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

Otmane Azeroual, Joachim Schöpfel, Dragan Ivanovic, Anastasija Nikiforova

Otmane Azeroual
German Centre for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany
ORCID: **0000-0002-5225-389X**

Joachim Schöpfel
GERiiCO-Labor, University of Lille, 59650 Villeneuve-d'Ascq, France
ORCID: **0000-0002-4000-807X**

Dragan Ivanovic
University of Novi Sad, Novi Sad 21000, Serbia
ORCID: **0000-0002-9942-5521**

Anastasija Nikiforova
University of Tartu, Institute of Computer Science, Narva mnt 18, 51009 Tartu, Estonia
ORCID: **0000-0002-0532-3488**

Type of contribution: research paper

CRIS 2022 conference call for papers https://eurocris.org/cris2022-conference-call-proposals

Extended abstract

Today, researchers 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. Then, an effective mechanism should be employed to ensure faster value creation from these organizations' data, respecting this growing data variety, i.e. heterogeneous data. The processing of electronic data plays a central role in modern society. Data in general is an elementary component of operational processes in companies and scientific organizations. They are also the basis for decision making. Poor quality research information can negatively impact results and decisions. The quality of research information, or the trustworthy and reliability of the data, is critical. This is closely linked to the topic of data lake and data wrangling. 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. The raw data are not cleaned, validated, or transformed; in fact, they are original data in their original format. In addition to text or numeric data, the data lake can also record images, video or other data formats - there are no restrictions on data types. The data can be stored in the cloud or locally. When storing research data/ 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 different data structures and thus allows to store a variety of data within the memory. This means that 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 aim is to convert complex data types and data formats into structured data without programming efforts. In other words, users should be able to prepare and transform their research information without the need of using the ETL (Extract-Transform-Load) tools or familiarity and use of programming languages (e.g. Java, Python or SQL). These transformations are automatically suggested after reading the data based on machine learning algorithms that greatly speeds up this process.

Consolidation of research information improves data quality and reduces duplicates between systems, and provides flexibility and scalability in connecting and processing different data sources. 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. 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. This research paper sets out the concept of a data lake with data wrangling process and shows how it can be used in CRIS to clean up data from heterogeneous data sources as it is ingested and integrated. In this paper, an architectural model is first designed and specified, which analyses the research information, adjusts it and transforms it into CRIS (see Fig. 1). 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. This ensures that 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 information system and the desired or required target quality, these individual steps should be carried out several times. In many cases, data wrangling is a continuous process that repeats itself repeatedly at regular intervals.

