EMPOWERING RESEARCH COMMUNITIES

Accelerating RDI Maturation in Switzerland

Carlos Vivar Rios
Swiss Data Science Center,
EPFL & ETH Zürich
Switzerland
carlos.vivarrios@epfl.ch
0000-0002-8076-2034

Oksana Riba Grognuz
Swiss Data Science Center,
EPFL & ETH Zürich
Switzerland
oksana.riba@epfl.ch
0000-0002-2961-2655

***Abstract* – Open Research Data (ORD) is gaining momentum in Switzerland, driven by the Swiss National ORD Strategy. To accelerate progress, the Swiss Data Science Center (SDSC) has established a dedicated ORD Engagement and Services team to support research communities in maturing their RDIs through a structured approach and targeted initiatives.**

We recognize that RDIs evolve through distinct levels: from isolated data, code, and models to standardized and findable digital research assets, then interoperable building blocks, culminating in an integrated ORD ecosystem. To facilitate this journey, we offer a flexible blueprint compatible with multiple domains and the latest good practices in data science. This blueprint serves as a practical starting point for communities to establish and evolve their RDIs.

To demonstrate of our RDI maturation framework, we introduce ORD hackathons focused on agile, cost-efficient and community-centric solutions. By bringing together RDI builders and users, these hackathons offer a platform for rapid collaborative prototyping and the evaluation of new systems.

This combined approach of a structured framework and hands-on hackathons empowers researchers to create FAIR, interoperable, and integrated RDI ecosystems.

***Keywords* – Research Data Infrastructure (RDI), Open Research Data (ORD), RDI Maturation, FAIR Principles, Community Engagement**

I. FROM RESEARCH ASSET TO AN OPEN RDI: A FLEXIBLE BLUEPRINT

A. Key Elements of an ORD RDI

The Swiss National Strategy defines Research Data Infrastructures (RDIs) as Research Infrastructures built around data or the capacity to work with data [1]. RDIs ideally foster community building and operate under Open Research Data (ORD) principles.

Drawing from this definition, we can identify essential components for any successful RDI:

- **Catalogues of Digital Assets:** Searchable and well-structured catalogues with API access to facilitate data discovery and integration.
- **Frictionless Access:** Easy access to data and software via downloads or remote execution.
- **Traceability:** Clear lineage of derived works, automatically integrated into the catalogues.
- **Community Governance:** Active participation and decision-making by the community.
- **Transparent Documentation:** Clear guidelines on operational management and usage.
- **User-Friendly Portal:** An intuitive interface for human interaction.
- **Coexistence:** Multiple RDIs can serve diverse needs and communities

B. Challenges and the Need for a Flexible Approach

The evolution of scattered datasets and software into a mature RDI is a complex process requiring active community participation. However, engaging researchers and fostering this evolution often requires significant resources and coordination. As a solution, we propose a flexible, low-resource blueprint to empower communities to easily initiate and operate RDIs, emphasising agile, low-cost and community-driven approaches. One effective strategy for achieving this is through hackathons that explore the evolution of isolated research assets into a mature and functional RDI.

This is a multi-stage process, as illustrated in Figure 1. Initially, research data and code often exist in isolation,

SRDD 2024: Swiss Research Data Day, Chur (Switzerland).

DOI: 10.nnnn/nnnnnnn.nnnnnnn

lacking standardisation or clear metadata ("Isolated Data & Code"). Through efforts in organisation, documentation, and the adoption of community standards, these assets become "Standardized and Findable," facilitating discovery and access. The next stage involves creating "Interoperable & Reusable Building Blocks" where data and software can seamlessly interact through standardised interfaces and protocols, enabling the development of workflows and integrated services. Ultimately, the goal is to achieve an "Integrated ORD Ecosystem," a collaborative environment where multiple interoperable RDIs exchange data and services, maximising the value of research assets and driving innovation.

