COMBINING DATA LAKE AND DATA WRANGLING FOR ENSURING DATA QUALITY IN CRIS

Otmane Azeroual[1*],

Joachim Schöpfel[2],

Dragan Ivanovic[3],

Anastasija Nikiforova[4,5]

(1) German Centre for Higher Education Research and Science Studies (DZHW), Germany

(2) GERiiCO-Labor, University of Lille, France

(3) University of NoviSad, Serbia

(4) University of Tartu, Institute of Computer Science, Estonia

(5) European Open Science Cloud Task Force «FAIR metrics and data quality», Belgium

15th International Conference on Current Research Information Systems (CRIS2022)
Dubrovnik, Croatia, May 12-14, 2022

BACKGROUND AND MOTIVATION

Today, billions of data sources continuously generate, collect, process, and exchange data. With the rapid increase in the number of devices and information systems in use, the amount and variety of data are increasing. This is also the case for the research / scientific domain.

Researchers as the end-users of RIS should be able to integrate ever-increasing volumes of data into their institutional database such as Current Research Information Systems (CRIS), regardless of the source, format or amount/ size of research information, where the data quality, flexibility and scalability in connecting and processing different data sources are crucial.

an effective mechanism should be employed to ensure faster value creation from these data

BACKGROUND AND MOTIVATION

an effective mechanism should be employed to ensure faster value creation from these data

DATA LAKE + DATA WRANGLING

This study sets out the concept of a data lake with data wrangling process to be used in CRIS to clean up data from heterogeneous data sources as it is ingested and integrated.

euro CRIS
THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

DATA LAKE

DATA LAKE ✓ vs DATA WAREHOUSE

Data	Users	Use cases
unstructured	Data Scientists, Data Analysts	Stream Processing, Machine Learning, Real time analysis

Raw

Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI

Data	Users	Use cases
Structured	Business Analysts	Batch Processing, BI, Reporting

Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc

Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/

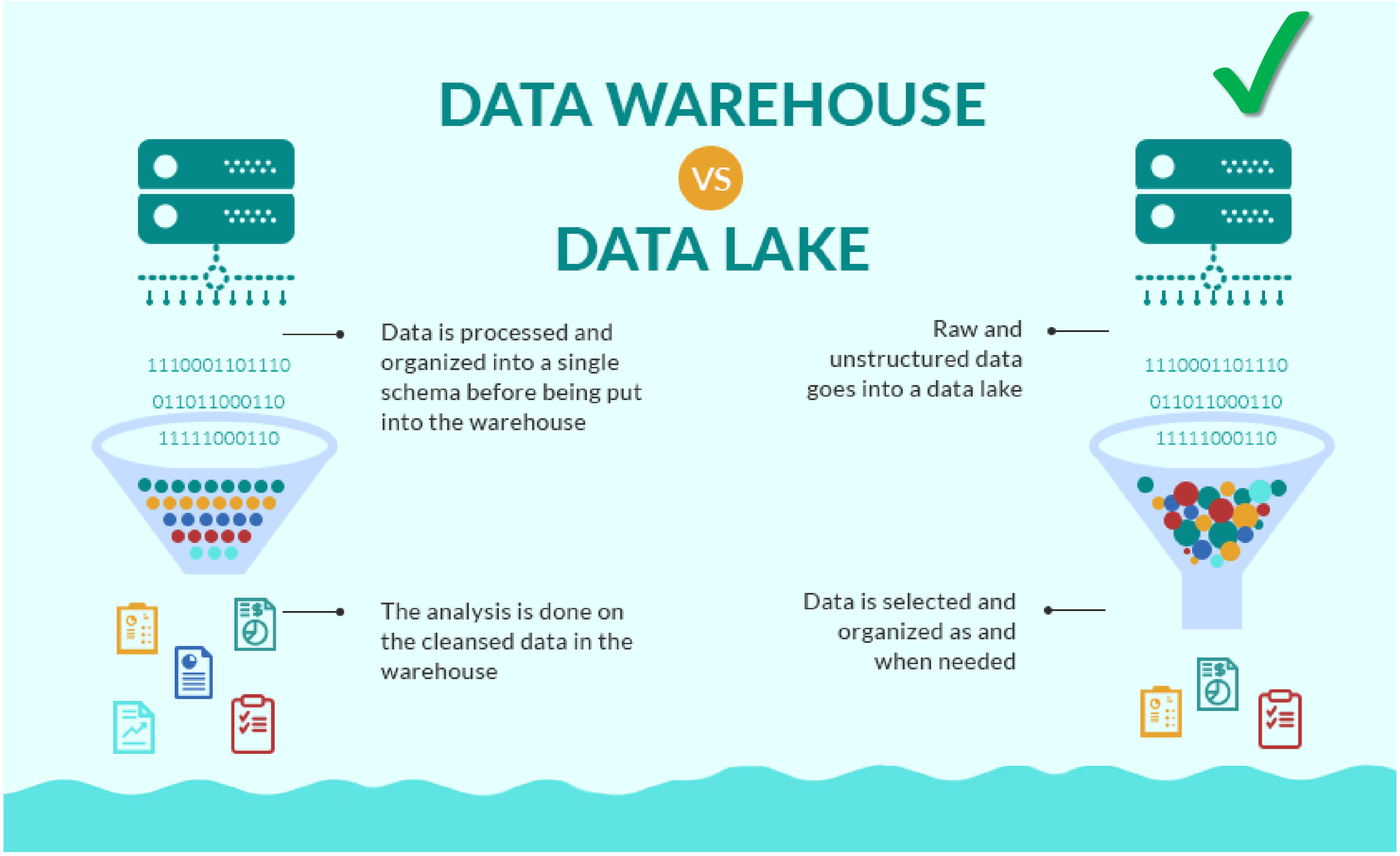
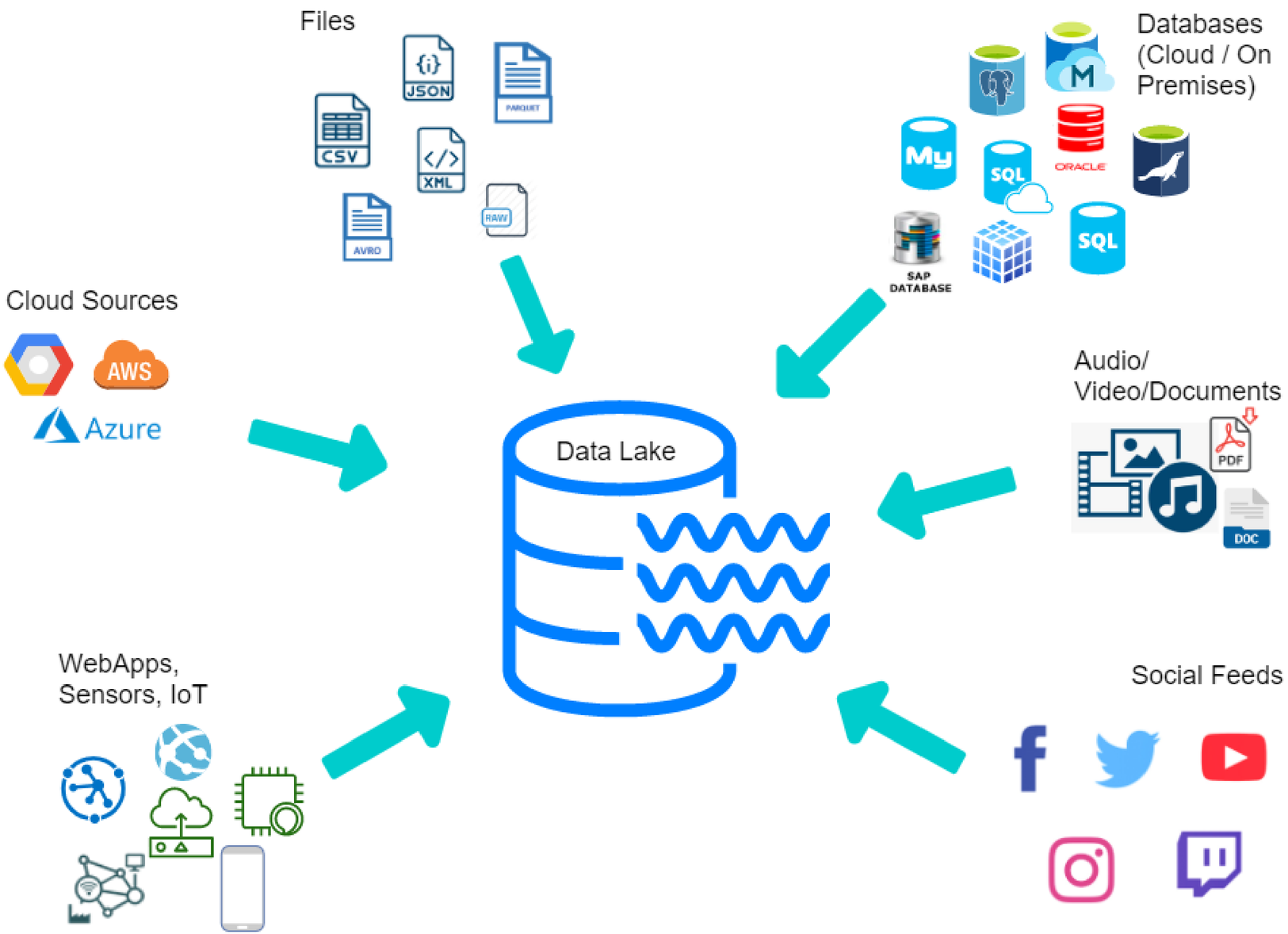
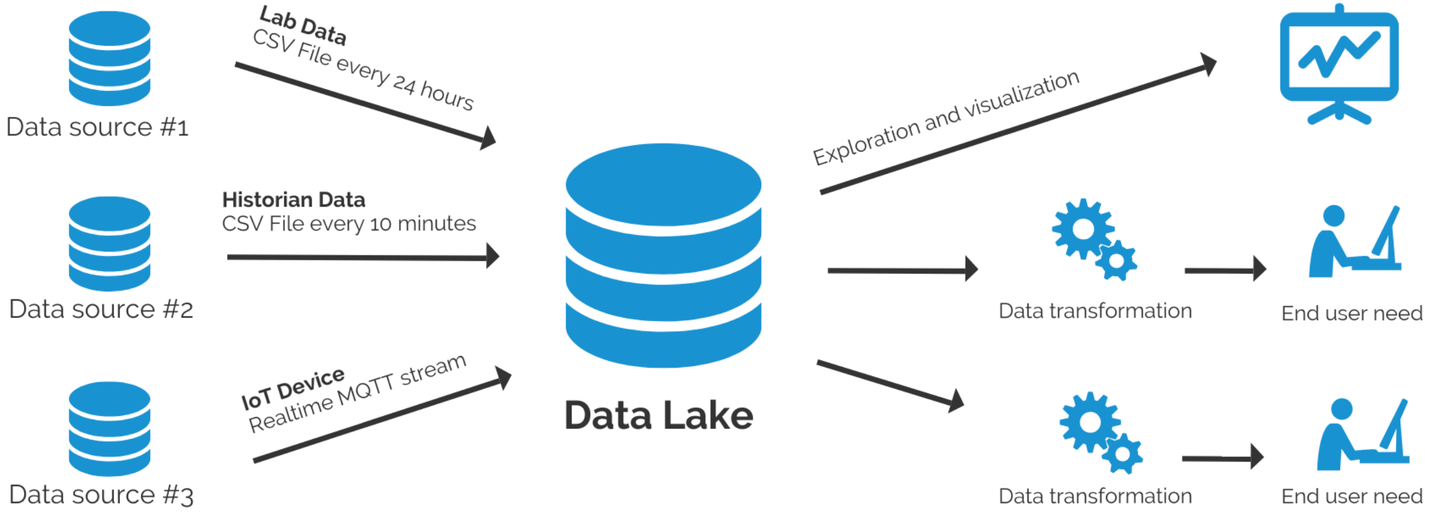


Image source: https://www.grazitti.com/blog/data-lake-vs-data-warehouse-which-one-should-you-go-for/, https://www.qubole.com/data-lakes-vs-data-warehouses-the-co-existence-argument/



- **Data Lake provides a scalable platform for storing and processing large amounts of research data from various sources in their original raw format, regardless of their type, i.e., structured or unstructured data or text, numeric, images, video etc.**
- **The raw data are not cleaned, validated, or transformed ➔ they are original data in their original format.**

Image source: https://towardsdatascience.com/augment-your-data-lake-analytics-with-snowflake-b417f118



- **When storing data / research information, the completeness of data and reduction of the cycle time between data generation and availability are important.**
- **The lack of pre-processing does not slow down data supply and does not lead to data loss.**

