Dataversifying Natural Sciences: Pioneering a Data Lake Architecture for Curated Data-Centric Experiments in Life & Earth Sciences

Genoveva Vargas-Solar, Jérôme Darmont, Alejandro Adorjan, Javier A. Espinosa-Oviedo, Carmem Hara, Sabine Loudcher, Regina Motz, Martin Musicante, José-Luis Zechinelli-Martini

Dataversifying Natural Sciences: Pioneering a Data Lake Architecture for Curated Data-Centric Experiments in Life & Earth Sciences

Genoveva Vargas-Solar[1], Jérôme Darmont[2], Alejandro Adorjan[4], Javier A. Espinosa-Oviedo[1,3], Carmem Hara[5], Sabine Loudcher[2], Regina Motz[6], Martin Musicante[7] and José-Luis Zechinelli-Martini[8]

[1]CNRS, Univ. Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221, France
[2]Université de Lyon, Lyon 2, UR ERIC 5 avenue Mendès France, 69676 Bron Cedex, France
[3]CPE Lyon, 43 Blvd. du 11 Novembre 1918, 69616 Villeurbanne Cedex, France
[4]Unversidad ORT, Montevideo, Uruguay
[5]Universidade Federal do Paranà, Dept. de Informatica, Curitiba - PR, 81531-980, Brazil
[6]Instituto de Computación (INCO) Facultad de Ingeniería, Universidad de la Repúbica, Uruguay
[7]Universidad Federal Rio Grande do Norte, DIMAP, Natal, Brazil
[8]Fundación Universidad de las Américas, Puebla Exhacienda Sta. Catarina Mártir s/n 72820 San Andrés Cholula, Mexico

Abstract
This vision paper introduces a pioneering data lake architecture designed to meet Life & Earth sciences' burgeoning data management needs. As the data landscape evolves, the imperative to navigate and maximise scientific opportunities has never been greater. Our vision paper outlines a strategic approach to unify and integrate diverse datasets, aiming to cultivate a collaborative space conducive to scientific discovery. The core of the design and construction of a data lake is the development of formal and semi-automatic tools, enabling the meticulous curation of quantitative and qualitative data from experiments. Our unique "research-in-the-loop" methodology ensures that scientists across various disciplines are integrally involved in the curation process, combining automated, mathematical, and manual tasks to address complex problems, from seismic detection to biodiversity studies. By fostering reproducibility and applicability of research, our approach enhances the integrity and impact of scientific experiments. This initiative is set to improve data management practices, strengthening the capacity of Life & Earth sciences to solve some of our time's most critical environmental and biological challenges.

Keywords
Life and Earth sciences, data-driven experiments, data lake, data curation

1. Introduction

These days, it is relatively easy and inexpensive to acquire massive amount of data, even in continuous mode. This has been no different for experimental and observational sciences like Life & Earth sciences. Accessibility to data about the Earth and its biodiversity, with varying levels of provenance, quality and reliability, opens up the possibility of constructing different perspectives on the phenomena observed, leading to scientific conclusions with different depths that target a wide range of knowledge consumers (civilians, decision-makers, scientists).

Traditional *schema-on-write* approaches, such as the Extraction, Transformation and Loading (ETL) process, are ineffective for the data management requirements of these experimental sciences. Data lakes are becoming increasingly common for the management and analysis of massive data. Data lakes are repositories that store raw data in its original format. They can be well adapted for storing data harvested from digital sources (observation stations), social media, Web and in situ collectors.