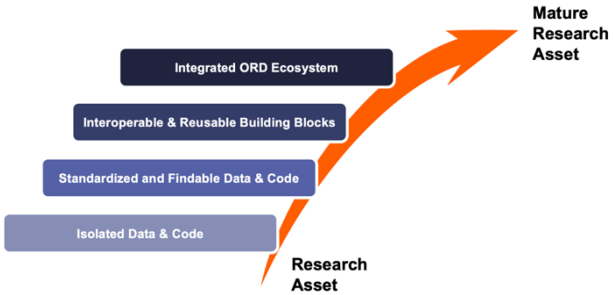

Figure 1. The maturation process of research asset.

II. Accelerating RDI Maturation Through Hackathons

Hackathons are intensive, time-limited events where individuals or teams come together to collaborate on creative problem-solving, innovation, or software development projects. Typically lasting between 24 to 72 hours, hackathons create a dynamic environment where participants—often programmers, designers, data scientists, and subject-matter experts—work collaboratively to build prototypes, develop solutions, or explore novel ideas. The event fosters experimentation and learning by providing resources such as datasets, APIs, and mentorship, and it often culminates in presentations or pitches of the developed solutions to judges or stakeholders.

Hackathons provide a focused and collaborative environment where researchers, developers, and other stakeholders can prototype and test new ideas rapidly. By bringing together RDI builders and users, hackathons can foster community ownership, accelerate the development cycle, and ensure that RDIs are truly responsive to the needs of their users.

Hackathons are not limited to coding; they are increasingly used across diverse fields, including healthcare, climate science, and education. These events encourage participants to focus on specific challenges, such as addressing social issues, advancing sustainability, or leveraging new technologies. By emphasising creativity, collaboration, and rapid iteration, hackathons act as catalysts for innovation, bringing together interdisciplinary expertise and energy to solve pressing problems or explore futuristic concepts.

We can divide the output of a hackathon into two main categories: functional demonstration and conceptual designs. Functional demonstrations showcase one or several RDI features, allowing participants to explore new technologies or procedures without the constraints of a production environment. This means they can focus on functionality rather than production-specific details like security and scalability. Conceptual designs, on the other hand, prioritize the high-level vision, often making assumptions about technical feasibility. These projects excel at identifying novel ways to engage the community or connect existing initiatives. In practice, most projects blend elements of both categories to varying degrees, depending on the team's focus and interests.

A. Prototyping an RDI with GitHub

To reduce the resource requirements for initial RDI prototyping, we propose a blueprint for RDI implementation based on GitHub [2]. This platform serves as a space to share connectors between different research assets, contribute to the catalogue of related assets, host documentation and applications, and, finally, offer an environment where the community can collaborate and redefine their rules of interaction and evolution.

GitHub has been successfully used for software development for decades, but its use can be extended to any digital project. In a GitHub project, collaborators can perform modifications (Commits) and propose these changes to debate with the rest of the community in pull requests (PRs). If the PR gets approved, it gets merged into the official project, and some automatism can be triggered, i.e., the update of a webpage or reanalysis of datasets.

Our prototype proposal leverages GitHub repositories to create an RDI hub, fostering community-driven development through committed changes and public discussions. These modifications can be categorised into:

- **Constitutional:** Changes related to RDI governance.
- **Curations:** Additions, deletions, or modifications to assets.
- **Content:** Updates to the portal's content or associated applications.

A central Hub repository hosts the catalogues, apps, static pages, and discussions, serving as the core of the RDI (Figure 2). This approach provides a practical starting point for communities to build their RDIs, promoting collaboration, transparency, and open data principles. By prioritising community engagement and leveraging existing tools, we aim to facilitate the development of sustainable and impactful RDIs in Switzerland and beyond.

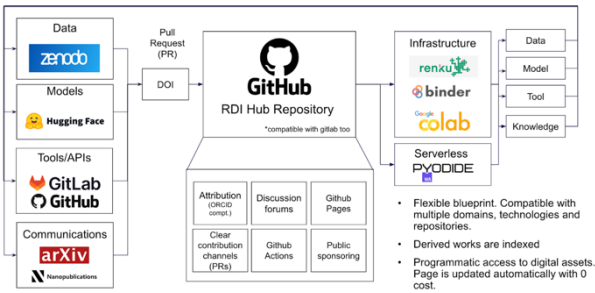

Figure 2. Diagram showing how a GitHub repository can be used to tailor research assets and infrastructure.

