Experiversum: an Ecosystem for Curating and Enhancing Data-Driven Experimental Science

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

[1] CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221, France
genoveva.vargas-solar@cnrs.fr
[2] Federal University Rio Grande do Norte, DIMAP, Natal, Brazil
{umberto.costa,martin.musicante}@ufrn.br
[3] Université Lumière Lyon 2, ERIC, France
{jerome.darmont,sabine.loudcher}@univ-lyon2.fr
[4] Université Claude Bernard Lyon 1, ERIC, France
javier.espinosa@univ-lyon1.fr
[5] Federal University of Parana, Brazil
carmemhara@ufpr.br
[6] Fundación Universidad de las Américas Puebla, Mexico
joseluis.zechinelli@udlap.mx
[7] Universidad de las República, Uruguay
regina.motz@gmail.com

Abstract. This paper introduces Experiversum, a lakehouse-based ecosystem that supports the curation, documentation and reproducibility of exploratory experiments. Experiversum enables structured research through iterative data cycles, while capturing metadata and collaborative decisions. Demonstrated through case studies in Earth, Life and Political Sciences, Experiversum promotes transparent workflows and multi-perspective result interpretation. Experiversum bridges exploratory and reproducible research, encouraging accountable and robust data-driven practices across disciplines.

Keywords: Data and experiment curation · Reproducible research · Lakehouse architecture · Data processing pipelines · Metadata.

1 Introduction

Massive data production is increasingly vital in experimental sciences such as life, earth, social sciences and humanities, where large-scale, cost-effective data acquisition is now possible. Such fields generate diverse datasets of varying quality, enabling multifaceted analyses. Traditional schema-on-write methods such as ETL (Extraction, Transformation, Loading) struggle with such heterogeneity. Data lakes provide a flexible alternative by storing raw data in original formats, but require effective metadata extraction to integrate data and ensure reproducibility.

Open science demands not just data sharing, but also the documentation of experimental context, including conditions and decisions. This requires detailed metadata that captures both data and the knowledge production process. The main challenge is twofold: designing metadata models that represent both data and processing workflows, and implementing ELT (Extraction, Loading, Transformation) pipelines that support experiment curation and track how decisions impact outcomes. Metadata must serve as an execution guide for ELT processes to ensure reproducibility.

This paper introduces Experiversum, a lakehouse prototype system that applies a metamodel to curate and manage data-driven experiments. Experiversum enables researchers to explore, analyse and reuse experiments with rich metadata, in alignment with open science principles. Consequently, the remainder of the paper is structured as follows. Section 2 reviews related works on metadata, provenance and reproducibility. Section 3 introduces the Experiversum ecosystem. Section 4 details the system's architecture, curation processes and exploration functions. Section 5 presents use cases in social, earth, and life sciences. Section 6 concludes and outlines future work.

2 Related Works

This section reviews key approaches for curating experimental data and processes, covering storage and management systems such as data warehouses, data lakes, lakehouses and dataverses [15, 18]. We also compares data lake solutions used in earth, life, and social sciences.

The Evolving Practice of Data Curation. Data curation has evolved from focusing on preservation and quality control [10, 13] to a value-added process that includes metadata enrichment and contextualization [14]. In fields of earth sciences and biodiversity, this shift supports reusability and clear provenance [5]. Modern platforms such as dataverse combine automation with expert oversight to support the full research lifecycle [20].

Infrastructure for Modern Research. Managing today's research data, ranging from structured tables to unstructured content, requires flexible systems. Data warehouses are optimized for structured analytics, but lack support for diverse formats [3]. Data lakes address this with schema-on-read flexibility [9]. However, without proper governance, data lakes risk becoming "data swamps" [2]. The lakehouse model combines the strengths of both warehouses and lakes [1], while dataverses offer curated, citable storage [6, 12]. Our work advocates for integrating a lakehouse and a dataverse approaches in earth and life sciences.

Discipline-Specific Data Challenges. Different disciplines require tailored infrastructures. In natural sciences, repositories support metadata standards for reproducibility [16]. In social sciences, data is often qualitative and harder to standardize. Data lakes offer needed flexibility [8], but must preserve context and consider ethical issues, especially with personal or indigenous data [4]. Hybrid solutions aim to balance scale and detail [7, 19].