The extraction of value through data-driven experiments in the Life & Earth sciences is determined by two main elements:

- The maintenance of metadata gathering the conditions under which experiments are performed (quantitative perspective) to preserve the memory of the experimental process of knowledge production process, and to enable understanding and reproducibility.
- An open science perspective that can go beyond data sharing and must consider the sharing of

Published in the Workshop Proceedings of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024, Paestum, Italy).
*Genoveva Vargas-Solar.
†The authors' list is alphabetical except for the first two authors.
genoveva.vargas-solar@cnrs.fr (G. Vargas-Solar); jerome.darmont@univ-lyon2.fr (J. Darmont); aadorian@gmail.com (A. Adorjan); javier.espinosa@liris.cnrs.fr (. J. A. Espinosa-Oviedo); carmemhara@ufpr.br (C. Hara); sabine.loudcher@univ-lyon2.fr (S. Loudcher); rmotz@fing.edu.uy (R. Motz); mam@dimap.ufrn.br (M. Musicante); joseluis.zechinelli@udlap.mx (J. Zechinelli-Martini)

know-how, decision-making, expertise, project management, and people within the projects that define the research must be considered.

This vision paper introduces our approach to designing and building a data lake for collecting and integrating data and meta data of Life & Earth sciences' data-driven experiments.

The remainder of the paper is organised as follows. Section 2 gives a general overview of approaches that address curating and managing knowledge in Life & Earth sciences. Section 3 describes the challenges associated with curating data and data-driven experiments in Life & Earth sciences often guided by researchers. In particular, the section gives the general challenges for building data lakes containing curated data and producing knowledge derived from data-driven experiments. Section 4 introduces the general principle for building, maintaining and exploiting a data lake. The data lake allows the creation of "dataverses" that can export the history of the development of experimental processes that lead to knowledge in Life & Earth sciences. Finally, Section 5 concludes the paper and discusses future work.

2. Related work

We introduce the main topics and approaches that underline the vision of maintaining and sharing data to perform data-driven experiments: data harvesting tools, data curation techniques, data labs, data lakes, science lakes and dataverses.

2.1. Data harvesting

Data available on the Web play a determining role in decision-making in personal and corporate life. Collecting and storing this data in a structured model helps integrate them with other sources and use the dataset in various applications, such as event detection and sentiment monitoring. Online newspapers are essential sources of information, accessed daily by thousands of people.

Various works in the literature report manual efforts to extract data from pages on the *Web* [1, 2]. However, these efforts have been eased by applying Web scraping techniques. Some work complements automated extraction processes to obtain clean and analysed data by implementing curation procedures [3]. Among the various existing tools available on the *Web* for data extraction, we can highlight ParseHub[1] is a web scraping tool that facilitates data extraction from websites through an interactive click-based interface, saving the data directly to the cloud in JSON and CSV formats. It navigates through continuation pages and captures complete news articles, with the ability to collect data based on specific character sequences. 80legs[2] offers sequential data extraction from websites. Octoparse[3] simplifies the data extraction process by enabling users to create a scraping workflow with clicks. It includes features like URL and string lists for targeted scraping and ready-to-use templates for popular sites like Amazon and Google. FactExtract [3] is tailored for aggregating content from specific Senegalese news sources, boasting automatic language detection for ten languages, data cleaning, and analysis, all whilst avoiding data duplication. This tool, which utilises Python's Newspaper library, also features automated daily updates for the news content it monitors. *ENoW* - News Data Extractor from the *Web*[4] is a news scrapping system that explores online newspapers. *ENoW* receives search strings as input and stores in a relational database data extracted from the news and their full content.

2.2. Data curation

According to Garcov et al., [4], research data curation is "preparing research data and artefacts for sharing and long-term preservation". Research repositories are the standard for publishing data collections to the research communities. Datasets at an early collection stage are generally not ready for analysis or preservation. Thus, extensive preprocessing, cleaning, transformation, and documentation actions are required to support usability, sharing, and preservation over time [5]. Curated data collections have the potential to drive scientific progress [6], are relevant for reproducibility and improve the reliability of sciences [7]. However, data curation introduces challenges for supporting data-driven applications [8] adopting quanti-qualitative methods. For example, research challenges curating material across time, space and collaborators [7]. Quantitative and qualitative research methodologies apply ad-hoc data curation strategies that keep track of the data that describe the tools, techniques, hypothesis, and data harvesting criteria defined a priori by a scientific team.

