PROCEEDINGS OF SCIENCE

Enhanced Data Management Tools for ILDG and PUNCH4NFDI

Giovanni Pederiva,[a,b,*] **Basavaraja Bheemalingappa Sagar,**[c] **Daniel Knüttel**[d] **and Christian Schmidt**[c]

[a] *Jülich Supercomputing Center, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 54245 Jülich, Germany*

[b] *Center for Advanced Simulation and Analytics (CASA), Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany*

[c] *Universität Bielefeld, Fakultät für Physik, Universitätsstrasse 4, 33615 Bielefeld, Germany*

[d] *Fakultät für Physik, Universität Regensburg, Universitätsstrasse 31, 93040 Regensburg, Germany*

E-mail: g.pederiva@fz-juelich.de

We present the status and plans to develop as part of the PUNCH4NFDI project the tooling for a user-friendly and modern interface for the International Lattice Data Grid (ILDG).

The 41st International Symposium on Lattice Field Theory (LATTICE2024)
28 July - 3 August 2025
Liverpool, UK

*Speaker

PoS(LATTICE2024)450

1. Introduction

Research Data Management (RDM) and data sharing are crucial in the current landscape of scientific efforts, including for Lattice Quantum Chromodynamics (LQCD) ensembles of gauge field configurations. The International Lattice Data Grid (ILDG) [1] initiative aims at the federation of services and infrastructure to support storage and sharing of lattice data according to the FAIR principles (Findable, Accessible, Interoperable, and Reusable) to ensure data usability and longevity. New developments in the context of the PUNCH4NFDI project enabled the modernization of the ILDG services and architecture, allowing researchers to collaborate efficiently through standardized tools and workflows for data management.

2. PUNCH4NFDI

The PUNCH4NFDI consortium, funded by the German National Research Data Infrastructure (NFDI) program, supports around 9000 scientists across fields such as particle physics, astro-particle physics, hadron physics, and nuclear physics. RDM forms a core component of the consortium's work, addressing high data volumes through distributed infrastructures and task-specific tools [2].

2.1 Architecture and Community Impact

PUNCH4NFDI's federated data infrastructure is built to align with FAIR principles, emphasizing interoperability. Its architectural design draws from existing frameworks, like the AARC Blueprint Architecture [3], which provides guidance on scalable and modular data solutions. The ILDG middleware and services which are developed as part of the PUNCH4NFDI project integrate secure access controls, metadata management, and long-term data archiving, creating an ecosystem that not only supports lattice QCD but also serves as a model for multidisciplinary research.

Similar requirements and strategies to store and extract information from experimental data, simulations and observations can be leveraged to design and realize distributed data infrastructures and services. One of the goals of PUNCH is to develop middleware and services for a federated infrastructure for storage and long-term archiving of (meta-) data following the FAIR principles.

By pooling expertise across Germany's physics communities, PUNCH4NFDI helps standardize data practices, ensuring that datasets are documented, curated, and shared in ways that facilitate reuse. The consortium aligns itself well with NFDI's nationwide goals, serving as a blueprint for effective RDM across scientific fields.

3. International Lattice Data Grid (ILDG)

ILDG was established to facilitate data sharing and interoperability in lattice QCD research, connecting regional grids across the world and ensuring consistent data access for LQCD datasets. In particular, the rich and extensible metadata schemata for gauge configurations and ensembles, and the standardized file format for gauge configurations ensure that data is accessible, easily locatable, and documented, reducing redundancies across research projects. The ILDG's goals reflect the FAIR principles and is aligned with the global push toward improved RDM practices, underscoring the importance of accessibility and reuse in scientific research [4].

3.1 ILDG-Wide Services and Standards

ILDG is a federation of several autonomous regional grids forming a single Virtual Organization (VO). Each regional grid is responsible for operating its services and managing its datasets according to the common specifications and standards. The ILDG board guides decisions on common policies and strategies, and oversees the two working groups: The Metadata Working Group (MDWG) for the development of e.g. the community-wide metadata schemata and format specifications, and the Middleware Working Group (MWWG) to define requirements and specifications for the interoperability of the services and infrastructure across regions. In the recent years, a major effort has begun to improve and modernize ILDG, transitioning to the so-called ILDG 2.0 [5].

In 2024, ILDG transitioned to token-based authentication via the Indigo IAM (Identity and Access Management) service [6], which supports OpenID Connect (OIDC) and OAuth2 protocols, shifting away from grid certificates. This approach improves access control, simplifying user authentication, reflecting the ILDG's commitment to adapting technologies that enhance user experience, security, and scalability.

3.2 Regional Grid Services

Managing large volumes of data across distributed networks requires comprehensive services that support metadata management, storage, and access control. Each ILDG regional grid employs a standard architecture comprising three key components:

- *Metadata Catalog (MDC):* Enables public search and download of metadata, allowing researchers to identify and locate specific datasets.
- *File Catalog (FC):* Maps data identifiers to storage locations, optimizing data retrieval and maintaining metadata consistency.
- *Storage Elements (SE):* Supports long-term data retention on physical storage media, such as disks or tape archives.