- **The concept of data lake allows to store a variety of data within the memory**

BUT

- **there is a need to clean up dirty data and enrich them in a pre-processing process, where data wrangling is found to be suitable for these purposes.**

- **The goal is to convert complex data types and data formats into structured data without programming efforts ➔ users should be able to prepare and transform their research information without the need of using the ETL tools or familiarity and use of programming languages, where these transformations should be automatically suggested after reading the data based on machine learning algorithms that greatly speeds up this process.**

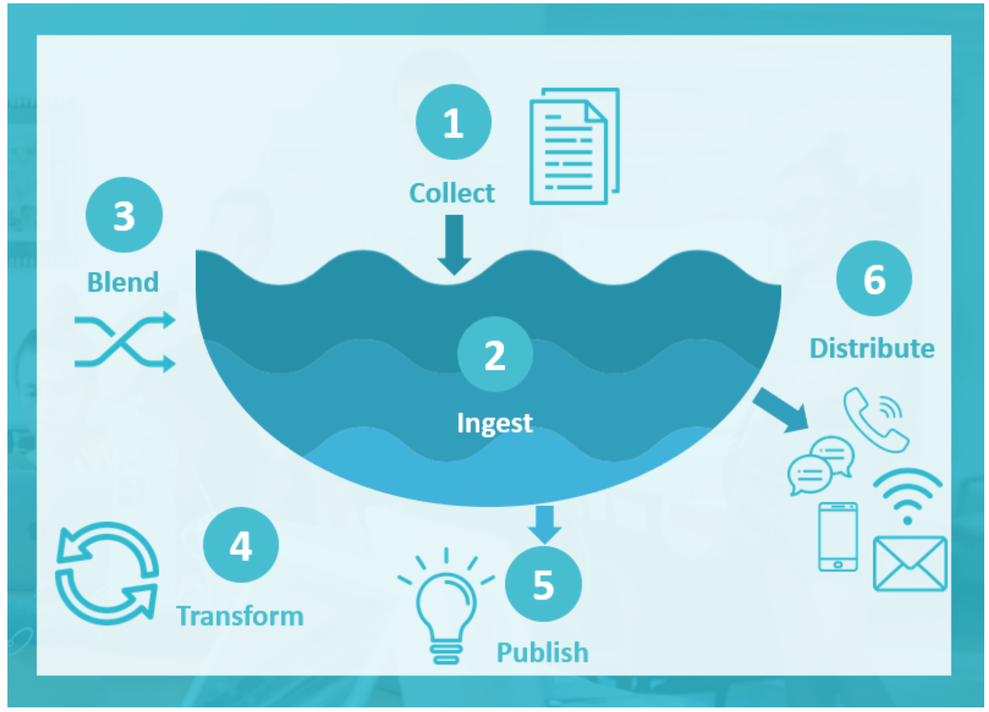


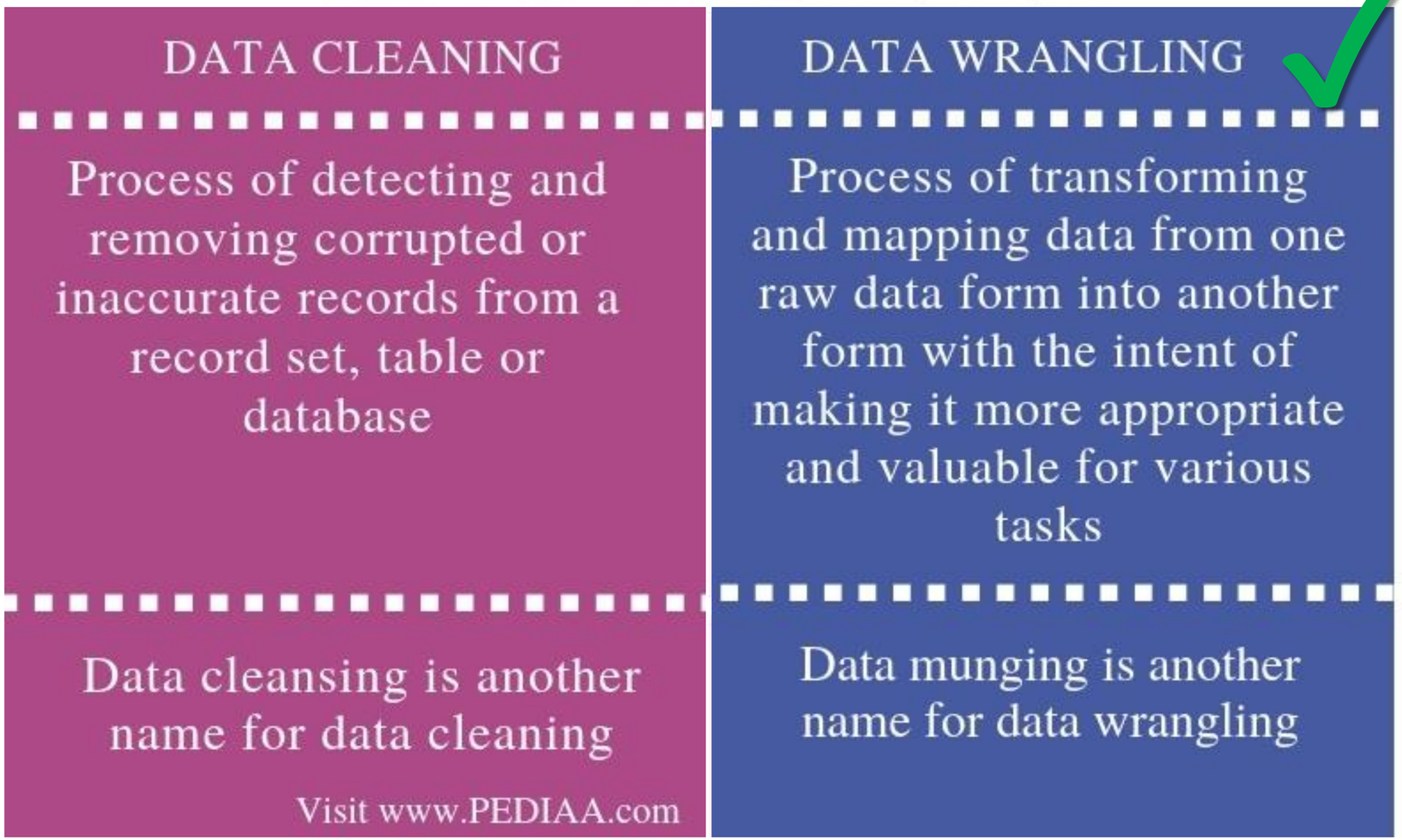
Image source: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.ecloudvalley.com%2Fwhat-is-datalake-and-datawarehouse%2F&psig=AOvVaw1vGuz42Qfu2-0J_bpmhpbJ&ust=1651908882490000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCNiaquD6yvcCFQAAAAAdAAAAABAO