Several software tools that apply statistical techniques and machine learning algorithms are available for qualitative researchers. Woods et al. [9] argue that Computer-Assisted Qualitative Data Analysis Software (CAQDAS) is a well-known tool for qualitative research. These tools support qualitative techniques and methods for applying Qualitative Data Analysis (QDA). ATLAS.ti [10], Dedoose [11], MAXQDA [12], NVivo [13] implement the REFI-QDA standard, an interoperability exchange for-

[1] https://www.parsehub.com/

[2] https://80legs.com/

[3] https://www.octoparse.com/

[4] L Reips, M Musicante, G Vargas-Solar, ATR Pozo, C.S Hara, ENoW-Extrator de Dados de Notícias da Web, Demonstration Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados, 2023, 78-83

mat. CAQDAS [14] researchers and practitioners can perform annotation, labelling, querying, audio and video transcription, pattern discovery, and report generation. Furthermore, CAQDAS tools allow the creation of field notes, thematic coding, search for connections, memos (thoughtful comments), contextual analysis, frequency analysis, word location and data analysis presentation in different reporting formats [15]. The REFI-QDA (Rotterdam Exchange Format Initiative)[5] the standard allows the exchange of qualitative data to enable reuse in QDAS [16]. QDA software such as ATLAS.ti [10], Dedoose [11], MAXQDA [12], NVivo [13], QDAMiner [17], Quirkos [18] and Transana [19] adopt REFI-QDA standard.

We assume that data curation consists of identifying, systematizing, managing, and versioning research data, considering versioning artefacts an essential component of tracking changes along the research project.

2.3. Data labs

Data science environments provide data labs like Kaggle[6] and Dryad[7] with stacks of services for (externalised) data storage, tagging and exploring tools. These environments allow a collective sharing space of highly curated data collection maintenance tools. There are specialised repositories like DataOne[8] and data repositories re3data[9].

DataONE (Data Observation Network for Earth) is a community-driven project that provides access to various environmental and ecological data across multiple member repositories. It is designed as an innovative framework aimed at facilitating research and enabling scientists and researchers to preserve, access, use, and increase the impact of their data. The platform provides robust data management tools, ensuring datasets' preservation and integrity. DataONE underscores data stewardship as a federated resource and supports scientific collaboration and reproducibility. It is invaluable for researchers seeking to address complex environmental challenges through shared data and knowledge.

Re3data is a global registry of research data repositories that offers a comprehensive directory for researchers seeking to access, store, share, and manage their datasets. It represents a variety of academic disciplines and provides detailed information about each repository, such as access policies, standards, and contact details. re3data promotes data sharing, visibility, and reuse as a critical reference point for finding suitable repositories for data deposition. The platform enhances transparency in research data management. It supports open science by guiding users to trustworthy and reliable repositories, thereby facilitating the discovery of high-quality data across different scientific fields.

2.4. Data lake, science lake and dataverse

Data lakes are expansive storage repositories that hold vast raw data in their native format until needed. Stein and Morrison [20] emphasised their potential for scalability and flexibility in handling big data from various sources. In recent studies, Dixon in 2010[10] defined the term and its initial application in big data analytics. Quix et al. (2016) [21] delved into the architectural considerations and challenges such as data governance and metadata management.

Science lakes, an offshoot of data lakes, are tailored specifically for the scientific community to address the need for interdisciplinary research, data management and complex analytics. Russom (2016) [22] suggested that science lakes provide a more discipline-specific approach to data handling, enabling better metadata curation and domain-specific data models, which are crucial for reproducibility in scientific research.

A data lake is a vast storage system that houses extensive volumes of raw data in its original format. This versatile system accommodates a range of data types, including structured, semi-structured, and unstructured forms. Data lakes are essential in environments focused on big data analytics and are designed to manage data characterised by large volume, high velocity, and diverse variety from multiple sources. They are commonly utilised for advanced data processing activities such as machine learning and predictive analytics. Unlike traditional databases following the schema-on-write approach, data lakes follow the schema-on-read approach, providing flexibility in how data is formatted and used.

