Efficient metadata management with the AMI ecosystem

Fabian Lambert[1,*], *Jérôme* Odier[1], *Jérôme* Fulachier[1], and *Pierre-Antoine* Delsart[1]

[1]Laboratoire de Physique Subatomique et de Cosmologie (LPSC) de Grenoble

Abstract. The ATLAS Metadata Interface (AMI) is a comprehensive ecosystem designed for metadata aggregation, transformation, and cataloging. With over 25 years of feedback in the LHC context, it is particularly well-suited for scientific experiments that generate large volumes of data.
This paper explains, in a general manner, why managing metadata is essential regardless of the experiment's scale. It then presents the different AMI ecosystem's components and their main functionalities, particularly the Web interfaces for searching data based on metadata criteria. Finally, it discusses the deployment of a functional demo, its subsequent scaling up, and how to integrate it into a data production system.

1 Metadata challenges in modern sciences

1.1 What are metadata?

Metadata are often described as "data about data." They provide essential information about other data, including its characteristics, content, and context. This information is critical for understanding, organizing, finding, and managing data effectively.

Metadata can originate from a variety of sources. For instance, it might include the version of the software used to generate or process the data, or annotations made by a physicist, such as references to an article in which the data was utilized. These metadata elements help contextualize the data and are crucial for its long-term usability.

1.2 Why Metadata are essential for science

Despite their importance, metadata management is often overlooked in scientific collaborations. However, robust metadata practices are essential for producing reproducible science.

The challenge of reproducibility in science has been well-documented. A 2016 article in Nature[1], based on a survey of approximately 1,500 scientists, revealed that over 70% of researchers had been unable to reproduce another scientist's results, and more than 50% had failed to reproduce their own results.

There are several reasons for this reproducibility crisis, one of which involves software-related issues. Reproducing data generated by a program requires that the program runs under precisely the same conditions as it did originally. Tools like GUIX[2] or Nix[3] can help address these challenges by ensuring consistent computing environments. However, another critical factor is the accurate characterization of data task that relies heavily on metadata.

*e-mail: ami@lpsc.in2p3.fr

1.3 Metadata and the FAIR principles

To enable reproducible science, metadata must align with the FAIR principles. These principles emphasize the need for data to be:

- **Findable**: Metadata plays a key role in characterizing data, enabling users to know where the data is stored and, crucially, what it represents.
- **Accessible**: Metadata should provide clear information on how the data is preserved and the procedures for accessing it.
- **Interoperable**: Metadata, like data itself, must be interoperable. This means it should adhere to standards that allow it to "speak" a universal language, facilitating integration across diverse systems.
- **Reusable**: Metadata must clearly indicate whether data can be reused, under what conditions, and if any restrictions apply.

Adhering to the FAIR principles ensures that metadata supports reproducibility by providing comprehensive documentation and accessibility to both the data and its context.

2 AMI and metadata challenges

The AMI ecosystem[4][5][6] is developed with these principles in mind to address challenges related to metadata management. Over two decades of experience within the ATLAS collaboration[7], along with additional years of work with collaborations like NIKA2[8] and n2EDM[9], have significantly influenced its design and functionality.

Key questions guids the development process: how to enable physicists to select the data they need, how to interact effectively with multiple heterogeneous sources, and how metadata can ensure the long-term reusability of data even after experiments conclude. The latest version of AMI is designed to help physicists characterize and manage the data they find most useful.

3 The AMI Ecosystem in a Nutshell

The AMI ecosystem is essentially a software suite composed of several components that can function independently or together. Its main components include:

- **AMI Web Framework (AWF)**: A modern JavaScript framework that provides developers with controls to create web applications for selecting and displaying data.
- **Backend Microservices**: Developed in Java, the backend offers commands to produce interoperable outputs in XML, JSON, or CSV formats. It interacts with any type of database and provides introspection of the database structure, masking complexity from end users.
- **Task Server**: A distributed system akin to a super CRON, which enables metadata extraction programs to run from primary sources, as well as reprocessing tasks, ultimately saving metadata in AMI.
- **Clients**: Available in various programming languages, with Python being the most widely used, particularly in the ATLAS collaboration.
- **Metadata Query Language (MQL)**: A specialized query language oriented toward metadata, allowing users to select and query data sources without requiring expertise in databases or their underlying structures.