Beyond codebases and catalogues, GitHub offers powerful capabilities beyond serving as a repository for codebases or catalogues. Its integration with GitHub Pages enables automated deployment of web content, such as documentation or dashboards, triggered directly by the approval of pull requests. For more advanced use cases, GitHub Pages can host lightweight applications running in the browser via WebAssembly or even provide SQL-ready endpoints powered by SQLite. This transforms GitHub into a versatile platform that not only stores code and data but also delivers functionality, significantly reducing user friction.

The emerging technologies are rapidly evolving, with notable use cases already demonstrating their potential. For example, GPU models can now be deployed directly in a user's browser, enabling on-device inference over data that never leaves the user's computer. This approach enhances both performance and privacy, showcasing the potential for GitHub to become a dynamic, multi-functional space that empowers developers and researchers alike.

B. Community-Driven Projects

A core principle of our approach is to foster community-driven projects. This means empowering researchers to utilise existing RDIs and actively participate in their development and evolution. To achieve this, clear community flows are essential. These flows, defined through agreements and policies, guide individual contributions and ensure a cohesive, sustainable collaborative research ecosystem. Specific contribution guidelines and automated processes via GitHub Actions facilitate this process.

These flows should establish clear criteria for cataloguing research assets, specifying acceptable resource types, quality standards, and scientific value. Each accepted asset should be accompanied by a data model (e.g., a model card in YAML format for machine learning models) that the community can re-evaluate as needed.)

A GitHub-based workflow anchors this model. Issue tracking and pull requests foster transparency, constructive feedback, and inclusive decision-making. Public and focused issues invite collective problem-solving, while pull requests are structured venues for peer review and knowledge exchange.

This framework allows the community to adapt and evolve over time. Through regular retrospectives, open discussions, and flexible governance mechanisms, the RDI remains responsive to emerging scientific challenges and opportunities. This continuous evolution paves the way for a more mature and impactful RDI.

This community driven approach exemplified by successful open-source communities, such as the Automatic 1111 platform [3] for generative AI and the Napari image research platform [4], fosters robust and evolving ecosystems. Transparent contribution guidelines and inclusive processes are key to their success. Instead of traditional top-down infrastructure management, this approach fosters dynamic contribution networks built on trust, clear communication, and recognition of diverse contributions, ultimately driving innovation. Adopting this model allows us to cultivate adaptive, collaborative knowledge ecosystems equipped to address the complexities of modern scientific research.

III. ORD For The Sciences: A Practical Case Analysis

The "ORD for the Sciences" hackathon, jointly organised by the Swiss Data Science Center (SDSC) and EPFL Open Science, offered a compelling example of how hackathons can contribute to RDI development. This two-day event brought together over 80 participants from academia, the private sector, and public institutions to explore the role of ORD in advancing research. The hackathon featured three main tracks:

a) Research Data Infrastructure (RDI) track: Promote the development of RDI implementations by research communities around open science digital assets,

b) Data Science track: Encourage the development of digital resources for RDIs and

c) Outreach track: Explore new ways of communicating research results and engaging communities from RDIs.

Prior to the hackathon, participants submitted project proposals aligned with these tracks. The organising team reviewed and refined these proposals to ensure feasibility within the two-day timeframe. Each project description included goals, research resources, and required skills.

Participants were encouraged to adopt an open-source license (MIT by default) and adhere to FAIR (Findable, Accessible, Interoperable, Reusable) principles [5], including obtaining persistent identifiers, such as Digital Object Identifier (DOI) and ensuring open, well-documented access to data and code.

Prior to the hackathon, the organising committee facilitated team formation based on participant preferences, skill complementarity, and a balance of expertise levels. This resulted in 80 participants forming 14 teams, with seven teams focusing on the RDI track. Participants then began coordinating online through a dedicated Discord channel. To further support participants, the organising committee provided two webinars and ongoing technical support,

allowing them to focus on brainstorming and project development. These teams then convened at EPFL for two days of intensive collaboration (see table below).