Dataverse. The concept of dataverse takes the notion of data lakes further by creating a networked space where data is stored, actively managed, and shared within the scientific community. A dataverse is a data repository platform for publishing, citing, and discovering datasets. It enables researchers to publish, cite, and discover datasets while providing metadata and tools to ensure others can understand and use data. Dataverses are often domain-specific and support the principles of open science, providing features such as data version control, digital object identifiers (DOIs) for citation, and tools for data analysis within the platform. They are community-driven and emphasize the accessibility and reusability of research data.

The most prominent example is the open-source Dataverse project developed by the Institute for Quantitative Social Science at Harvard University. The Dataverse

[5] https://www.qdasoftware.org
[6] kaggle.com
[7] https://datadryad.org/stash
[8] https://www.dataone.org/about/
[9] https://www.re3data.org
[10] https://jamesdixon.wordpress.com/2014/09/25/data-lakes-revisited/

Project, initiated by King [23], provides an open-source platform for sharing, preserving, citing, exploring, and analysing research data. It focuses on data citation and reproducibility, as discussed by Crosas [24], who highlighted the platform's role in fostering collaboration and open science.

Different academic institutions have built their dataverses for sharing and disseminating experimental scientific results, including the data collections they curate: University of Arizona[11], the Different universities and academic institutions have promoted their dataverses like the University of Hamburg[12], the University of Michigan[13] and the Grenoble Dataverse[14].

Summary. Together, these systems represent a shift toward more open, integrated, and efficient ecosystems for data management, offering novel solutions to the challenges posed by the vast amounts of data generated in modern research. They move away from traditional databases and toward more fluid, dynamic systems that can accommodate the ever-changing landscape of big data and scientific research.

A dataverse and a data lake are concepts related to data storage and management but serve different purposes and are designed with varying cases of use in mind. While a dataverse is a scholarly platform aimed at curating, sharing, and preserving research data with rich metadata and community collaboration features, a data lake is a more generalised and scalable storage solution for raw data to support diverse data analytics and processing workflows.

2.5. Data lakes and data verses in Life & Earth sciences

Dataverses in Life & Earth sciences are specialised digital infrastructures designed to address specific data management needs for these scientific domains. They provide a structured yet flexible environment where datasets can be stored, accessed, shared, and analysed. These dataverses typically offer robust metadata standards and tools to ensure their data are well-described, making them discoverable and usable for various research purposes.

In Life Sciences, dataverses often focus on genomics, proteomics, clinical trials, and other biological data, integrating various sources of information to aid in complex analyses like phenotype-genotype correlations. For Earth Sciences, dataverses might concentrate on geospatial data, climate models, seismic activity records, and ecological data, supporting efforts to understand and model the Earth's dynamic systems.

[11] https://arizona.figshare.com
[12] https://www.fdm.uni-hamburg.de/en/fdm.html
[13] https://www.icpsr.umich.edu/web/about/cms/2365
[14] https://scienceouverte.couperin.org/cellule-data-grenoble-alpes/

These repositories support open science by promoting data sharing across disciplinary boundaries. This feature enables researchers to replicate studies and build upon existing work, which is fundamental for advancing knowledge. They also facilitate interdisciplinary collaboration, allowing experts from different fields to contribute to and draw from a collective data pool. For instance, a dataverse in these fields might include a combination of high-throughput experimental data, field observations, and simulation outputs. The combination of openness and rigorous data management positions dataverses as critical resources in pursuing scientific discovery in Life & Earth sciences.

In life and earth sciences, data lakes are pivotal for consolidating scientific data collected from various biodiversity studies and geological events like earthquakes. Once curated, processed, and analysed, this data contributes significantly to data-driven experiments underpinned by well-established protocols. The harvested data enriches the data lake and supports the creation of detailed, curated views for dissemination through dataverses.