4 Typical Usage of AMI

A typical data and metadata processing workflow (fig.1) demonstrates how the AMI ecosystem integrates within a broader data production system and aggregates heterogeneous data sources into a dedicated metadata database.

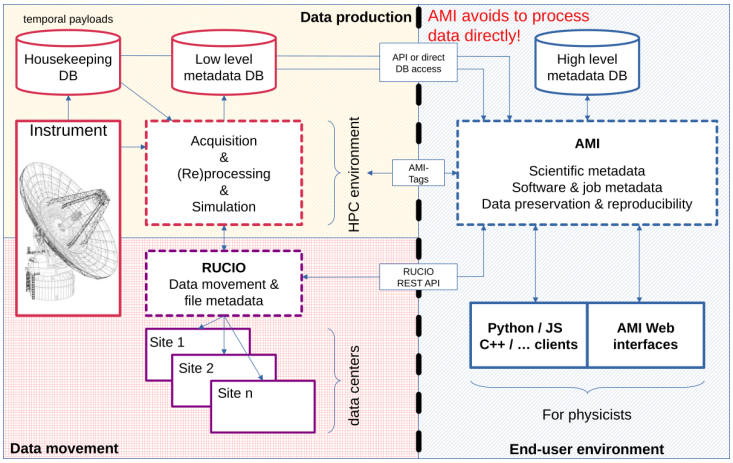

Figure 1. A typical data and metadata workflow.

In most experiments, data generation begins with an instrument, and the first source of metadata is often a configuration database ("HouseKeeping DB"), containing information about the instrument's configuration or simulation setup. During data production, software running analysis, reprocessing, or simulation tasks generates additional metadata, such as task status, which is typically stored in a separate database.

A data placement tool (e.g., Rucio) often manages data movements and backups, tracking the physical storage, size, and other details of the data. It provides metadata related to data location and management.

AMI collects metadata from these disparate sources and stores it in a centralized metadata database. Additionally, physicists contribute valuable metadata, such as references to scientific papers citing or utilizing the data.

When an experiment concludes, instruments may be dismantled, and associated software may become obsolete. Even if data is archived, tools to access it may no longer be available. A dedicated metadata database ensures data remains discoverable, characterized, and accessible long after an experiment ends, simplifying future access and reuse.

5 End-users and AMI

The AMI ecosystem is designed to enable physicists to effectively characterize their data while accommodating a wide range of user profiles and expertise levels. These users can include those who require web or command-line interfaces and others who prefer to write scripts for direct data access. Many users, however, may not possess advanced knowledge of databases or SQL, nor familiarity with the structure of the various databases they need to access.

To address these diverse needs, AMI provides multiple solutions. User-friendly web interfaces allow for intuitive point-and-click functionality, while client tools cater to advanced users seeking to automate workflows through scripts. The Metadata Query Language (MQL) further simplifies data access by abstracting database complexity, enabling users to retrieve information without requiring expertise in database systems or structures.

6 MQL: A Metadata-Oriented Language

The Metadata Query Language (MQL) was specifically developed to provide an intuitive and user-friendly method for querying metadata. Unlike SQL, which operates on database objects such as tables and requires knowledge of their relationships (fig.2), MQL is oriented around metadata entities. These entities encapsulate characteristics relevant to scientific datasets, such as dataset status or keywords (fig.3).

```
SELECT *
FROM DATASET, DATASET_KEYWORDS
WHERE
  DATASET.STATUS = 'VALID' AND DATASET_KEYWORDS.KEYWORD = '1jet'
  AND
  DATASET_KEYWORDS.DATASETFK = DATASET.IDENTIFIER
```

Figure 2. An SQL query with join clauses.

```
SELECT *
WHERE
  DATASET.STATUS = 'VALID' AND DATASET_KEYWORDS.KEYWORD = '1jet'
```

Figure 3. An MQL equivalent query on dataset entity.

MQL queries only have SELECT and WHERE clauses, i.e., without a FROM clause, allowing users to focus on the information they want to retrieve. The underlying AMI system handles the translation of these queries into the specific database commands needed. By eliminating the need for users to understand database structures or implement joins, MQL makes metadata access more accessible to a broader audience, including those without technical expertise in database systems.