Innovations and Remaining Challenges. Emerging technologies such as conversational analytics using Large Language Models (LLMs) are reshaping interaction with data[8]. While promising, LLMs raise concerns about accuracy and trust[9]. Interoperability remains difficult across disciplines and data types [17]. Ultimately, success depends not only on technical solutions but also on institutional support and user adoption [11].

3 Curating Data-Driven Experiments

A data-driven experiment consists of three key elements: (1) raw data from empirical sources, (2) the research team responsible for data selection, methods, and validation, and (3) contextual metadata describing collection, processing, and analysis conditions. Curation ensures all components are documented for transparency and reproducibility. We define a metadata model structured around three concepts: raw content, experimental specifications and context. Figure 3 in Appendix A illustrates this model.

Level 1: Raw content. The blue classes in Figure 3 represent data ingested or produced during experiments. Metadata are extracted through automated and manual processes, capturing summaries, distributions and structure, e.g., column types and format. Each release is profiled, e.g., licensing, size and provenance; and can include tabular, textual or signal data. Items can also be annotated with multimedia or textual comments to enhance interpretability.

Level 2: Experimental specifications. This level documents actions performed on datasets, whether manual or automated. Actions produce artefacts or models, with metadata describing structure, execution and provenance. Parameters, evaluation criteria and validation protocols are recorded to trace how and why actions were performed or repeated.

Level 3: Experiment context. Metadata describe the research team's composition , e.g., roles and seniority; responsibilities and the guiding research question. It captures the decision-making context and provides a basis for comparing experiments.

4 Experiversum: experiments and data universum environment

The Experiversum environment ensures preservation, documentation and reproducibility of scientific experiments using a lakehouse infrastructure (Figure 1).

Extraction and Loading. Raw data such as seismic signals or social media posts are ingested in their original format, e.g., signals, text and media. Subsets are grouped into catalogues based on attributes such as ingestion date, size, format and quantitative traits.

[8] Nguyen 2024; Kerner 2023; Dubey 2024
[9] Ghodsi et al., 2023

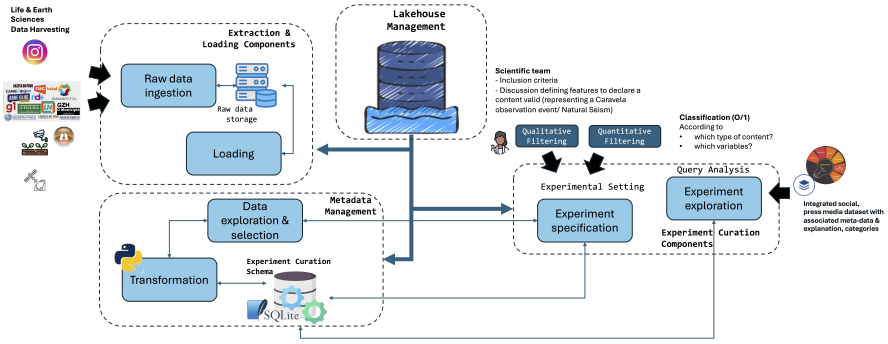

Fig. 1. Experiversum Architecture

Metadata Management. This module links metadata from raw and processed data to the experimental context, supporting reproducibility. Scientists can navigate datasets using quantitative summaries and relevant descriptors. Metadata is extracted following the curation model, normalized, and stored in a metadata repository, enabling comparison and exploration across experiments.

Experiment Curation. Researchers can specify experimental parameters such as selection criteria, team roles, questions and performance constraints. They can explore experiments on similar topics conducted under different conditions, review methods and assess outcomes—supporting comparative analysis and enhancing reproducibility.

Experiversum management. This component orchestrates Experiversum pipelines, managing seamless data flow from ingestion to analysis. It ensures consistency, performance, and smooth transitions across the infrastructure.

Extraction and Loading Pipeline. The EL pipeline handles data ingestion, cleaning and transformation before loading it for analysis—crucial for any data workflow. Key steps include (i) data extraction retrieves raw data from sources such as APIs, sensors, databases or unstructured files, e.g., seismic logs and social media; (ii) data cleaning removes duplicates, errors and inconsistencies; (iii) data enrichment adds contextual details such as timestamps or geolocations reliability; (iv) data loading stores data with associated structural and quantitative metadata in organized collections.