Our vision emphasises the importance of developing and maintaining data lakes with partially curated content in life and earth sciences, facilitating the continuous cycle of experimental data feeding back into the lake and subsequently sharing via dataverses.

3. Maintaining and sharing earth and life sciences knowledge: challenges

Various data on life and earth sciences have been acquired from different sources [25]. Integrated access to data collections and their curated versions can facilitate their maintenance, analysis and experimentation. It can also demonstrate knowledge of the discipline with its vocabulary, concepts and relationships in a synthetic way.

Curation, maintenance and exploration of data collections in the data lake calls for proposing techniques for exploring data collections that can be explored and enriched while producing new data and analytical results. Data curation also means keeping track of the type of experiments carried out on the data, their results and the conditions under which they were carried out. Maintaining a catalogue of data-related questions and experiments can promote open science, share data and knowledge, and share the data and knowledge the scientific community has gained from it [26]. This information should also be stored in the data lake.

Challenge 1: How to structure and organise life and earth sciences metadata? Metadata modelling is a way of structuring and organising earthquakes and biodi-

versity. The metadata model must make the content of a data lake findable, accessible, interoperable and reusable (FAIR principles [27]). Metadata can represent the data's structural, semantic and contextual aspects (provenance, conditions and assumptions under which the analytical results are obtained, i.e., the metadata driving the analysis). Most proposed models are based on logic or structured by graphs [28, 29] that can be specialised in seismic geophysical data and biodiversity. Besides, associating metadata can be achieved by considering quantitative and qualitative perspectives through data curation. Combining quantitative and qualitative approaches allows for a meta-model of the content used and produced in experiments and the conditions in which the content is produced, chosen, validated and considered representative knowledge for the domain of study.

Challenge 2: How to integrate data in the data lake? Since the experiments require several data collections, integrating the data into the data lake must be part of a pipeline that includes data discovery, exploration, selection and integration. This process should be designed based on the requirements of life and earth science experiments [25]. The heterogeneity of the data (text, signals, multimedia, proprietary formats from seismographs), the speed of the data often produced in the form of streams in the case of seismic sensors in addition to the volume are aspects that require original contributions in the design, maintenance and exploration of the data lake.

Challenge 3: How to integrate data in the data lake considering scientists' needs? The researcher's intervention, defined as a researcher-in-the-loop (RITL) [30], is a crucial aspect of human intervention to assess content concerning (i) the conditions in which it is produced and (ii) to make decisions about the new tasks to perform and the way a research project will move forward. RITL is a case of Human-in-the-loop (HITL), where the primary output of the process is a selection of the data, not a trained machine-learning model. HITL is crucial for handling supervision, exception control, optimisation, and maintenance [31, 32]. Under a RITL approach, a human sees all data points in the relevant selection at the end of the process. Using RITL requires a systematic solid way of working[15]. This characteristic is critical for designing content curation for quantitative and qualitative research methods.

Scientific content should be extracted and computed, including data, analytics tasks (manual and AI models), and associated metadata. This curated content allows the produced knowledge to be reusable and analytics results to be reproducible [33], thereby adhering to the FAIR principles [34].

[15] https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

4. Towards a curation approach for building a Life & Earth sciences data lake

Figure 1 illustrates the principle of our vision concerning the way a life and earth sciences data lake can be built, maintained and exploited. Our approach is based on the quantitative and qualitative curation of data harvested digitally and *in situ* (left-hand side of the figure). Heterogeneous raw data is gathered and stored in the data lake. Then, algorithms (statistical and Artificial Intelligence) and researchers can process, filter and classify data. This filtering process produces and stores meta-data in the data lake. Data exploration and integration (cleaning and engineering) processes can be performed on data samples from the data lake. They can be used for experimental purposes to produce content associated with the data stored in the data lake. Clean and curated data associated with meta-data representing the quantitative and qualitative perspective of the experiments can then be shared in a data verse (right-hand side of the figure).