7 Interfaces of the AMI Ecosystem

7.1 Search by Criteria

The default data selection tool in AMI (fig.4) allows users to refine their search by clicking on entity characteristics displayed on the left side. Each criterion adapts to its type (e.g., string, boolean, or number), enabling intuitive filtering. The system executes an MQL query behind the scenes, simplifying interaction for point-and-click users.

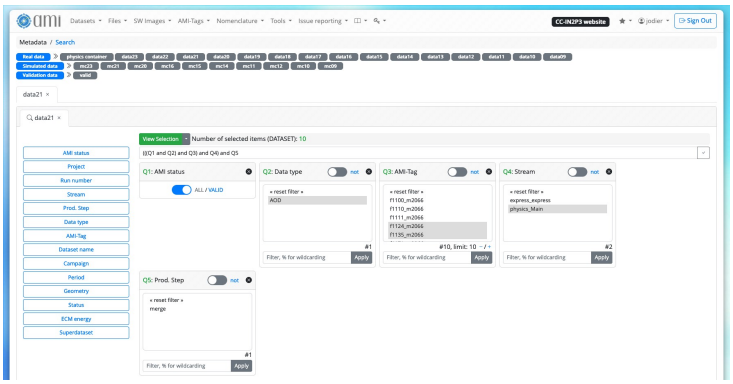

Figure 4. Search by criteria Web interface.

7.2 Search result

Search results appear in a customizable table format (fig.5). Features include grouping, filtering, and linking metadata elements, such as datasets and their associated files, through configurable icons.

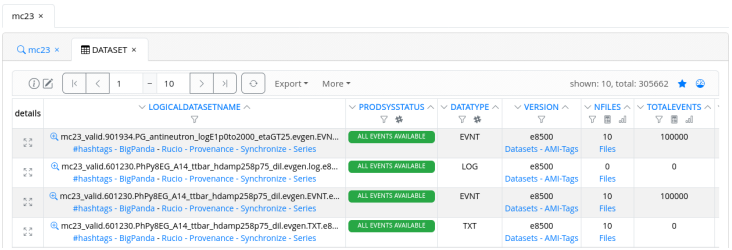

Figure 5. Search result Web interface.

7.3 Search Modeler

Administrators can use this intuitive interface (fig.6) to build search tools for end users. By defining a target entity, the system suggests potential characteristics for search criteria, automatically generating the user interface. This automation is made possible by AMI's database introspection capabilities.

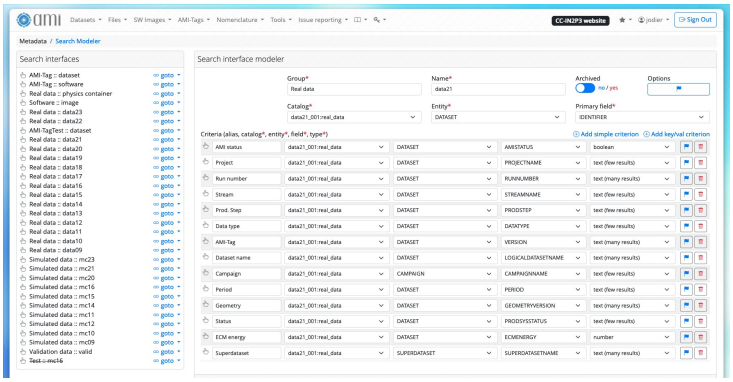

Figure 6. The search modeler Web application.

8 Interacting with Microservices

8.1 Web Interaction

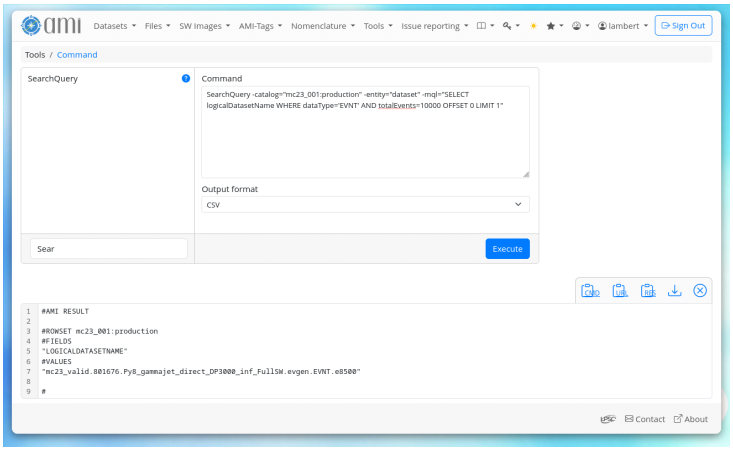

Figure 7. Executing MQL query from a Web application.

