate research information from different areas of the organization. Once successfully implemented, data can be retrieved, managed and made available and accessible to everyone within the entity. A data lake and data wrangling can be implemented to improve and simplify IT infrastructure and architecture, governance and compliance. They also provide valuable support for predictive analytics and self-service analysis by making it easier and faster to access a large amount of data from multiple sources.

The concept of data lake and data wrangling contributes to the understanding of research information from its location to the state (structure, quality, value etc.). This aspect is key to supporting different user groups and analytics, since the proper organization of the data lake makes it easier to find the research information the user needs. Managing the research information that has already been pre-processed offers the greatest potential for increased efficiency and cost saving, as preparing research information is the most time-consuming part of data analysis. In addition, by providing pre-processed research information, users with limited or no experience in data preparation (low level of data literacy) can be supported and analyses can be carried out faster and more accurately.

5. Conclusion and Future Works

Research information can have a huge multi-dimensional impact. But before they can be properly used, they must go through a series of processes. The developed concept of the data lake and data wrangling allows to store data in a raw format. An essential step is the so-called data wrangling (also referred to as data munging) – data cleaning and sorting at the beginning of each data analysis. While in most cases it is expected to be done in an automated way that is the goal, depending on the use-case, separate steps such as conversion of raw data will be done manually. Research information only goes through a verification and enrichment process of a data wrangling process, before it enters the data lake and is integrated into CRIS (see a practical example in our presentation[§]).

The presented concept provides a logical basis for implementation in the CRIS. It allows the modeling of the basic security mechanisms to ensure a certain level of quality control within a CRIS. While it is based on the concept of a data lake enriched with data wrangling, there are many similarities with the concept of the data lakehouse, including

[§] https://dspacecris.eurocris.org/handle/11366/1957

improved data security, reduced data redundancy, and data reliability achieved by means of data governance and data wrangling, which are typically considered to be weaknesses of data lakes.

The developed model is the first version of such a solution. It is possible and even recommended to expand it with additional elements from other security models. The Brewer-Nash model is particularly interesting in this respect and will be considered in our future work. This allows to avoid conflicts of interest and can be used in a data lake to separate different datasets.

In addition, the concept presented in this study offers many possibilities, including but not limited to more efficient use of research information in organizations. In addition to completely new applications and the resulting business opportunities, it primarily ensures the democratization of research information, i.e. having the right research information available at the right time, but in case they are not, making them easily and quickly available. This allows to ensure the right data basis for further analytics and data-driven decision-making. However, while the proposed solution is capable of solving the above discussed problems (at least partly), another concept requires in-depth attention – FAIRness of CRIS. We argue that FAIR principles can be applied not only to the data and information but also to the whole RIS or CRIS contributing to the improvement of its FAIRness, which, however, is a dual or bidirectional process, where CRIS promotes and contributes to the FAIRness of data and infrastructures, and FAIR principles push for further improvement in the underlying CRIS. This is another direction of our future work, while the first steps in this direction are taken in (Azeroual et al. 2022).

References

1. Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. 11th Annual Conference on Innovative Data Systems Research (CIDR '21), January 11–15, 2021, Online. https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
2. Azeroual, O.; Saake, G.; Abuosba, M.; Schöpfel, J. (2019). Solving Problems of Research Information Heterogeneity During Integration – Using the European CERIF and German RCD Standards as Examples. Information Services & Use, 39(1–2):105–122. https://doi.org/10.3233/ISU-180030
3. Azeroual, O. (2020). Data Wrangling in Database Systems: Purging of Dirty Data. *Data*, 5(2):50. https://doi.org/10.3390/data5020050
4. Azeroual, O. (2022). Big Research Information in Data Lake. *Academia Letters*, Article 4532. https://doi.org/10.20935/AL4532
5. Azeroual, O.; Schöpfel, J.; Pölönen, J.; Nikiforova, A. (2022). Putting FAIR principles in the context of research information: FAIRness for CRIS and CRIS for FAIRness. 3rd Workshop on Reframing Research (submitted 2022, https://refresh2022.infrascience.isti.cnr.it/)
6. Begoli, E.; Goethert, I.; Knight, K. (2021). A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4643-4651). https://doi.org/10.1109/BigData52589.2021.9671534
7. Endel, F.; Piringer, H. (2015). Data Wrangling: Making data useful again. *IFAC-PapersOnLine*, 48, pp. 111–112. https://doi.org/10.1016/j.ifacol.2015.05.197
8. Fang, F. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 820–824. https://doi.org/10.1109/CYBER.2015.7288049
9. Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. (2019). Leveraging the Data Lake: Current State and Challenges. In: Ordonez C., Song IY., Anderst-Kotsis G., Tjoa A., Khalil I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2019, Lecture Notes in Computer Science, 11708. Springer, Cham. https://doi.org/10.1007/978-3-030-27520-4_1
10. Gorelik, A. (2016). The Enterprise Big Data Lake. Hrsg. von T. McGovern. O'Reilly Media, Inc., ISBN: 9781491931554
11. Hai, R.; Geisler, S.; Quix, C. (2016). Constance: An intelligent data lake system. In: *Proceedings of the 2016 International Conference on Management of Data*, ACM, pp. 2097–210. https://doi.org/10.1145/2882903.2899389
12. Jeffery, K. G. (2004). The new technologies: can CRISs benefit?. *7th International Conference on Current Research Information*