euro CRIS
THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

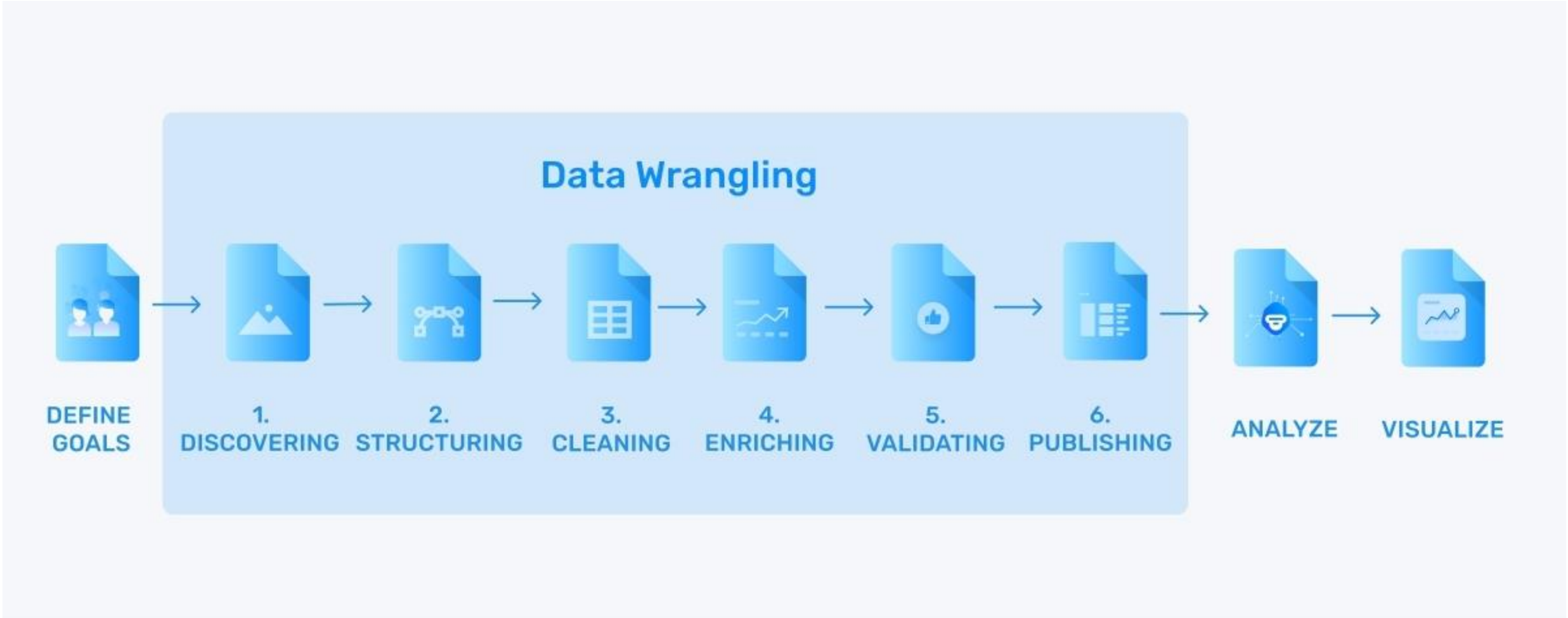
DATA WRANGLING

euro CRIS THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

DATA WRANGLING VERSUS DATA CLEANING



https://pediaa.com/what-is-the-difference-between-data-wrangling-and-data-clean



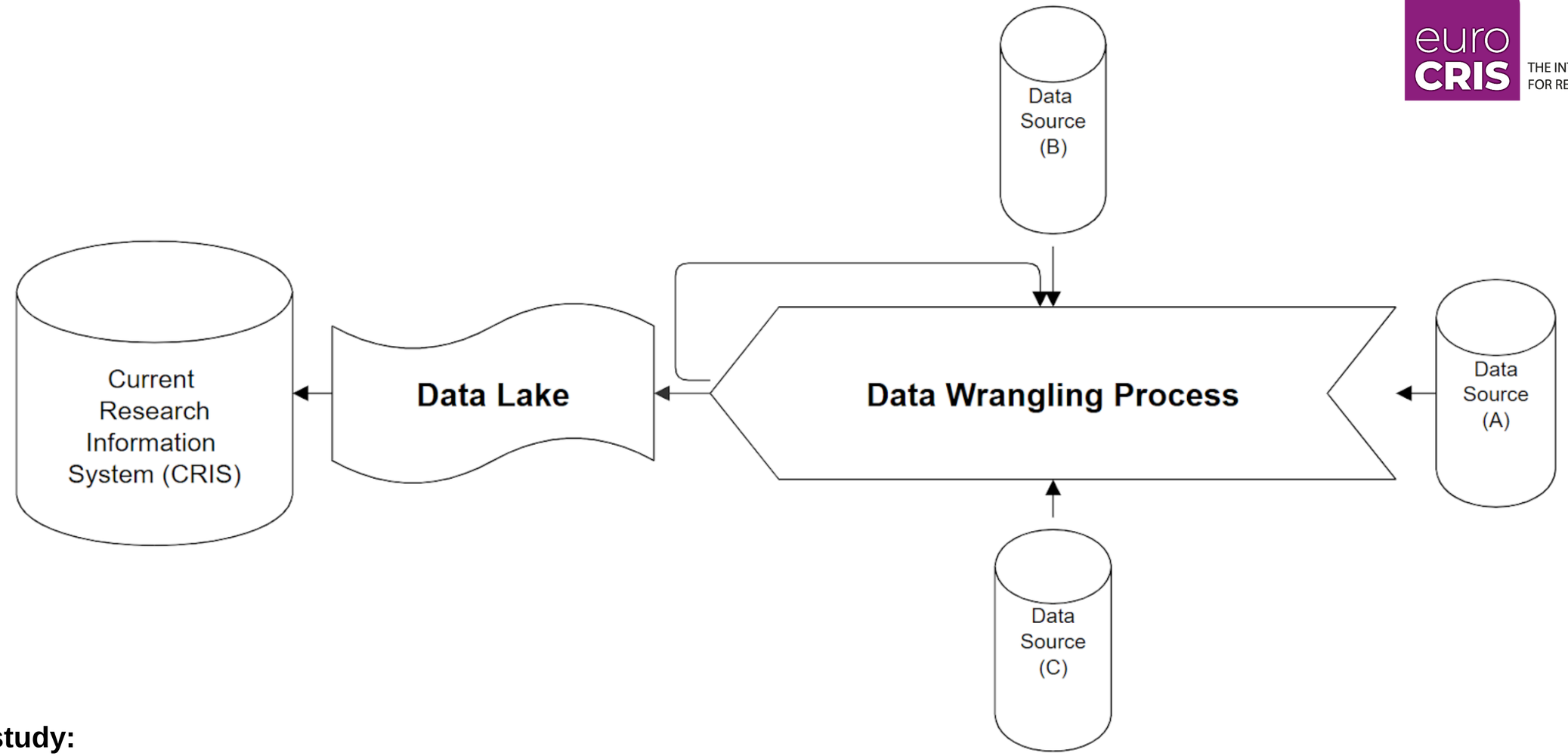
With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

