Metadata definition and database implementation for searchable Raman-spectroscopy data handling

Irio Lavagno
Tiberlab srl
Rome, Italy
irio@lavagno.org

Alessandro Pecchia
CNR-ISMN
Consiglio Nazionale delle Ricerche
Rome, Italy
alessandro.pecchia@cnr.it

Alessia Di Vito
Dept. of Electronic Engineering
University of Rome Tor Vergata
Rome, Italy
alessia.di.vito@uniroma2.it

Matthias Auf der Maur
Dept. of Electronic Engineering
University of Rome Tor Vergata
Rome, Italy
auf.der.maur@ing.uniroma2.it

***Abstract*—Optical nanoscale metrological techniques like Raman spectroscopy produce large amounts of data, especially if employed as a default characterisation tool in industrial production lines. Proper management of huge amounts of data is needed, in order to allow for their further use in data analysis, automatization or in machine learning applications. The definition of metadata and the design of the database structure are paramount in these contexts. This work presents the prototype development of such data management system, showcasing effective data and metadata management in compliance with EU defined standards for scientific research. The data handling of tip-enhanced Raman spectroscopy (TERS) measurements is used as a benchmark. Metadata and data format definitions as well as the database structure implementation are described.**

***Index Terms*—metadata, database, FAIR data management, tip-enhanced Raman spectroscopy**

I. Introduction

Ongoing research efforts [1] are dedicated to the development of Scanning Probe Microscopy (SPM) platforms capable of operating within production lines. A significant volume of data is anticipated in the development, testing, and automation of an SPM-based tool with real-time capabilities suitable for a real industrial production environment. Moreover, the instrumentation generated in this framework will gather various kinds of research data, which must be handled in accordance with EU defined standards [2]. Therefore, the establishment of a robust data environment, conforming to EU defined standards for the efficient management of experimental data, is crucial.

Here, we present the prototype of a data management system that allows to store, manage, and query the data and metadata generated by instruments, particularly Raman spectrometers. We describe the implementation, the technical choices, and the data flow within the application. The FAIR (findable, accessible, interoperable, re-usable) data management required by EU standards for research data is addressed.

II. Metadata and Common Data Format

The effective management of research data requires the formal definition of metadata to characterize the experimental and data management environment. Information entities, including the instrument used, the experimental setup, the specific sample, and the generated raw data, are described by different metadata elements. In order to ensure findability, we aim at a set of standard terms for the metadata, which might be conveniently chosen to coincide with the ones being defined within NFDI4Chem (Vibrational Spectroscopy Ontology) [3], [4] or CHARISMA [5] or similar community-driven standardization initiatives. In our implementation, metadata are extracted directly from the instruments' data files, where possible, and translated to a standard format. However, since not all instruments provide the same set of metadata, usually intervention of the user might be necessary to integrate it with additional input.

III. Database Structure Definition

To build the primary container for the files generated by different instruments, the HDF5 database architecture [6], [7] has been chosen, which supports fast I/O processing and storage, as well as management, processing, and storage of structured heterogeneous data. HDF5 metadata management was too limited for our needs, however, so we implemented a Volume Object Layer (VOL) [8] that stores the metadata into a SQLITE3 database [9]. This approach allows to perform more sophisticated queries on the metadata.

A schematic view of the designed data management system is sketched in Figure 1. The Web Application offers an easy user interface to the application. The REST API offers a programmatic access to the application. The Data Application Interface is the core part through which data and metadata are inserted into the database. HDF5 is where the data are stored. VOL is a plugin that adds the ability to query a SQL database to the HDF5 storage. SQLite is the library implementing a portable database. VFS is a virtual filesystem through which is possible to interact with the application (alternative to the web or REST interface). Instruments and datafiles are the actual instruments that generate the data, not part of the application but inserted in the scheme to show the interaction.

A. Data Access Interface

To make data retrieval and manipulation more accessible and intuitive, we added a Java program as a front-end towards the HDF5 library itself. The Java program offers two different ways to interacting with the data, through a virtual file system

We wish to acknowledge the support of European project Challenges, under grant agreement number 861857.