Systems, Antwerp, May 13–15. http://hdl.handle.net/11366/311

13. Jeffery, K.; Asserson, A. (2009). Institutional repositories and current research information systems. *New Review of Information Networking*, 14(2): 71–83. https://doi.org/10.1080/13614570903359357
14. Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; van Ham, F.; Riche, N.H.; Weaver, C.; Lee, B.; Brodbeck, D.; Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4): 271–288. https://doi.org/10.1177/1473871611415994
15. Kutay, J. (2022). Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns. [Online]. [cit. 12.03.2022] https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/
16. Low, W. L.; Lee, M. L.; Ling, T. W. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems*, 26(8): 585–606. https://doi.org/10.1016/S0306-4379(01)00041-2
17. Mathis, C. (2017). Data Lakes. *Datenbank Spektrum*, 17, 289–293. https://doi.org/10.1007/s13222-017-0272-7
18. McCallum, Q.E. (2012). Bad Data Handbook; O'Reilly Media: Sebastopol, Canada, ISBN: 9781449321888
19. Miloslavskaya, N.; Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, 88, pp. 300–305. https://doi.org/10.1016/j.procs.2016.07.439
20. Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing*, 8(3): 391–432. https://doi.org/10.22364/bjmc.2020.8.3.02
21. Oreščanin, D.; Hlupić, T. (2021). Data Lakehouse a Novel Step in Analytics Architecture. *44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, IEEE, (pp. 1242–1246). https://doi.org/10.23919/MIPRO52101.2021.9597091
22. Otto, B.; Lee, Y.W.; Caballero, I. (2016). Information and data quality in networked business. *Electron Markets*, 21, pp. 79–81. https://doi.org/10.1007/s12525-011-0062-2
23. Rattenbury, T.; Hellerstein, J.; Heer, J.; Kandel, S.; Carreras, C. (2017). Principles of Data Wrangling: Practical Techniques for Data Preparation, 1st ed.; O'Reilly Media: Sebastopol, CA, USA.
24. Ravat, F.; Zhao, Y. (2019). Data lakes: Trends and perspectives. *International Conference on Database and Expert Systems Applications*, pp. 304–313. Springer, Cham. https://doi.org/10.1007/978-3-030-27615-7_23
25. Ravat, F.; Zhao, Y. (2019). Metadata management for data lakes. *European Conference on Advances in Databases and Information Systems (ADBIS 2019)*, vol. 1064, pp. 37–44. Springer, Cham. https://doi.org/10.1007/978-3-030-30278-8_5
26. Redman, T. (1998). The impact of poor data quality on the typical entreprise. *Communications of the ACM*, 41(2): 79–82. https://doi.org/10.1145/269012.269025
27. Schöpfel, J.; Prost, H.; Rebouillat, V. (2017). Research data in current research information systems. *Procedia Computer Science*, 106, 305–320. https://doi.org/10.1016/j.procs.2017.03.030
28. Schöpfel, J.; Azeroual, O.; Saake, G. (2020). Implementation and user acceptance of research information systems: An empirical survey of German universities and research organisations. *Data Technologies and Applications*, 54(1): 1–15. https://doi.org/10.1108/DTA-01-2019-0009
29. Sharma, B. (2018). Architecting Data Lakes - Data Management Architectures for Advanced Business Use Cases. 2. Aufl. O'Reilly Media, Inc., ISBN: 9781491952597
30. Strong, D.M.; Lee, Y.W.; Wang, R.Y. (1997). Data quality in context. *Communications of the ACM*, 40(5): 103–110. https://doi.org/10.1145/253769.253804
31. Terrizzano, I. G.; Schwarz, P. M.; Roth, M.; Colino, J. E. (2015). Data Wrangling: The Challenging Journey from the Wild to the Lake. *7th Biennial Conference on Innovative Data Systems Research (CIDR'15)*, January 4–7, Asilomar, California, USA. https://www.cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf
32. Wang, R.Y.; Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers? *Journal of Management Information Systems*, 12(4): 5–33. http://www.jstor.org/stable/40398176
33. Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2): 58–65. https://doi.org/10.1145/269012.269022

34. Yair, G. (2021). Managing Minds: The Challenges of Current Research Information Systems for Improving University Performance. In: Sinuany-Stern, Z. (eds) Handbook of Operations Research and Management Science in Higher Education. International Series in Operations Research & Management Science, vol. 309. Springer, Cham. https://doi.org/10.1007/978-3-030-74051-1_4
35. Zhao, Y.; Megdiche, I.; Ravat, F (2021). Data Lake Ingestion Management. ArXiv, abs/2107.02885, pp. 1–12. https://doi.org/10.48550/arXiv.2107.02885