Tagging Experimental Processes Pipeline. improves reproducibility and collaboration by assigning structured tags to experimental workflows. It consists of three tasks. (i) experiment specification records metadata such as experiment ID, name and date to link processes and data; (ii tagging applies algorithmic or user-defined tags to annotate datasets and processes; (iii) tag storage maintains tag traceability for reuse and reference.

Transformation Pipeline converts raw or semi-structured data into usable formats aligned with the metadata model. It consists of three tasks. (i) structuring maps text, signals and media to metadata entities; (ii) contextual enrichment

adds metadata to reflect experimental settings; (iii) preparation formats data for analytics, machine learning or further experimentation.

Exploring and Querying Processes. The exploration and analytics pipeline enables users to query, analyze and visualize curated datasets. It supports exploratory data analysis (EDA), statistical modeling and machine learning to uncover insights. There are five tasks.

- Experiment Querying and Retrieval: access datasets by filtering parameters such as time, location or experiment settings for efficient data selection.
- Filtering and Aggregation: refine data by extracting relevant subsets and aggregating across dimensions, e.g., time and region) to produce summary metrics.
- Descriptive and Predictive Analytics: perform statistical analysis (averages, correlations, trends) and advanced tasks (classification, regression, clustering, anomaly detection...) for pattern discovery and forecasting.
- Data Visualisation: display results using graphs, charts and heatmaps to simplify interpretation, trend spotting and anomaly identification.
- Collaboration and Sharing: share results, export outputs and integrate findings into reports or publications to support teamwork and dissemination.

5 Use Case-Based Validation

The first prototype of Experiversum is implemented using SQLite3 as the storage backend. Pipelines are developed in Python and three demonstration scenarios (biodiversity, seismic data and graffiti analysis) are built using Flask, Bootstrap and executable Jupyter notebooks.

Tracking "Caravelas Portuguesas" along the Brazilian Coast. This use case classifies sightings of the jellyfish Physalia physalis along Brazil's coast[10].

Raw data extraction and loading. Instagram posts tagged with relevant hashtags (#aguaviva, #caravelaportuguesa, etc.) are extracted, converted into CSVs containing metadata (ID, source, location, media URL) and uploaded into Experiversum.

Data transformation. CSV headers are mapped to our metadata model, e.g., experiment, media, content and tags). Unstructured and imprecise geo-temporal data, e.g., "last summer" or inaccurate locations, is cleaned and corrected. Each transformation is registered and results in derived datasets.

Experimental settings. Two research teams collaborate: data scientists use machine learning models to classify posts, while biologists manually tag and define classification categories. Settings include inclusion criteria, e.g., location/time) gender of the affected person, model calibration and performance thresholds.

Exploration and querying. Users query jellyfish occurrences by time and region, explore ecological associations and compare human and machine learning classifications to study methodological differences.

Classification of Seismic Activity in Northeast Brazil. This case curates seismic data to differentiate natural from anthropogenic events, producing validated bulletins summarising seismic activity.