Harvested data, models and knowledge integration. Various life and earth sciences data have been harvested from different sources. Since they are heterogeneous and produced at different paces (continuous and in batch), our approach proposes an integration approach based on a pivot meta-representation. The principle is to present a general meta-model of their content and process them for extracting technical, structural and semantic meta-data. This abstract representation provides integrated access to data collections and curated versions under a global knowledge graph and can promote their maintenance, analysis, and experimentation. It can also show the knowledge of the discipline with its vocabulary, concepts, and relations in a synthetic manner. The data lake can be pivotal in collecting, processing, and exporting raw data in a curated view.

Curation, maintenance, and exploration of data collections for bringing data value from in situ observations and experiments. Since data acts as a backbone in modelling phenomena for understanding their behaviour, it is critical to developing good collection and maintenance: which are available data collections? Are they complete? Which is their provenance? In which conditions were they collected? Have they been processed? In which cases have they been used, and what are the associated results? We propose techniques to explore data collections using graphs that can be explored and enriched while new data and analytics results are produced. Data curation also means keeping track of the type of experiments run on data, their results, and the conditions in which they were performed. Maintaining a

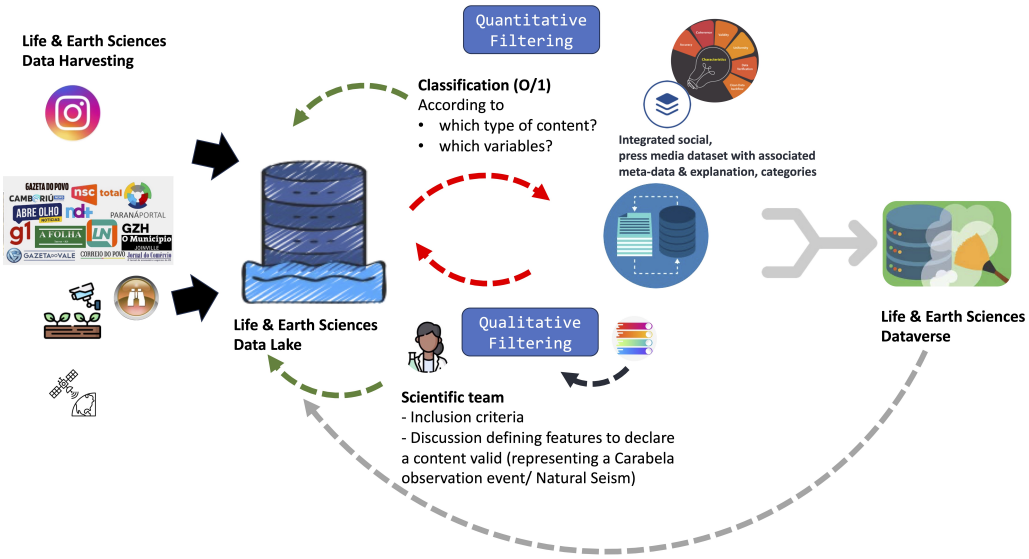

Figure 1: General overview of the curation approach for building, maintaining and exploiting a data lake.

catalogue of data-related questions and experiments can promote open science and share data and the knowledge that the scientific community has derived from it.

Modelling and simulating experiments to answer questions in life and earth sciences. Answering research questions through data-driven experiments implies:

- Designing ad hoc experiment artefact models and programming languages for enabling friendly, context-aware, and declarative construction of experiments in life and earth sciences.
- Collecting execution of experiments data (raw input data, prepared datasets, experiments' tasks calibration and associated results).

Pilot experiments. The data lake will be tested in real scenarios through collaboration with domain experts in seismology and biodiversity studies in Brazil. The entry point will be two pilot experiments, namely:

1. the classification process of seismic signals collected by stations through different observations to detect "natural" and human-made earthquakes in the northern human-made earthquakes in the northern region of Brazil;
2. the classification of *in situ* observations of the "carabela portuguesa"[16] and modelling its behaviour on the Brazilian coast.