To support and document the hackathon process, the SDSC developed the sdsc-hackathons.ch platform, which comprehensively supported all event stages. This dedicated site served multiple purposes: tracking participant progress, showcasing project developments, and providing a centralised resource for documentation and community engagement.

A. Projects Developed During the Hackathon

The hackathon showcased diverse scientific resources that exemplified the multifaceted nature of Research Data Infrastructures (RDIs) across different disciplines. These projects demonstrated the critical importance of Open Research Data by representing a broad spectrum of data characteristics, highlighting the unique challenges and opportunities in data management.

- **SzCORE - Epilepsy Benchmarks [6]:** This project highlighted the impact of open data on healthcare by developing a platform for benchmarking epileptic seizure detection algorithms.
- **Brain Score [7]:** This platform aims to yield accurate, machine-executable computational models of how the brain gives rise to the mind.
- **ERA5 [8]:** a comprehensive reanalysis dataset from ECMWF, provides global atmospheric, land, and ocean variables with a resolution of approximately 31 km and hourly data spanning several decades, resulting in a multi-petabyte-scale archive.
- **The Happy Whale and Dolphin Project [9]:** is a machine learning competition aimed at developing models to identify individual whales and dolphins from images, leveraging advanced computer vision techniques to aid marine life research and conservation by improving the tracking and monitoring of these species globally.
- **DemocraSci [10]:** This initiative fosters transparency and accountability in democratic processes by leveraging data science techniques, such as natural language processing and machine learning, to analyse and visualise political discourse, public opinions, and decision-making processes, enabling evidence-based insights into democratic systems.

Using these already established scientific resources spanning diverse domains—from healthcare and neuroscience to climate research, marine biology, and political science—seven teams leveraged the hackathon's collaborative framework to develop innovative Research Data Infrastructures (RDIs).

Each team creatively approached their chosen domain, transforming existing datasets into more accessible, interactive, and powerful research platforms. Among these, Team AWORD stood out with their GitLab-based data indexing platform, integrating metadata directly into repositories. This approach ensured version control and fostered community collaboration through familiar tools, embodying principles of transparency and accessibility.

Another notable project leveraged SzCore, a platform to benchmark epileptic seizure detection algorithms. By hosting sensitive datasets in a secure infrastructure and enabling portable containerised execution of algorithms, the team created an open-source, collaborative ecosystem that emphasised traceability and fairness in evaluating models running on sensitive data. This project demonstrated how a concrete, focused effort, connected with the right infrastructure can have a big impact on healthcare innovation.

The Brain-Score visualisation project brought a unique perspective to neuroscience research. By leveraging Python scripts and interactive D3.js-based visualisations, the project provided an intuitive interface for exploring the connection between AI models and brain function. Its modular architecture underscored the potential for interoperability, a cornerstone of a successful RDI.

Processing large-scale multidimensional data, such as ERA5, presented another critical challenge addressed during the hackathon. One team proposed combining Zarr and TrinoDB to efficiently query and process such datasets. This method showcased the need for tailored approaches for RDIs in handling high-dimensional data.

With a similar problem but another approach CoralORD tackled the challenges of managing ERA5 datasets by creating a solution for remote access and analysis of large-scale, software-defined datasets locally. By integrating DuckDB and Croissant [11], the project demonstrated a scalable approach to data querying, offering insights into how RDIs could support big data challenges independently of the discipline.

The team behind Saving Willy demonstrated how RDIs can bridge domain-specific needs and broad accessibility. By cataloguing cetacean images with metadata and offering ML tools through a Streamlit interface hosting in Hugging Face Spaces, the project emphasized traceability and community engagement. Their solution brought to the platform a model developed for a challenge, enhancing its reusability

Finally, in the political studies domain, the Neo4J graph navigation platform introduced a novel way to explore legislative data spanning multiple Swiss parliamentary periods by using large language models (LLMs). This demonstrated how a dataset shared in Zenodo can be deployed and adapted to other interfaces to allow its reuse by non-technical communities.

These projects highlighted the RDI model's versatility, demonstrating how it can be adapted to different domains while maintaining core principles like FAIR data, community governance, and transparency. Furthermore,

they underscored the power of hackathons to catalyse innovative, community-driven solutions that advance open research.