All services adhere to open standards like HTTP, SAML, and OAuth2, and follow the further ILDG-wide specifications to enable interoperability and easy integration across networks. A centralized XML document maintains the URLs for regional services of the ILDG.

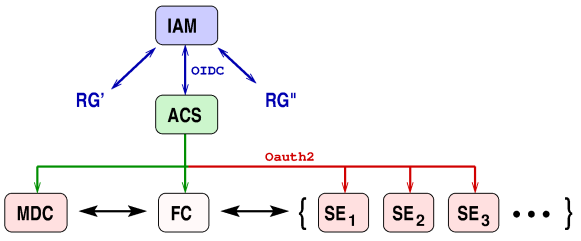

Figure 1: Schematic representation of the different services of the ILDG within a regional grid.

3.3 User Workflow

User interactions within ILDG are structured around two roles: data consumers and data providers. Each role follows a distinct workflow for accessing or contributing to the shared datasets. Schemas of the workflows are reported in Fig. 2

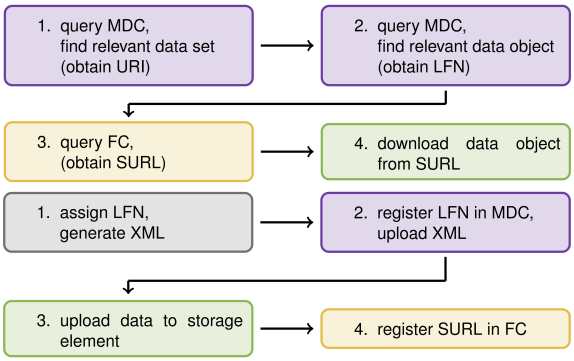

Figure 2: Workflow for data providers and consumers in ILDG. Here, LFN refers to the "Logical Filename" of a data object which holds (one or more) gauge field configurations, i.e. the LFN is a unique identifier, which is independent of the storage location of the data; while SURL refers to the Storage URL, i.e. the qualified path of a data object on a specific SE.

3.4 PUNCH4NFDI Contributions to ILDG 2.0

PUNCH4NFDI has enabled critical developments for the modernization and extension of ILDG and the Regional Grid in continental Europe (LDG) since 2022. Key contributions include a modern, enhanced and more flexible implementation of the catalog services and user interfaces for searching and generation of metadata.

Bringing together the experience and concepts from 20 years of ILDG with new and broader requirements from the multidisciplinary PUNCH use-cases has led to an intense cooperation between ILDG Working Groups and Lattice QCD members of PUNCH. The new middleware and services are designed as flexible building blocks for a FAIR and distributed (meta-) data infrastructure and are to a large extent agnostic to the specific metadata schema or data. The new metadata catalog (MDC) for ILDG 2.0 offers several innovative features:

- *Freely Configurable Metadata Schemata:* Handling of multiple (meta-) data collections, each defined by a distinct XML schema.
- *XPath Queries:* Support complex search functions to find and extract specific (meta-) data sets.
- *RESTful API:* Ensures seamless interaction with other ILDG tools and services.
- *Containerized Deployment:* Simplifies deployment across regional grids, for simpler adoption.

Furthermore, the transition from grid certificates to token-based authentication allows fine-grained access control for all services, improves user experience and enhances data security. The ILDG approach, which uses token exchange between an access control service (ACS) and an IAM (or other AAI), might also be an interesting solution for other communities within PUNCH4NFDI.

4. User Tools and Interfaces for ILDG Services

In order to make ILDG services more usable and attractive for scientists, a set of convenient user tools needs to be developed. These will hide most of the complexity of the service architecture

and streamline the workflows for uploading, querying and downloading data sets. For listing and querying of the ILDG metadata catalogs, a basic web service, called FacetNavi [7], has been developed by the Japanese Lattice Data Gird (JLDG) and is interoperable with other regional grids. One of the goals of PUNCH4NFDI is to develop more powerful and user-friendly tools and interfaces for both command-line and graphical usage.

4.1 Graphical User Interface for Metadata Generation, Queries and Management

The interaction with metadata is essential for data consumers, who are searching for a specific data set, an ensemble in the LQCD use-case, as well as for data providers, who must generate and upload new metadata. A graphical user interface (GUI) under development allows web-based metadata interactions, replacing complex XML and XPath commands with a user-friendly, web-form-based system. The interface is parametric in the metadata schema and guides the users through the metadata hierarchy and fields, thereby generating automatically the corresponding search requests to one or multiple MDC instances.

Since LQCD ensembles are collections of similar configurations, one option in the GUI is to generate XML metadata template files, which can then be appropriately filled with specific values for each individual configuration before upload. A similar approach for markup generation is also explored in JLDG [7].

The GUI enhances data accessibility by abstracting technical details contained in the XML Schema Definition (XSD) files, thus lowering barriers for researchers with limited XML or XPath knowledge. This improvement ensures broader participation and efficient data sharing within the lattice QCD community and beyond.