[10] https://es.wikipedia.org/wiki/Physalia_physalis

Experiment setup. Participants include seismographs (data generation), data collectors (retrieval), junior analysts (event detection) and senior analysts (review and bulletin publication).
Data extraction and loading. SAC files are uploaded and validated. Amplitude values (by axis: X, Y, Z) are extracted and stored along with metadata such as station_id, channel_id and timestamps.
Data transformation and tagging. Junior analysts plot waveforms to detect and tag events. Each event is annotated by station, year and magnitude. Analysts identify P and S wave arrivals. A triangulated event list forms the basis of the official bulletin, validated by a senior analyst.
Exploration. Waveforms and results are visualised through a Web interface. Analysts can share or publish curated outcomes.
Graffiti Analysis for Political Messaging. A two-member team (junior + senior) conduct qualitative analysis to classify political graffiti across a city.
Research framing. Over two cycles, the team refines the central question: "Can political messages be traced through graffiti?" They define inclusion criteria and political graffiti indicators through discussion.
Data collection. The junior researcher photographed 1,050 graffiti images across districts. After review, 546 were validated and shared on Instagram (link upon acceptance).
Analysis. Manual classification is complemented by unsupervised machine learning (k-means and hierarchical clustering via Orange). Results from both are iteratively refined and interpreted collaboratively.
Results. Narratives and metadata are compiled through successive review rounds. Final deliverables include classifications, summaries and reproducibility documentation.
Lessons Learned. Developing an experiment curation system reveals key insights into the challenges and benefits of structuring data-driven research. It underscores how data, metadata and decisions intersect, and the importance of systematic curation for transparency, reproducibility and collaboration.
Curation and Reproducibility. Experiversum supports the curation of varied data types (seismic signals, social media and multimedia) through ingestion, transformation and tagging pipelines. These pipelines enrich content with contextual metadata, enabling reproducible experiments and traceable results. *Lesson*: Metadata models are crucial for linking data with experiments, ensuring interpretability beyond storage.
Data Transformation and Tagging. While structured data such as seismic signals are easily processed, tagging unstructured content, e.g., social media, prove more difficult, requiring advanced techniques. *Lesson*: Automated tagging suits structured data. Unstructured sources need robust Natural Language Processing (NLP) methods.
Using Metadata to Understand Experiments. Figure 2 illustrates metadata-driven queries across three use cases. The first chart shows political graffiti labels by annotator, with juniors contributing most tags, indicating their key role in interpretation. The second chart compares human and machine classifications in seismic

monitoring, showing 90% agreement but highlighting some discrepancies needing expert review. The third chart displays confidence scores for species classification, while many fall in the 0.8–1.0 range, lower-confidence cases (<0.6) point to the need for manual checks. Such visualisations show how curated metadata improves analysis, validation and understanding across complex experiments.

6 Conclusion and Future Work

This paper introduces Experiversum, a lakehouse-based platform for curating, exploring and reproducing data-driven experiments. Experiversum integrates ELT pipelines with a structured metamodel to link raw data to experimental intent, enabling workflow reuse across disciplines. Case studies with biodiversity and seismic data highlight its flexibility for interdisciplinary research. The main insight is that reproducibility requires preserving full experimental context, not just raw data. Our metadata model and curated workflows improve traceability and reuse across diverse data types.

Future work includes extending metadata coverage, using NLP and graph techniques for tagging, adding privacy-aware analytics, and deploying the platform in real infrastructures to support open, collaborative science.

Acknowledgements. This work was funded by project LETITIA, Lyon Computer Science Federation (FIL). http://www.vargas-solar/letitia

References

1. Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M.: Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the 11th Conference on Innovative Data Systems Research (CIDR) (2021), https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
2. Becker, C., Genschel, U., Siegfried, T.: From data swamp to data lakehouse: Metadata management in interdisciplinary research. Journal of Information Management and Data Science **5**(2) (2022)
3. Bimonte, S., Coulibaly, F.A., Rizzi, S.: An approach to on-demand extension of multidimensional cubes in multi-model settings: Application to iot-based agro-ecology. Data & Knowledge Engineering **150**, 102267 (2024). https://doi.org/10.1016/j.datak.2023.102267, https://doi.org/10.1016/j.datak.2023.102267
4. Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., et al.: The care principles for indigenous data governance. Data Science Journal **19**, 43 (2020)
5. Cheney, J., Chiticariu, L., Tan, W.C., et al.: Provenance in databases: Why, how, and where. Foundations and Trends® in Databases **1**(4), 379–474 (2009)
6. Crosas, M., King, G., Honaker, J., Sweeney, L.: Automating open science for big data. The ANNALS of the American Academy of Political and Social Science **659**(1), 260–273 (2015)