Users can execute MQL queries via a web application by entering commands in a form and selecting output formats like XML, JSON, or CSV (fig.7).

8.2 Shell Interaction

Advanced users can execute the same queries from a shell script after installing the AMI client (fig.8). This method supports automation and script-based workflows.

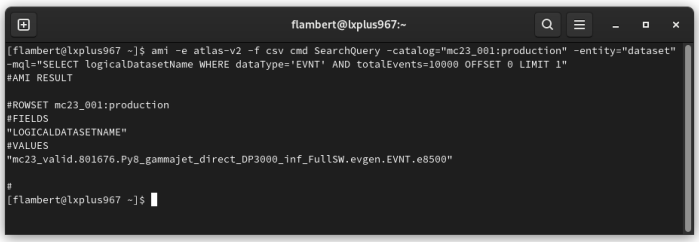

Figure 8. MQL query command executed from a shell.

8.3 Python Interaction

Developers can embed AMI commands in Python scripts, integrating metadata queries into larger workflows for automation or complex analyses (fig.9).

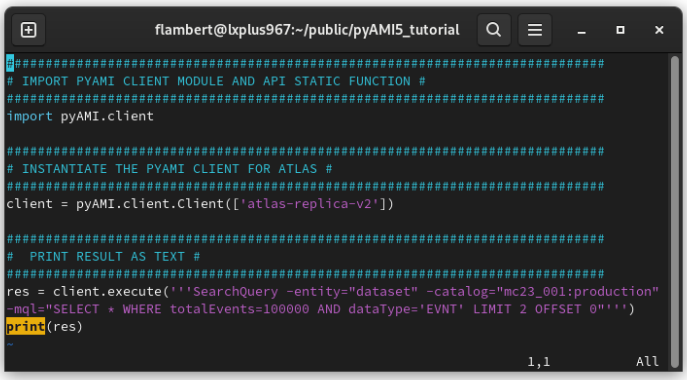

Figure 9. MQL query command executed from Python.

9 Task Server: Streamlining Metadata Extraction

9.1 Task Server Interface

The task server enables the execution of programs to extract metadata from primary sources. The interface displays the program being run along with an execution report (fig.10).

9.2 Pipelined Tasks

An experimental feature allows users to configure chained tasks, where outputs from one task can serve as inputs for another (fig.11). This capability aims to streamline workflows by automating complex metadata processes.

10 Explore AMI

To try AMI, visit the official Web site[10], explore the online demo[11][12] updated hourly. Documentation[13] and additional resources are also available to support new users.

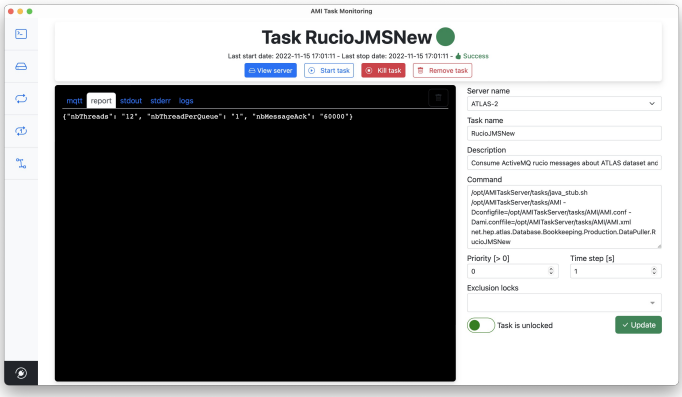

Figure 10. Configuring a recurrent task.