Source: https://monkeylearn.com/blog/data-wrangling/, https://www.altair.com/what-is-data-wrangling/

euro CRIS
THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

DATA LAKE + DATA WRANGLING
=
DATA QUALITY IN CRIS



In this study:

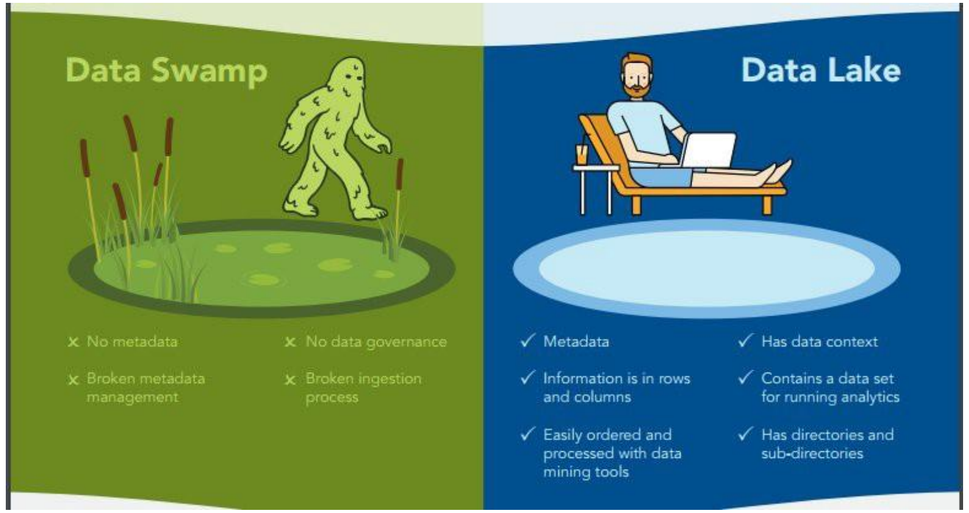
- **an architectural model is first designed and specified, which analyzes the research information, adjusts it and transforms it into CRIS;**
- **a data lake makes both structured and unstructured data available in a reliable, trustworthy, secure and controlled way;**
- **the data wrangling process is used to verify and improve the quality of data, which also protects data from misuse ➔ data are properly updated, retained, and eventually deleted according to the stage of its lifecycle.**

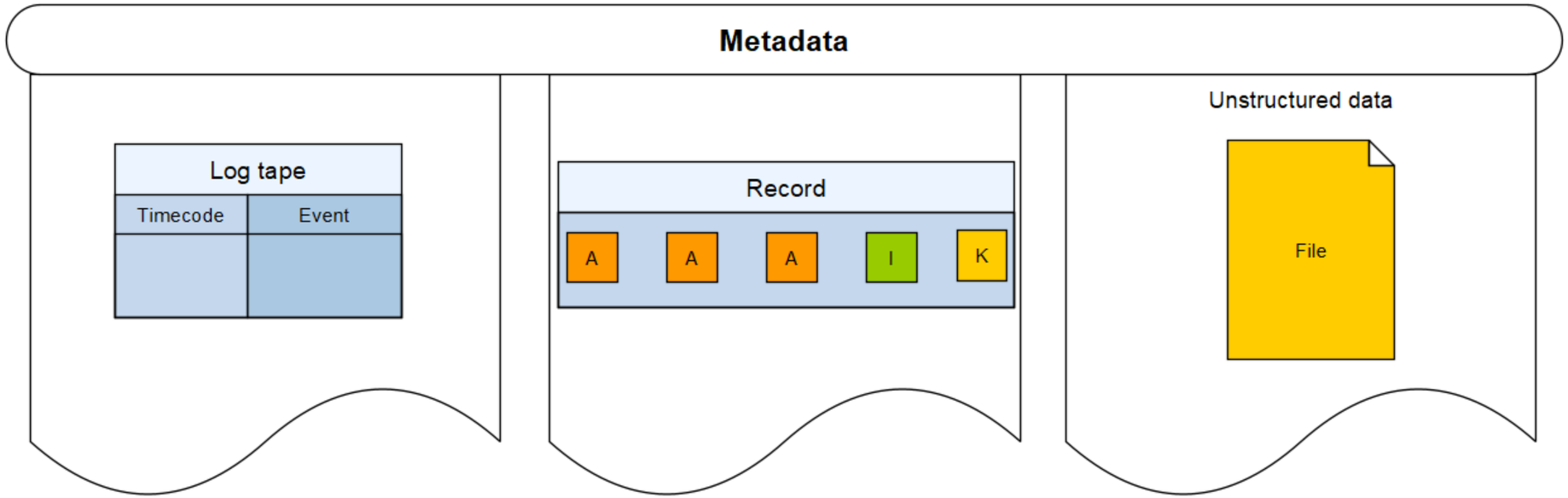
The data wrangling process consists of several sequential steps. Depending on the IS and the desired or required target quality*, these individual steps should be carried out several times ➔ data wrangling is a continuous process that repeats itself repeatedly at regular intervals.