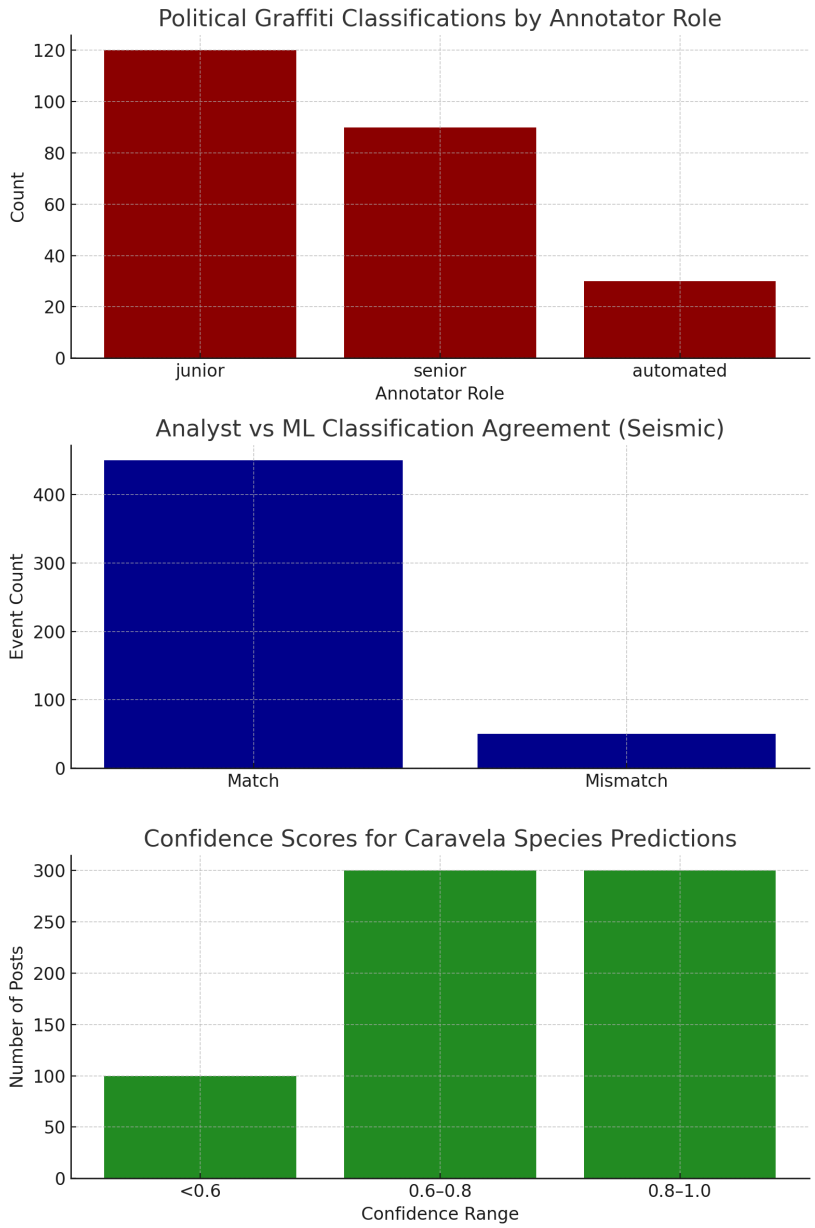

Fig. 2. Query visualisation in Experiversum

7. Dunning, A., van Erp, M., Skarpelis, C.: Advancing fair data practices in the social sciences: Lessons from the sshoc project. Data Science Journal **20**, 1–9 (2021)
8. Hai, R., Geisler, S., Quix, C.: Conquering the data lake: A research agenda. Proceedings of the 2016 International Conference on Information Systems (2016)
9. Hegde, M., Smit, C., Pilone, P., Petrenko, M., Pham, L.: Use of schema on read in earth science data archives. In: 2017 American Geophysical Union (AGU) Fall Meeting (2017), https://agu.confex.com/agu/fm17/meetingapp.cgi/Paper/311915
10. Higgins, S.: The dcc curation lifecycle model. International journal of digital curation **3**(1), 134–140 (2008)
11. Jagadish, H., Abiteboul, S., Buneman, P.e.a.: The big data conundrum in the social sciences. Communications of the ACM **64**(3), 68–77 (2021)
12. King, G.: An introduction to the dataverse network as an infrastructure for data sharing (2007)
13. Lord, P., Macdonald, A.: e-science curation report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision (2003)
14. Palmer, C.L., Renear, A.H., Cragin, M.H.: Purposeful curation: Research and education for a future with working data (2008)
15. Rocha, H.F., Nascimento, L., Camargo, L., Noernberg, M., Pozo, A.T.R., Hara, C.S.: Identifying occurrences of the cnidarian physalia physalis in social media data. Computer Science and Information Systems **21**(4), 1887–1911 (2024)
16. Russom, P.: Data warehouse modernization. TDWI Best Pract Rep (2016)
17. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. Journal of Intelligent Information Systems **56**(1), 97–120 (2021)
18. Vargas-Solar, G., Darmont, J., Adorjan, A., Espinosa-Oviedo, J.A., Hara, C., Loudcher, S., Motz, R., Musicante, M., Zechinelli-Martini, J.L.: Dataversifying natural sciences: Pioneering a data lake architecture for curated data-centric experiments in life & earth sciences. arXiv preprint arXiv:2403.20063 (2024), https://arxiv.org/abs/2403.20063
19. Vargas-Solar, G., Darmont, J., Adorjan, A., Espinosa-Oviedo, J., Hara, C.S., Loudcher, S., Motz, R., Musicante, M.A., Zechinelli-Martini, J.: Dataversifying earth sciences: Pioneering a data lake architecture for curated data-centric experiments in life and earth sciences. In: Palpanas, T., Jagadish, H.V. (eds.) Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference co-located with the EDBT/ICDT 2024 Joint Conference, Paestum, Italy, March 25, 2024. CEUR Workshop Proceedings, vol. 3651. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3651/DARLI-AP-13.pdf
20. Zgolli, A., Collet, C., Madera, C.: Metadata in data lake ecosystems. Data Lakes **2**, 57–96 (2020)