Table 1. Analysis of hackathon projects submitted to the RDI track at the ORD for the Sciences Hackathon.

Project	Catalogues of Digital Assets	Frictionless Access	Traceability	Community Governance	Transparent Documentation	User-Friendly Portal	Coexistence
Team 1. AWORD [12]	Provides a data index based on GitLab. Metadata is located within the same repository, potentially limiting scalability.	Uses Decap for entry edition and GitLab authentication.	Though use of GitLab offers version control capabilities.	Encourages community collaboration through GitLab.	Provides detailed documentation, including a glossary, hosted with Quarto Markdown.	Uses Quarto Markdown for the frontend, enabling simple and readable interfaces.	Focused on FAIR data principles but does not explicitly mention coexistence with other RDIs.
Team 2. RDI to benchmark epileptic seizure detection algorithms [13]	Hosts datasets of people with epilepsy to enable benchmarking.	Allows developers to upload seizure detection algorithms, leveraging GitHub and Docker for easy access.	Builds a leaderboard and benchmarking framework to maintain algorithm evaluation traceability.	Emphasizes open-source principles and encourages collaboration via GitHub discussions and issues.	Documentation is available via GitHub repositories, but not extensively detailed.	Not explicitly mentioned but leverages GitHub pages for visualization.	Can potentially integrate with broader healthcare datasets and benchmarks.
Team 3. Discovering AI and brain connections via interactive visualizations on Brain-Score [14]	Brain-Score hosts hundreds of models and datasets, acting as a central catalog for the neuroscience community.	Provides access to models and benchmarks via an interactive interface.	Uses Python scripts and statistical metrics for traceability in benchmark performance.	Encourages community-driven development and comparison of AI-brain models.	Offers clear and concise documentation, making data and tool usage straightforward.	User-friendly portal with interactive D3.js-based visualizations, improving accessibility.	Not explicitly mentioned, but its modular nature supports potential integration with other platforms.
Team 9. Processing Large Scale Multidimensional S3 Data on Distributed SQL Engines [15]	Proposes methods to query highly dimensional data stored in Zarr or similar formats, supporting catalog-like functionality.	Aims to enhance accessibility and usability of high-dimensional datasets via distributed SQL engines (e.g., TrinoDB).	The use of Zarr and TrinoDB supports traceable queries.	Promotes open-source methodologies and encourages future community-driven improvements.	Documentation could benefit from additional clarity but uses community-based feedback.	Not explicitly mentioned.	Potentially coexists with other RDIs by enhancing cloud-native data formats and processing.
Team 10. Saving Willy (the Orca) with Data! [16]	Provides a centralized hub for cataloguing and sharing cetacean images and metadata.	Researchers can upload and access images, alongside using ML models via a Streamlit interface.	Includes dataset DOIs for traceability.	Actively encourages community engagement through guidelines and Huggingface integration.	Offers some documentation and templates for model integration on Huggingface.	Provides a user-friendly interface in Streamlit, making it accessible to researchers.	Can coexist with other biodiversity or marine datasets as a domain-specific tool.

Team 12. A platform for semantic navigation and visualization of Neo4J graphs [17]	Its provide access to the knowledge graph with information extracted for four consecutive legislative Swiss periods (48 to 51), from year 2007 to 2023.	It aimed to produce an LLM interface to the knowledge graph and an easy way to run the application using python environment.	The dataset was hosted in Zenodo and have a DOI.	Open to contributions from the community via github issues.	The documentation offered can be improved but it contains basic running instructions	They offered a streamlit based interface that it's able to connect to the Neo4j instance.	It coexists with other RDIs on political studies and acts as a tool to improve accessibility to the data.
Team 13. CoralORD: A replicated, self-updating software-defined dataset for machine learning applications [18]	This project explored the organization of ERA5 datasets and derived works.	It generated a solution to handle large software-defined datasets with remote access capabilities. By using DuckDB and Zarr it allows to perform analysis without the need of downloading the whole dataset	CoralORD uses Croissant to define dataset metadata and subset descriptors for structured and scalable data management.	It supports asynchronous collaboration by using Git and SDDM to enable real-time dataset creation, download, and version control.	The documentation can be improved as there's only a README file.	It makes use of DuckDB in order to interact with the database. However, a proper user interface is missed.	This project offers a tool that can be used in other RDIs that needs to deal with big datasets.