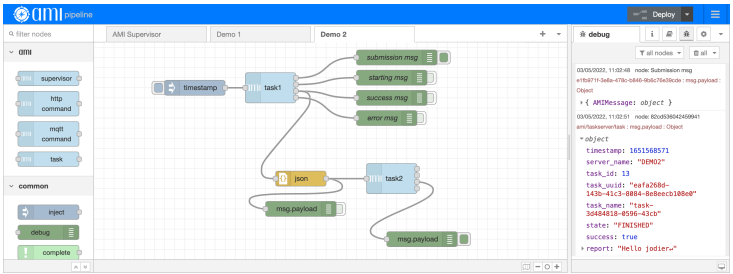

Figure 11. Chained task executions.

11 Conclusion

The ATLAS Metadata Interface (AMI) ecosystem, developed within the ATLAS experiment and other scientific collaborations, represents a crucial tool for managing and accessing scientific data. Its flexible and scalable approach enables the aggregation, transformation, and cataloging of metadata from heterogeneous sources while providing simplified and intuitive access for users of all expertise levels.

With the integration of web interfaces, backend services, and the dedicated Metadata Query Language (MQL), AMI offers a robust solution for researchers seeking to efficiently exploit their data in a reproducible manner. The alignment of AMI with the FAIR principles ensures the long-term sustainability, interoperability, and reusability of scientific data, while simplifying their discovery and management within large collaborations.

Through over 25 years of development and feedback, AMI continues to evolve to address the growing challenges of metadata management in modern scientific experiments. By facilitating metadata access and manipulation, AMI plays a key role in long-term data stewardship and ensuring the reproducibility of research.

References

[1] Baker, M. 1,500 scientists lift the lid on reproducibility. Nature **533**:452–454 (2016). https://doi.org/10.1038/533452a

[2] Vallet, N., Michonneau, D. & Tournier, S. Toward practical transparent verifiable and long-term reproducible research using Guix. Sci Data **9**, 597 (2022). https://doi.org/10.1038/s41597-022-01720-9

[3] Package manager and open source Linux distribution: https://nixos.org/ [accessed 2025-01-10]

[4] J. Fulachier, O. Aidel, S. Albrand, F. Lambert, Looking back on 10 years of the ATLAS Metadata Interface. Proceedings of the 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP) J. Phys.: Conf. Ser. **513**, 042019 (2013). https://doi.org/10.1088/1742-6596/513/4/042019

[5] J. Odier, O. Aidel, S. Albrand, J. Fulachier, F. Lambert, Evolution of the architecture of the ATLAS Metadata Interface (AMI). Proceedings of the 21st International Conference on Computing in High Energy and Nuclear Physics (CHEP) J. Phys.: Conf. Ser. **664**, 042040 (2015). https://doi.org/10.1088/1742-6596/664/4/042040

[6] J. Odier, F. Lambert, J. Fulachier, The ATLAS Metadata Interface (AMI) 2.0 metadata ecosystem: new design principles and features. Proceedings of the 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP) EPJ Web of Conf.: Conf. Ser. **214**, 05046 (2019). https://doi.org/10.1051/epjconf/201921405046

[7] The ATLAS Collaboration et al., The ATLAS Experiment at the CERN Large Hadron Collider. JINST **3**, S08003 (2008). https://doi.org/10.1088/1748-0221/3/08/S08003

[8] Calvo, M., Benoît, A., Catalano, A. et al., The NIKA2 Instrument, A Dual-Band Kilopixel KID Array for Millimetric Astronomy. J Low Temp Phys **184**, 816–823 (2016). https://doi.org/10.1007/s10909-016-1582-0

[9] Ayres, N.J., Ban, G., Bienstman, L. et al., The design of the n2EDM experiment. Eur. Phys. J. C **81**, 512 (2021). https://doi.org/10.1140/epjc/s10052-021-09298-z

[10] AMI Web site: http://ami-ecosystem.in2p3.fr [accessed 2025-01-10]

[11] AMI demo: http://demo.ami-ecosystem.in2p3.fr:667/ [accessed 2025-01-10]

[12] AMI demo repository: https://github.com/ami-team/AMIDemo [accessed 2025-01-10]

[13] AMI documentation: https://ami-ecosystem.in2p3.fr/doc/ [accessed 2025-01-10]