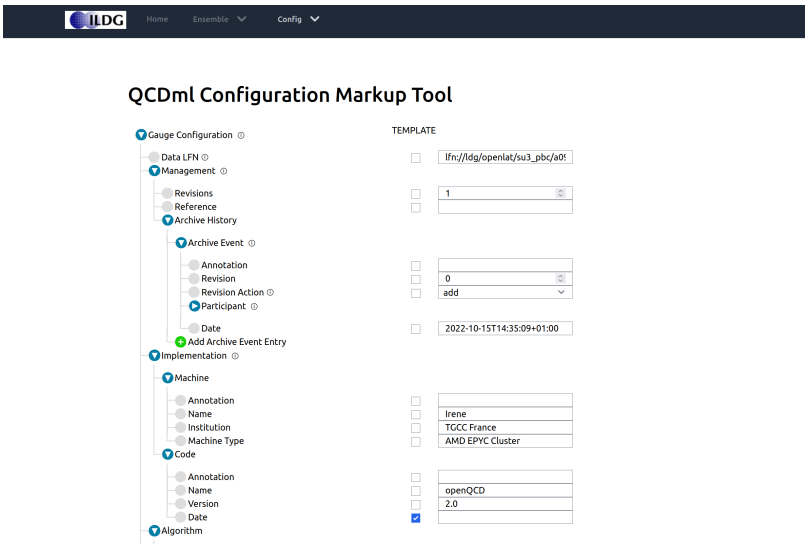

Figure 3: Example of graphical user interface for metadata interactions.

Currently, the interface is being extended to support token-based authentication to enable privileged operations as well, like the upload of a metadata file to the MDC or the registration of a new data file in the FC. This approach is complimentary to command-line-interface tools, which are more suited for scripting and, hence, handling of bulk operations on large data sets.

4.2 Metadata Harvesting Interface

The Protocol for Metadata Harvesting by the Open Archives Initiative, commonly referred to as OAI-PMH [8, 9], defines a standard method to collect information from metadata catalogs. This requires that the metadata in the catalog includes at least the equivalent information as defined by the Dublin Core [10, 11] and DataCite [12] schema.

To enable harvesting of the MDC in such use-case, we developed an add-on service which can transform metadata from the MDC and exposes it through an OAI-PMH-compliant interface. Metadata harvesters can then seamlessly utilize this interface to fetch information from the MDC.

While this interface might already be interesting for suitable metadata collections from PUNCH4NFDI use-cases, it is not yet exploited by ILDG, because the current QCDmlEnsemble metadata schema does not include all fields required by the Dublin Core and DataCite schema. This shortcoming should be addressed in the near future by the ILDG Metadata Working Group.

Conclusion

PUNCH4NFDI and the newly improved efforts in the ILDG exemplify the shift toward FAIR-compliant infrastructures in scientific research. By developing enhanced web services, and robust authentication mechanisms, both initiatives provide lattice QCD researchers and multidisciplinary communities with accessible, reliable data solutions. These advances contribute to a more collaborative and FAIR-compliant future, driving scientific discovery through shared resources and open standards.

Acknowledgments

The authors thank Hubert Simma and Olaf Kaczmarek for the helpful discussions. The authors gratefully acknowledge the used services provided by the PUNCH4NFDI Consortium. This work is funded by the PUNCH4NFDI project. The consortium is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 460248186.

References

[1] C. Davies, A. Irving, R. Kenway and C. Maynard, *International lattice data grid*, *Nucl. Phys. B Proc. Suppl.* **119** (2003) 225.

[2] PUNCH4NFDI consortium, *Particles, Universe, NuClei and Hadrons for the NFDI*, https://www.punch4nfdi.de, (2022).

[3] The Authentication and Authorisation for Research and Collaboration, *Aarc blueprint architecture*, https://aarc-community.org/architecture, (2014).

[4] The International Lattice Data Grid, *Hpc and data for lattice qcd*, https://hpc.desy.de/ildg, (2002).

[5] F. Karsch, H. Simma and T. Yoshie, *The International Lattice Data Grid – towards FAIR data*, *PoS* **LATTICE2022** (2023) 244 [2212.08392].

[6] The Authentication and Authorisation for Research and Collaboration, *Aarc blueprint architecture*, https://aarc-community.org/architecture, (2014).

[7] Japanese Lattice Data Grid (JLDG), *Qcdml faceted navigation 2*, https://www.jldg.org/facetnavi, (2018).

[8] P. Hunter and M. Guy, *Metadata for harvesting: The open archives initiative, and how to find things on the web*, *The Electronic Library* **22** (2004) 168.

[9] Open Archives Initiative, *The open archives initiative protocol for metadata harvesting*, https://www.openarchives.org/OAI/openarchivesprotocol.html, (2002).

[10] S. Weibel, J. Kunze, C. Lagoze and M. Wolf, *Dublin core metadata for resource discovery*, Tech. Rep. RFC Editor, No. rfc2413 (1998).

[11] The Dublin Core Metadata Initiative, *Dublin core schemas*, https://www.dublincore.org/schemas, (2003).

[12] International Data Citation Initiative e.V., *Datacite metadata schema 4.6*, https://datacite-metadata-schema.readthedocs.io/en/4.6, (2003).