In both cases, it is necessary to (i) apply statistical methods to investigate and unveil new patterns in seisms and biodiversity data, answering open problems or leading to new research questions; (ii) build predictive models to better describe or approximate phenomena, increasing the knowledge about our planet. The conditions in which statistics and prediction are performed, results, observations, interpretation and validation of the results are data to be integrated into the data lake.

Discussion. The originality of the work is to address the construction of a data lake that includes:

1. Raw collected data representing life and earth sciences phenomena (streams, batch, multimedia, proprietary).
2. Data produced along data-driven experiments adopting data science techniques including artificial intelligence algorithms (ML-driven data lakes).
3. Contextual data describing the conditions in which data are collected, and experiments are designed and enacted. The data lake will provide data curation modules for extracting metadata according to a well-adapted model and modules exploring data and using them for designing new experimentations, thereby adopting an open science perspective.

[16]The Portuguese caravel (Physalia physalis) is a monotypic colonial species of siphonophore hydrozoan of the family Physaliidae. It is commonly found in the open ocean in all warm waters of the world, especially in the tropical and subtropical regions of the Pacific and Indian Oceans, as well as in the Atlantic Gulf Stream. Its sting is dangerous and very painful https://es.wikipedia.org/wiki/Physalia_physalis.

5. Conclusions and future work

Our vision is that it is necessary to address fundamental research topics at the centre of Data Science, Big Data management and analytics for solving data-driven problems in life and earth sciences.

The contribution is the design and exploration techniques of a data lake with a well-adapted model for metadata about life and earth sciences experiments consuming and producing quantitative and qualitative data. An important work will be to define exploration operators and pipelines to exploit the content for further maintaining and developing new life and earth sciences experiments.

6. Acknowledgements

The work reported in this paper is done in the context of the LETITIA[17] project, funded by the *Fédération Informatique de Lyon*[18].

References

[1] G. Vargas-Solar, J.-L. Zechinelli-Martini, J. A. Espinosa-Oviedo, L. M. Vilches-Blázquez, Laclichev: Exploring the history of climate change in latin america within newspapers digital collections, in: New Trends in Database and Information Systems: ADBIS 2021 Short Papers, Doctoral Consortium and Workshops: DOING, SIMPDA, MADEISD, MegaData, CAoNS, Tartu, Estonia, August 24-26, 2021, Proceedings, Springer, 2021, pp. 121–132.

[2] L. S. do Nascimento, C. S. Hara, M. N. Junior, M. Noernberg, Redes sociais como uma fonte de dados alternativa para monitorar águas-vivas no brasil, in: Livro de Memórias do IV SUSTENTARE e VII WIPIS: Workshop internancional de Sustentabilidade, Indicadores e Gestão de Recursos Hídricos (Online) – Even3, Piracicaba, 2022.

[3] E. N. Sarr, S. Ousmane, A. Diallo, Factextract: automatic collection and aggregation of articles and journalistic factual claims from online newspaper, in: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2018, pp. 336–341.

[4] D. Garkov, C. Müller, M. Braun, D. Weiskopf, F. Schreiber, " research data curation in visualization: Position paper"(data) (2023).

[5] S. Lafia, A. Thomer, D. Bleckley, D. Akmon, L. Hemphill, Leveraging machine learning to detect data curation activities, in: 2021 IEEE 17th International Conference on eScience (eScience), IEEE, 2021, pp. 149–158.

[6] A. Zuiderwijk, R. Shinde, W. Jeng, What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption, PloS One 15 (2020).

[7] M. Vuorre, J. P. Curley, Curating research assets: A tutorial on the git version control system, Advances in Methods and Practices in Psychological Science 1 (2018) 219–236.

[8] M. Esteva, W. Xu, N. Simone, K. Nagpal, A. Gupta, M. Jah, Synchronic curation for assessing reuse and integration fitness of multiple data collections (2022).

[9] M. Woods, R. Macklin, G. K. Lewis, Researcher reflexivity: exploring the impacts of caqdas use, International Journal of Social Research Methodology 19 (2016) 385–403.