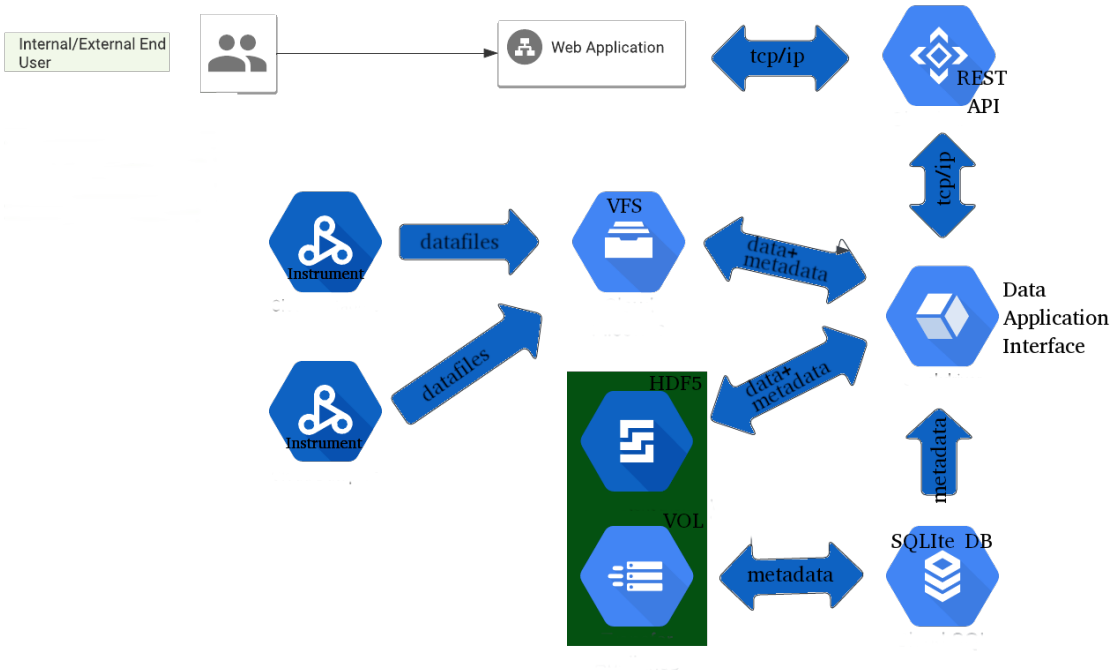

Fig. 1. Schematic diagram of the system architecture.

and the SQLite database. A virtual filesystem has been created (so far, it has been tested on Linux, but it is designed to work on Windows as well) in which directories are mapped to HDF5 groups, and files to HDF5 datasets, so it is possible to store a data file into the HDF5 file simply by copying (or dragging) it into a virtual directory. Metadata are automatically extracted and stored in the database if the inserted file is of a format recognized by the system. At the moment only files with the extension .wdf are processed, but more types will be added in the future.

IV. Conclusions

While the current implementation has some limitations and could benefit from improved user-friendliness, it serves as proof of concept for a data management system capable of gathering and processing TERS data from different instruments, achieving a level of homogeneity, and allowing for complex data queries. The next steps will focus on code reorganization, optimization, and refactoring of the VOL plugin and the Data Interface.

This paper contributes to the ongoing dialogue surrounding effective data management in scientific research projects, offering a practical example of how such systems can be developed and implemented.

References

[1] European Project Challenges, grant agreement number 861857.

[2] A. Ntziouni et al., Review of Existing Standards, Guides, and Practices for Raman Spectroscopy. Applied Spectroscopy 76(7), pp. 747–772, 2022.

[3] Chemistry Consortium in the NFDI https://www.nfdi4chem.de/index.php/objectives/

[4] Vibrational Spectroscopy Ontology (work in progress) https://nfdi4chem.github.io/VibrationalSpectroscopyOntology/

[5] Characterisation and HARmonisation for Industrial Standardisation of Advanced MAterials https://wiki.charisma.ideaconsult.net/

[6] The HDF Group. Hierarchical Data Format, version 5, 1997-NNNN, https://www.hdfgroup.org/HDF5/

[7] M. Folk et al., An overview of the HDF5 technology suite and its applications. Proceedings of the EDBT/ICDT 2011 workshop on array databases, 2011.

[8] O. Perevalova, Database VOL-plugin for HDF5. Diss. Universität Hamburg, 2017.

[9] SQLite, https://www.sqlite.org/index.html