SEVERAL ASPECTS AFFECTING A DATA LAKE

euroCRIS THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

Aspect	Description
Metadata	describes a dataset in more detail containing data about the origin, structure and content of the data + sorting, filtering or categorizing properties + are used for system management and administration
Data mapping	describes the context of the data ➔ integration map - a detailed specification of which application data from which data sources are linked / associated with which characteristics (mostly metadata)
Data lake context	describes the higher-level use case on which the data lake is based ➔ 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 and their context so that they can be better classified for analysis purposes. The context for records can be data origin, categorization, or other contextual feature in the metadata
Processing logging	refers to the raw data processing that takes place in the data lake. The data record and its metadata are manipulated in the process ➔ is of particular interest to data analysts to analyze data lake usage, data set and use case





- **Analog data: data sources automatically generate data in a specific predefined and therefore known data format. Due to the automatic generation, 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 but are significantly different from analog data in their origin - analog data typically represents physical measurement data, application data arises during the operation, transactions of an application (e.g. transmitted system data or analysis data). So-called "records" are used as a common storage solution for these data characterized by 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 DBMS.

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

DATA WRANGLING

euro CRIS THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

- 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.
- The use of data wrangling eliminates low-quality data, i.e. redundant, incomplete, inaccurate or incorrect data, etc., in order to preserve only high-quality research information from which the reliable and value-adding knowledge can be obtained.
- This adjusted research information is then entered into the appropriate target CRIS system to be used in further phases of the analysis (e.g. by analytical applications and protected from unauthorized access by access control).
- This should minimize the effort of analysis and enriching large volumes of data and metadata and achieve far-reaching added value in the procurement of information staff, developers and end users of the CRIS.

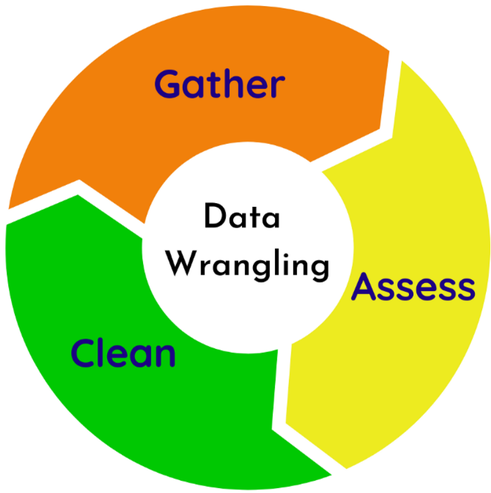
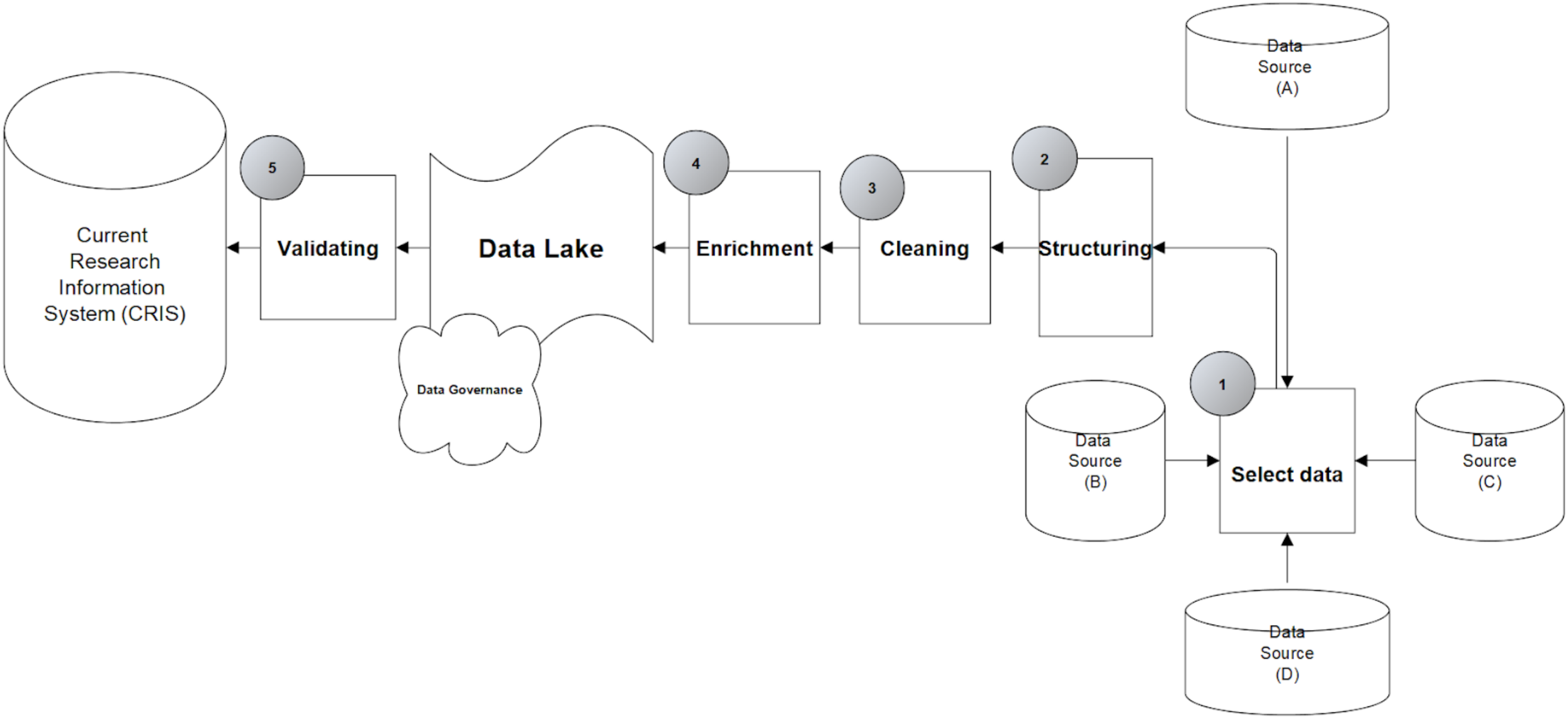


Image source: https://towardsdatascience.com/data-wrangling-raw-to-clean-transformation-b30a27bf4b3b