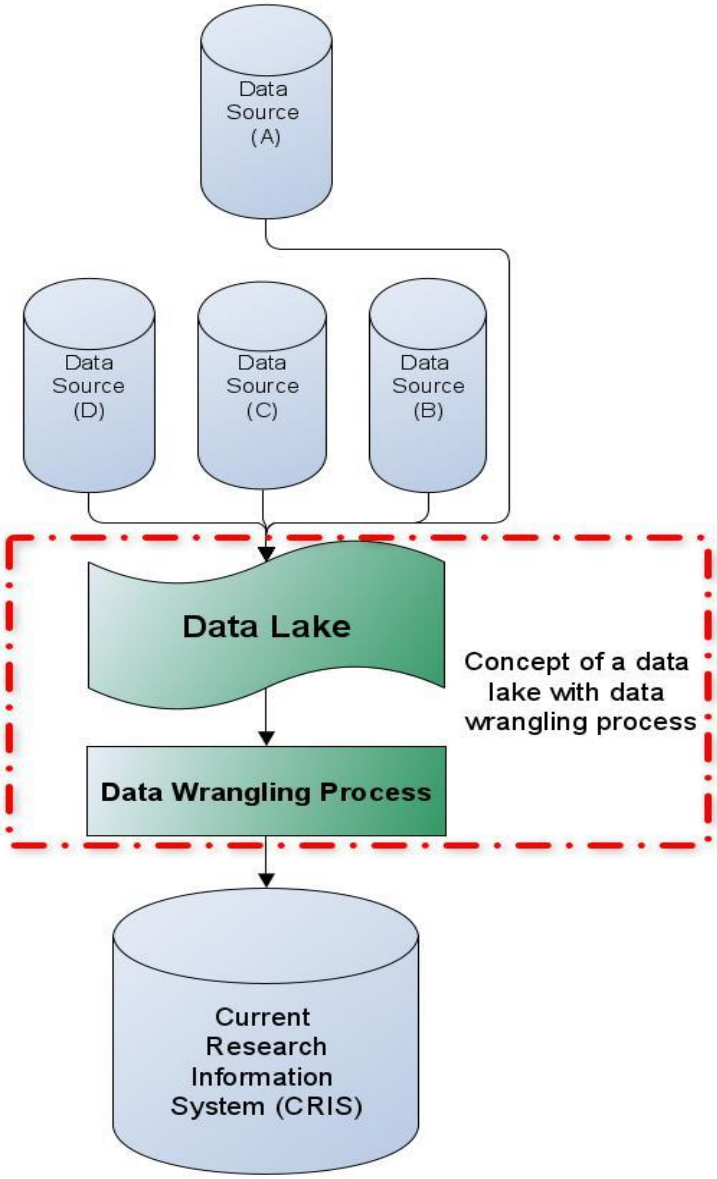

Fig. 1. Architecture of a controlled data lake with data wrangling process

Keywords

Current research information system, CRIS, research information, metadata management, dirty data identification, data quality, data curation, data management, data lake, data wrangling, research information system, RIS

References

- Azeroual, O. Data Wrangling in Database Systems: Purging of Dirty Data. *Data*, **2020**, 5(2):50.
- Endel, F.; Piringer, H. Data Wrangling: Making data useful again. *IFAC-PapersOnLine*, **2015**, 48, pp. 111–112.

- Fang, F. 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, **2015**.
- Giebler C., Gröger C., Hoos E., Schwarz H., Mitschang B. 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, vol. 11708. Springer, Cham, **2019**.
- Gorelik, A.. The Enterprise Big Data Lake. Hrsg. von T. McGovern. O'Reilly Media, Inc., **2016**.
- Hai, R.; Geisler, S.; Quix, C. Constance: An intelligent data lake system. In: *Proceedings of the 2016 International Conference on Management of Data,* ACM, pp. 2097–210, **2016**.
- Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; van Ham, F.; Riche, N.H.; Weaver, C.; Lee, B.; Brodbeck, D.; Buono, P. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *J. Inf. Vis.,* 10, pp. 271–288, **2011**.
- Kandel, S.; Paepcke, A.; Hellerstein, J.; Heer, J. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the International Conference on Human Factors in Computing Systems*, CHI 2011, Vancouver, BC, Canada, 7–12 May, pp. 3363–3372, **2011**.
- McCallum, Q.E. Bad Data Handbook; O'Reilly Media: Sebastopol, Canada, **2012**.
- Miloslavskaya, N.; Tolstoy, A. Big Data, Fast Data and Data Lake Concepts, *Procedia Computer Science*, vol. 88, pp. 300–305, **2016**.
- Otto, B.; Lee, Y.W.; Caballero, I. Information and data quality in networked business. *Electron. Mark.,* 21, pp. 79–81, **2016**.
- Rattenbury, T.; Hellerstein, J.; Heer, J.; Kandel, S.; Carreras, C. Principles of Data Wrangling: Practical Techniques for Data Preparation, 1st ed.; O'Reilly Media: Sebastopol, CA, USA, **2017**.
- Ravat, F.; Zhao, Y. Data lakes: Trends and perspectives. *International Conference on Database and Expert Systems Applications*, pp. 304–313. Springer, **2019**.
- Ravat, F.; Zhao, Y. Metadata management for data lakes. *European Conference on Advances in Databases and Information Systems*, pp. 37–44. Springer, **2019**.
- Redman, T. The impact of poor data quality on the typical entreprise. *Commun. ACM*, 41, pp. 79–82, **1998**.
- Sharma, B. Architecting Data Lakes - Data Management Architectures for Advanced Business Use Cases. 2. Aufl. O'Reilly Media, Inc., **2018**.
- Strong, D.M.; Lee, Y.W.; Wang, R.Y. Data quality in context. *Commun. ACM*, 40, pp. 103–110, **1997**.

- Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers? *J. Manag. Inf. Syst.*, 12, pp. 5–33, **1996**.
- Wang, R.Y. A product perspective on total data quality management. Commun. ACM, 41, pp. 58–65, **1998**.
- Zhao, Y.; Megdiche, I.; Ravat, F. Data Lake Ingestion Management. *ArXiv*, abs/2107.02885**,** pp. 1–12, **2021**.

Authors Profile

Otmane Azeroual, is researcher and project manager at the German Centre for Higher Education Research and Science Studies (DZHW) in Berlin. After studying Business Information Systems at the University of Applied Sciences (HTW) Berlin, he completed his Ph.D. in engineering informatics at the Institute for Technical and Business Information Systems (ITI), Database Research Group of the Otto von Guericke University Magdeburg and at the Department of Computer Science and Engineering of the University of Applied Sciences (HTW) Berlin. Since 2021, he has also been a lecturer in the Digital Business & Data Science course at the University of Europe for Applied Sciences. His research interests lie in the field of database systems, information systems, data quality management, artificial intelligence, business intelligence, big data, open data, IT security, cloud data management and Industry 4.0.

Joachim Schöpfel, is senior lecturer in Library and Information Sciences at the University of Lille (France), researcher at the GERiiCO laboratory and consultant at the Ourouk consulting firm. He was Manager of the INIST (CNRS) scientific library from 1999 to 2008, head of the University of Lille department of information sciences from 2009 to 2012 and director of the French Atelier National de Reproduction des Thèses (ANRT) from 2012 to 2017. He teaches library marketing, auditing, valorization and digitization of cultural heritage collections, intellectual property and information science. His research interests are scientific information and communication, especially open access and open science, research data and grey literature. He is member of euroCRIS, of the NDLTD board of directors.

Dragan Ivanovic, full professor at University of Novi Sad and research software engineer, has a large experience of working on CERIF-based systems development in the context of EC funded projects for euroCRIS. His specialties include: Data Analysis, Data modelling, Research Information Systems, digital and data repositories. He is a member of euroCRIS, and VIVO technical leader. Dragan published more than 50 research articles.

Anastasija Nikiforova, is a researcher (PhD in Computer Science – Data Processing Systems and Data Networking), whose research interests include, but are not limited to, data management with a particular focus on data quality, open (government) data, Smart city, Society 5.0, sustainable development, HCI, Industry 4.0 and digitization. She is an assistant professor of Information Systems at University of Tartu, a part of European Open Science Cloud Task Force “FAIR Metrics and Data Quality”, and visiting researcher at Delft University of Technology, Faculty Technology Policy and Management. Her previous experience includes the role of an assistant professor and researcher in the Innovation Laboratory at University of Latvia, IT-expert at the Latvian Biomedical Research and Study Centre, BBMRI-

ERIC Latvian National Node and an advisor for the Institute for Social and Political Studies (University of Latvia), where she was involved in six ERDF and Horizon 2020 projects with a focus on data management and software engineering.
She is an expert of the Latvian Council of Sciences in (1) Computer Science and Informatics, (2) Electrical Engineering, Electronics, ICT, (3) Social Sciences–Economics and Business, and COST – European Cooperation in Science & Technology. For promotion of open data and technologies the Latvian Open Technologies Association has recognized her as a person of the of the year in 2021. She serves as a PC for 10+ international conferences and invited reviewer for 10+ high-quality journals.