A Appendix A

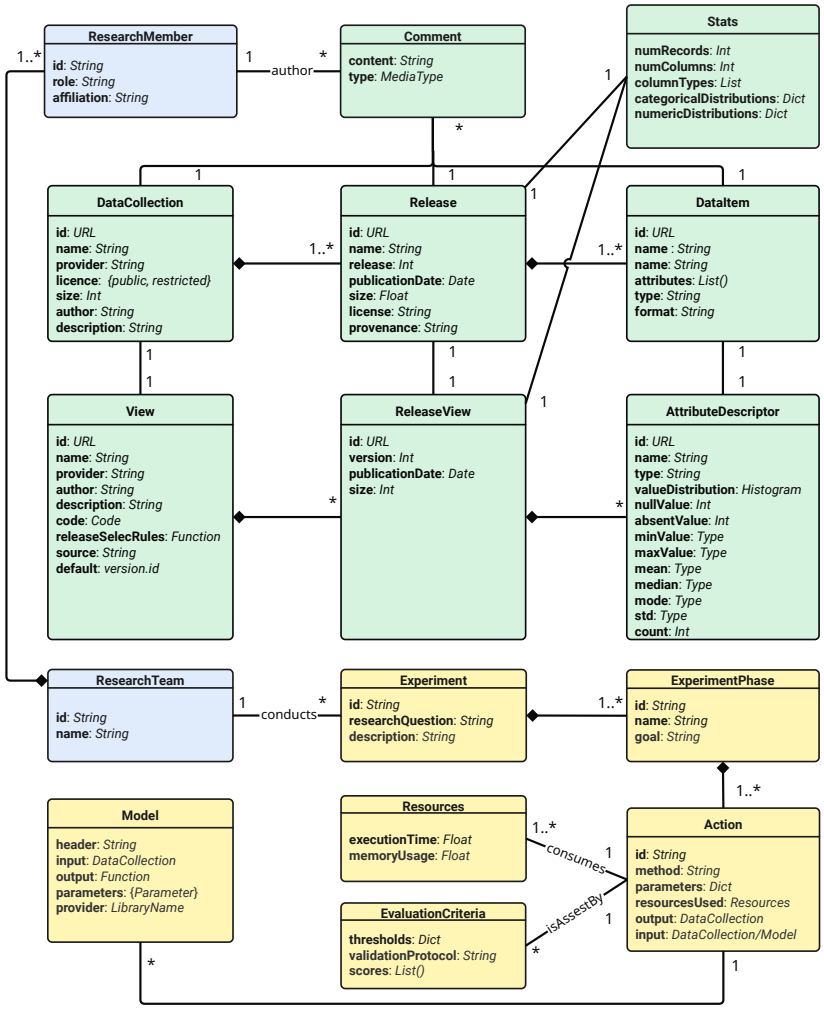

Fig. 3. Data Metamodel UML Class Diagram