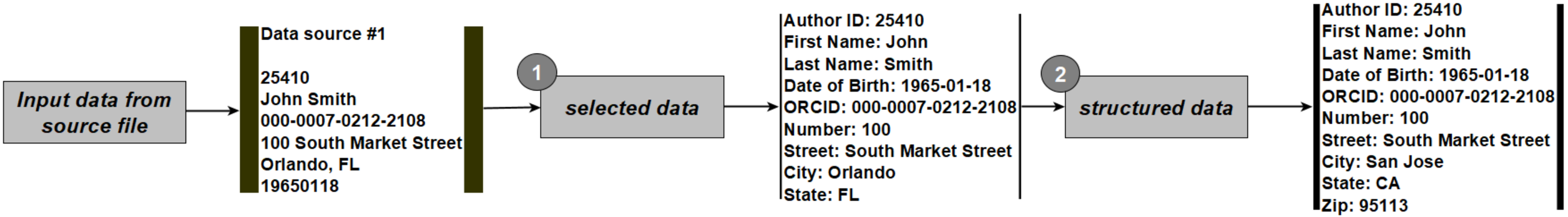


The data wrangling process (steps of the process are indicated by numbers) to prepare research information and integrate it into CRIS

Step	Description
Select data	**The required data records are identified in different data sources. 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**
Structuring	**In most cases, there is little or no structure in the data ➔ change the structure of the data for easier accessibility.**
Cleaning	**Almost every dataset contains some outliers that can skew the analysis results ➔ the data are extensively cleaned for better analysis *(processing of null values, removing duplicates and special characters, and standardization of the formatting to improve data consistency)***
Enrichment	**The data needs to be enriched - an inventory of the data set and a strategy for improving it by adding additional data should be carried out. The data set is enriched with various metadata:** **✓ Schematic metadata provide basic information about the processing and ingestion of data ➔ the data wrangler analyzes / parses data records according to an existing schema.** **✓ Conversation metadata are exchanged between accessing instances with the idea to document information obtained during the processing or analysis of these data for subsequent users.** **The recognized peculiarities/ features of a data set can be saved.**
***Data lake**	**The physical transfer of data in the data lake. Although data are prepared using metadata, the record is not pre-processed.** **The goal is to avoid a data swamp ➔ estimate the value of the data and decide on their lifespan depending on the data quality and its interconnectedness with the rest of the DB.** **Analyzes are not performed directly in the data lake, but only on the relevant data. To be able to use the data, the requester needs the appropriate access rights ➔ Data Wrangler performs data extraction, however, general viewing and exploration of the data should be possible directly in the data lake.**
***Data governance**	**The contents of the data lake, technologies and hardware used are subject to change ➔ an audit is required to take care of the care and maintenance of the data lake. The main principles / guidelines and measures that regulates data maintenance coordinating all processes in the data lake and responsibilities are defined**
Validating	**the data are checked one more time before they are integrated into the target CRIS to identify problems with the data quality and consistency of the data, or to confirm that the transformation has been successful.** **Verify that the values of the attribute are correct and conform to the syntactic and distribution constraints, thus ensuring high data quality AND document every change so that older versions can be restored, or history of changes can be viewed. 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 validating and structuring the data.***

At the end of this process, research information can be used by analytical applications and protected from unauthorized access by access control