[10] ATLAS.ti, ATLAS.ti, https://atlasti.com, last accessed April 2023.

[11] Dedoose, Dedoose, https://www.dedoose.com/, last accessed April 2023.

[12] V. Software, Maxqda, http://maxqda.com, last accessed April 2023.

[13] NVivo, Nvivo, https://www.qsrinternational.com/, last accessed April 2023.

[14] N. Chen, M. Drouhard, R. Kocielnik, J. Suh, C. Aragon, Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity, ACM Transactions on Interactive Intelligent Systems 8 (2018) 1–20.

[15] J. C. Evers, Current issues in qualitative data analysis software (qdas): A user and developer perspective, The Qualitative Report 23 (2018) 61–73.

[16] S. Karcher, D. D. Kirilova, C. Pagé, N. Weber, How data curation enables epistemically responsible reuse of qualitative data, The Qualitative Report 26 (2021) 1996–2010.

[17] QDAMiner, Qdaminer, https://provalisresearch.com/products/qualitative-data-analysis-software/, last accessed April 2023.

[18] Quirkos, Quirkos, https://www.quirkos.com, last accessed April 2023.

[19] Transana, Transana, https://www.transana.com, last accessed April 2023.

[20] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, B. Mitschang, Leveraging the data lake: Current state and challenges, in: Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019, Linz, Austria, August 26–29, 2019, Proceedings 21, Springer, 2019, pp. 179–188.

[21] R. Hai, C. Quix, M. Jarke, Data lake concept and systems: a survey, arXiv preprint arXiv:2106.09592 (2021).

[22] P. Russom, Data warehouse modernization, TDWI Best Pract Rep (2016).

[17] http://vargas-solar.com/letitia/

[18] https://fil.cnrs.fr

[23] G. King, An introduction to the dataverse network as an infrastructure for data sharing, 2007.
[24] M. Crosas, G. King, J. Honaker, L. Sweeney, Automating open science for big data, The ANNALS of the American Academy of Political and Social Science 659 (2015) 260–273.
[25] U. S. da Costa, J. A. Espinosa-Oviedo, M. A. Musicante, G. Vargas-Solar, J.-L. Zechinelli-Martini, Using provenance in data analytics for seismology: Challenges and directions, in: European Conference on Advances in Databases and Information Systems, Springer, 2022, pp. 311–322.
[26] A. Adorjan, G. Vargas-Solar, R. Motz, Towards a human-in-the-loop curation: A qualitative perspective, in: 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2022, pp. 1–8.
[27] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.
[28] E. Scholly, P. N. Sawadogo, P. Liu, J. A. Espinosa-Oviedo, C. Favre, S. Loudcher, J. Darmont, C. Noûs, goldmedal: une nouvelle contribution à la modélisation générique des métadonnées des lacs de données, Revue des Nouvelles Technologies de l'Information (2021).
[29] A. Diouan, E. Ferey, S. Loudcher, J. Darmont, C. Noûs, Métadonnées des lacs de données et principes fair, in: 18e journées Business Intelligence et Big Data (EDA 2022), 2022.
[30] R. Van de Schoot, J. de Bruin, Researcher-in-the-loop for systematic reviewing of text databases, Zenodo: SciNLP: Natural Language Processing and Data Mining for Scientific Text (2020).
[31] I. Rahwan, Society-in-the-loop: programming the algorithmic social contract, Ethics and information technology 20 (2018) 5–14.
[32] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: A state of the art, Artificial Intelligence Review 56 (2023) 3005–3054.
[33] J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, J. Greenberg, The role of metadata in reproducible computational research, Patterns 2 (2021) 100322.
[34] P. P. F. Barcelos, T. P. Sales, M. Fumagalli, C. M. Fonseca, I. V. Sousa, E. Romanenko, J. Kritz, G. Guizzardi, A fair model catalog for ontology-driven conceptual modeling research, Conceptual Modeling. ER 73 (2022).