IV. Analysis of Projects Based on the GitHub RDI Hub Approach

To assess the effectiveness of our GitHub-centric RDI approach, we analysed the hackathon projects, focusing on the following key aspects:

- **RDI Maturation Stage:** The hackathon projects used research assets at various stages of maturity, ranging from "Isolated Data & Code" to "Interoperable & Reusable Building Blocks." For example, the "Discovering AI and brain connections via interactive visualisations on Brain-Score" project enhanced an existing platform with new features, showcasing a step towards "Interoperable & Reusable Building Blocks." This highlights the potential of hackathons to support RDI development across different maturity levels (Figure 1).
- **Alignment with Core Components:** Several projects effectively utilised the core components of our approach. For instance, the "RDI to benchmark epileptic seizure detection algorithms" project employed GitHub to host datasets, code, and documentation, creating a centralised and accessible hub for the research community.
- **Flexibility and Adaptability:** While some projects adhered closely to our proposed approach, others showcased its adaptability. For example, the "Saving Willy (the Orca) with Data!" project integrated GitHub with Hugging Face and Streamlit to address the specific needs of cetacean conservation. This demonstrates the versatility of our approach in accommodating diverse research domains and technologies.
- **Community Engagement and Collaboration:** The hackathon fostered a strong sense of community, with many projects involving participants from different institutions and disciplines collaborating on innovative solutions. This collaborative spirit is a key enabler of successful RDI development, as it ensures that RDIs are responsive to the needs of their users and promote knowledge sharing.
- **Challenges and Opportunities:** The hackathon also highlighted some challenges and opportunities related to our proposed approach. Some projects faced difficulties in ensuring data privacy and security, particularly when dealing with sensitive data. This emphasises the need for continued development of tools and best practices for managing sensitive data within RDIs.

This analysis demonstrates the effectiveness of our GitHub-centric approach in fostering community-driven RDI development. The hackathon projects showcased a range of innovative solutions, highlighting the flexibility and adaptability of the blueprint. The event also underscored the importance of community engagement and collaboration in building sustainable and impactful RDIs.

V. Conclusion

The "ORD for the Sciences" hackathon highlighted the value of our GitHub-based blueprint for rapid initiation and development of Research Data Infrastructures (RDIs). Participants effectively utilized features such as version control, issue tracking, and automation workflows to create FAIR (Findable, Accessible, Interoperable, and Reusable) digital assets. This approach proved particularly effective for fostering community-driven development, enabling teams to establish traceable, transparent, and adaptable RDIs while addressing challenges such as metadata integration, data accessibility, and large-scale data processing.

Moreover, the hackathon served as an ideal environment for applying and testing our broader RDI maturation framework. Its collaborative format encouraged rapid

prototyping, interdisciplinary exchange, and creative problem-solving, leading to innovative projects across diverse domains like neuroscience, healthcare, and environmental science. These outcomes reinforce the value of hackathons as a key strategy for accelerating RDI maturation and fostering a culture of collaboration within research communities, ultimately driving impactful and sustainable advancements in open science.

In conclusion, the "ORD for the Sciences" hackathon provided valuable insights into the practical application of our low-resource GitHub-based blueprint and its role within our broader RDI maturation framework. By combining structured guidance with hands-on experimentation, we can empower researchers to create FAIR, interoperable, and integrated RDI ecosystems that drive scientific progress.