USE-CASE



- ✓ **Data formatting**
- ✓ **Correction of incorrect data (e.g. address data)**

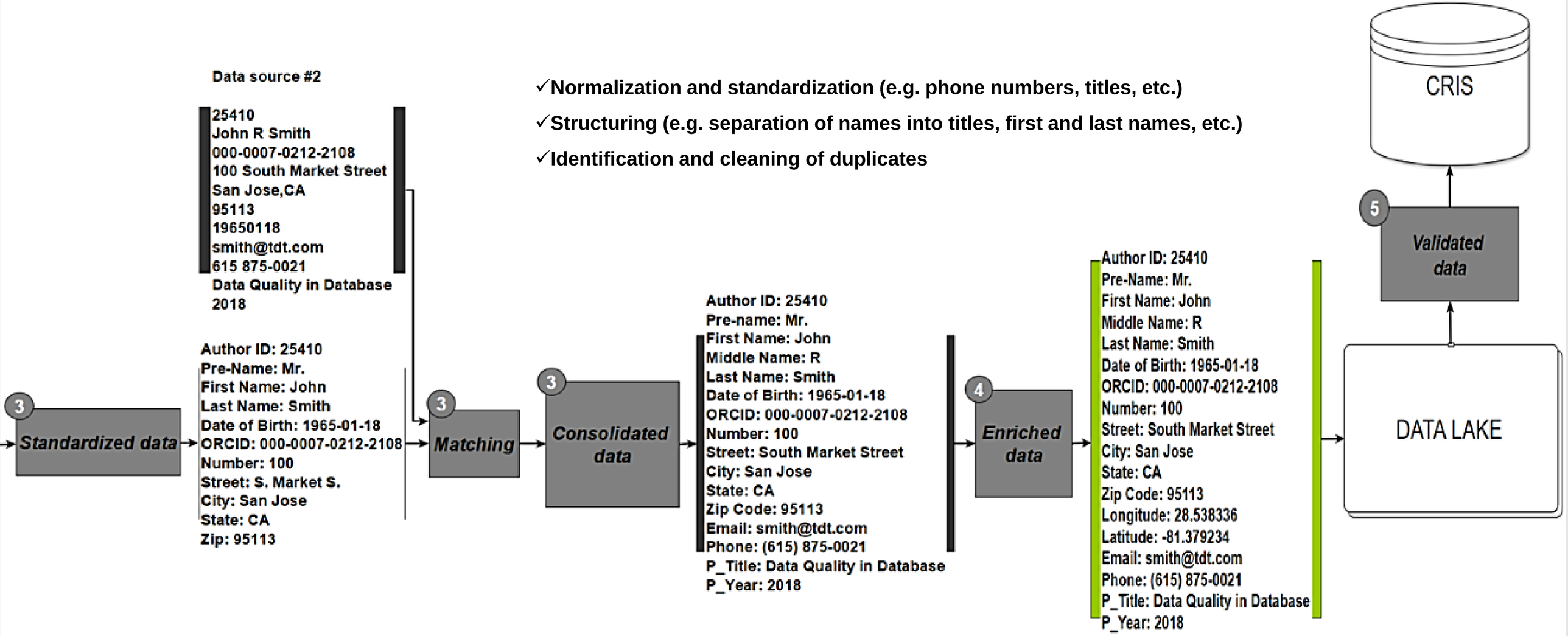
USE-CASE

euro CRIS THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

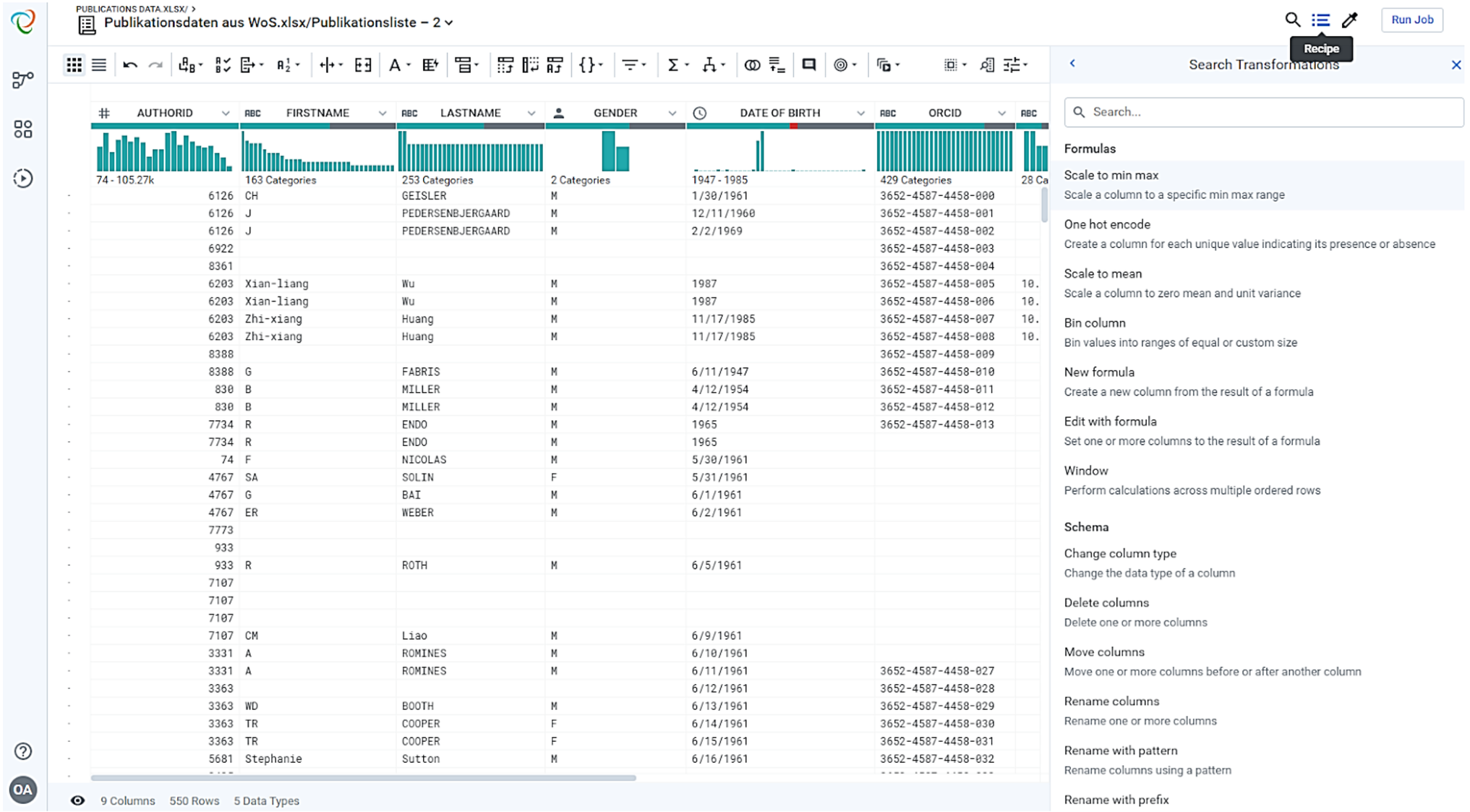
✓Normalization and standardization (e.g. phone numbers, titles, etc.)

✓Structuring (e.g. separation of names into titles, first and last names, etc.)

✓Identification and cleaning of duplicates



USE-CASE: TRIFACTA FOR DATA WRANGLING



CONCLUSIONS

- ✓ 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 in order to be useful remain critical but at the same time become more difficult to fulfill ➔ data wrangling can be seen 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 provide valuable support for predictive analytics and self-service analysis by making it easier and faster to access large amount of data from multiple sources.
- ✓ The proper organization of the data lake makes it easier to find the research information the user needs. Managing the data that have already been pre-processed results in an increased efficiency and cost saving, as preparing data for their further use is the most resource-consuming part of data analysis. By providing pre-processed research information, users with limited or no experience in data preparation (low level of data literacy) can be supported and analyzes can be carried out faster and more accurately.

LOG IN

EUROPEAN OPEN SCIENCE CLOUD

EOSC STRATEGIC AGENDA + ADVISORY GROUPS + GET INVOLVED + NEWS & EVENTS + ASSOCIATION + CONTACT

Complete the Data Quality survey!

12 April 2022

The Data Quality subgroup of the EOSC Association's Task Force Fair metrics and Data Quality created a survey intended to capture needs and expectations for data quality from a wide range of stakeholders. Your responses will help to identify common approaches and formulate recommendations to support data quality within EOSC (European Open Science Cloud).

Fill in the survey, it will take around **10 minutes** of your time and will be open for you to respond till **Thursday 5th May**.
By completing all answers, we are happy to acknowledge your contribution in the forthcoming EOSC data quality recommendations.

Please email us back (carlo.lacagnina@bsc.es) if you would like your name to be included in the formal acknowledgement. For any further feedback do not hesitate to contact us for an individual interview.
Your opinion is very important for shaping the EOSC!

Fill in the survey

euroCRIS THE INTERNATIONAL ORGANISATION FOR RESEARCH INFORMATION

" EOSC - Data Quality Task Force subgroup

This survey is intended to capture needs and expectations for data quality from a wide range of stakeholders. Your responses will help us to identify common approaches and formulate recommendations to support data quality within EOSC (European Open Science Cloud).

Your opinion is very important for shaping the EOSC!

The survey is split into three thematic blocks, the expected time to fill it in is 10 minutes and it will be open for you to respond till Thursday 5th May.

EUROPEAN OPEN SCIENCE CLOUD

Continue press Enter

EOSC Data Quality survey

EUROPEAN OPEN SCIENCE CLOUD

THANK YOU FOR ATTENTION!

QUESTIONS?

For questions or any other queries, contact us via email - Nikiforova.Anastasija@gmail.com, azeroual@dzhw.eu