REFERENCES

[1] ORD National Strategy Council (StraCo), *Research Data Infrastructures: A Distinct Characteristic in Research Infrastructures*, Version of October 31, 2023. [Online]. Available: https://openresearchdata.swiss/wp-content/uploads/2024/08/Concept-Paper-StraCo_V4_2023-10-23.pdf

[2] SDSC, "ORD for the Sciences Hackathon: pNeuma RDI Hub," GitHub repository, 2024. [Online]. Available: https://github.com/sdsc-ordes/ordfts-hackathon-pneuma-rdi-hub

[3] AUTOMATIC1111, "stable-diffusion-webui-extensions," GitHub repository, 2024. [Online]. Available: https://github.com/AUTOMATIC1111/stable-diffusion-webui-extensions

[4] N. Sofroniew *et al.*, "napari: a multi-dimensional image viewer for Python," [Computer software]. Available: https://github.com/napari/napari.

[5] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, article 160018, 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.18

[6] J. Dan *et al.*, "SzCORE: A Seizure Community Open-source Research Evaluation framework for the validation of EEG-based automated seizure detection algorithms," *Epilepsia*, vol. 65, no. 3, pp. 456–467, Mar. 2024. [Online]. Available: https://doi.org/10.1111/epi.18113

[7] M. Schrimpf *et al.*, "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?," *bioRxiv*, 2018. [Online]. Available: https://doi.org/10.1101/407007

[8] Copernicus Climate Change Service (C3S), "ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate," 2017. [Online]. Available: https://cds.climate.copernicus.eu/cdsapp#!/home

[9] T. Cheeseman, K. Southerland, W. Reade, and A. Howard, "Happywhale - Whale and Dolphin Identification," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/competitions/happy-whale-and-dolphin/.

[10] L. Brandenberger, 'DemocraSci - A Parliamentary Knowledge Graph (4 legislative periods)'. Zenodo, Oct. 15, 2024. doi: 10.5281/zenodo.13920293.

[11] M. Akhtar *et al.*, "Croissant: A Metadata Format for ML-Ready Datasets," in *Proceedings of the 2024 ACM SIGMOD International Conference on Management of Data (DEEM '24)*, Santiago, Chile, 2024, pp. 1–6. doi: 10.1145/3650203.3663326.

[12] ORD Hackathon 2024, "01_Matterhorn," GitLab repository, Oct. 2024. [Online]. Available: https://gitlab.com/ord-hackathon-2024/01_Matterhorn.

[13] E. Mazhar, D. J. L. Lobo, E. S. Wurbel, and C. Doret, "SzCORE: An Open Seizure Detection Benchmarking Platform," GitHub repository, 2024. [Online]. Available: https://github.com/esl-epfl/szcore.

[14] M. Schrimpf *et al.*, "brain-score.web," GitHub repository, 2019. [Online]. Available: https://github.com/brain-score/brain-score.web

[15] E. Boulle et al., "Processing Large Scale Multidimensional S3 Data on Distributed SQL Engines," GitHub repository, Oct. 2024. [Online]. Available: https://github.com/ml-ops-edu/hackhaton-10-2024-team9

[16] L. Vancauwenberghe *et al.*, "Saving Willy Project," Hugging Face, 2024. [Online]. Available: https://huggingface.co/Saving-Willy/cetacean-classifier

[17] R. Franken, et all, "De-Cypher: Natural language questions to Cypher (Neo4J) query language generator," GitHub repository, Oct. 2024. [Online]. Available: https://github.com/sdsc-ordes/hackathon-nordend

[18] J. Anderson et al, *CoralORD: Self-updating software-defined datasets for reproducible ML applications*. [Online]. Available: https://github.com/jagh/CoralORD

ACKNOWLEDGMENTS

We want to express our sincere gratitude to all the participants who contributed their time, skills, and creativity to the hackathon, making it a vibrant and impactful event.

Special thanks go to our hackathon co-organizers: Luis Salamanca (SDSC), Valerio Rosetti (SDSC), Noémie Mazaré (EPFL), and Aruni Senaratne (EPFL). Your dedication and teamwork were instrumental in bringing this event to life.

We also acknowledge the collaborative efforts of the Swiss Data Science Center (SDSC) and the EPFL Open Science Office, whose partnership was essential in shaping and executing the hackathon.

Finally, we thank the large language model GPT-4 for assistance with refining and editing the manuscript.

CONTRIBUTIONS

Carlos Vivar Rios: Conceptualization, investigation, and writing.

Oksana Riba Grognuz: Project administration, supervision, and writing.