Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 3502

A FAIR research data infrastructure for high-throughput digital chemistry

Alice Gauthier, †[a] Laure Vancauwenberghe, †[b] Jean-Charles Cousty, *[a] Cyril Matthey-Doret, [b] Robin Franken, [b] Sabine Maennel, [b] Pascal Miéville [a] and Oksana Riba Grognuz [b]

The growing demand for reproducible, high-throughput chemical experimentation calls for scalable digital infrastructures that support automation, traceability, and AI-readiness. A dedicated research data infrastructure (RDI) developed within Swiss Cat+ is presented, integrating automated synthesis, multi-stage analytics, and semantic modeling. It captures each experimental step in a structured, machine-interpretable format, forming a scalable, and interoperable data backbone. By systematically recording both successful and failed experiments, the RDI ensures data completeness, strengthens traceability, and enables the creation of bias-resilient datasets essential for robust AI model development. Built on Kubernetes and Argo Workflows and aligned with FAIR principles, the RDI transforms experimental metadata into validated Resource Description Framework (RDF) graphs using an ontology-driven semantic model. These graphs are accessible through a web interface and SPARQL endpoint, facilitating integration with downstream AI and analysis pipelines. Key features include a modular RDF converter and 'Matryoshka files', which encapsulate complete experiments with raw data and metadata in a portable, standardized ZIP format. This approach supports scalable querying and sets the stage for standardized data sharing and autonomous experimentation.

Received 7th July 2025
Accepted 21st October 2025

DOI: 10.1039/d5dd00297d

rsc.li/digitaldiscovery

1 Introduction

Swiss Cat+[1] West hub at EPFL (École Polytechnique Fédérale de Lausanne) is an automated laboratory designed to perform high-throughput chemistry experiments with minimal human input. This sophisticated setup allows the generation of large volumes of both synthetic and analytical data, far exceeding what would be feasible through manual experimentation.[2] Automation of experimental workflows not only increases throughput, but also ensures consistency and reproducibility of the resulting data.

In recent years, the field of chemistry has lagged behind other scientific disciplines, *e.g.* molecular biology, with impressive achievements such as AlphaFold,[3] in the application of artificial intelligence. A major contributing factor to this delay is the lack of comprehensive and standardized data.[4–6] Most available datasets focus solely on successful outcomes, often excluding unsuccessful synthesis attempts, which are equally informative for data-driven modeling. The absence of detailed and traceable negative data points creates significant limitations in training robust AI systems capable of learning the full experimental landscape. In view of improving this situation, several initiatives, such as the Open Reaction Database (ORD),[7] have been developed in recent years. ORD is a shared database where research groups can upload fully structured and digitally compatible chemical reaction data.

To ensure the highest degree of integrity, interoperability and reusability of data generated at the Swiss Cat+ West hub, experimental information must be systematically captured and linked across the entire workflow. This approach enables the reuse of high-quality structured data in initiatives such as ORD and supports the development of robust AI models in chemistry. To achieve this, a Research Data Infrastructure (RDI) has been developed for the Swiss Cat+ West hub utilizing open-source components. RDIs are community-driven platforms for standardizing and sharing data, code, and domain knowledge. They begin with fragmented or siloed research outputs and progressively transform into reusable, findable, and interoperable resources. The first step is to apply common standards and make the data findable and accessible to the wider community. Once such standards are in place, it becomes easier to develop reusable and interoperable building blocks that many actors can benefit from. By mutualizing data resources and tools, RDIs play a key role in assembling high-quality datasets for large-scale analysis.

The RDI is designed from the ground up to serve data to researchers in a way that adheres to the FAIR principles:[8,9]

[a]*Swiss Cat+ West Hub, Ecole Polytechnique Fédérale de Lausanne EPFL, 1015 Lausanne, Switzerland. E-mail: jean-charles.cousty@epfl.ch*

[b]*Swiss Data Science Center – Open Research Data Engagement & Services, EPFL, INN Building, Station 14, 1015 Lausanne, Switzerland*

† Authors contributed equally.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

findability, accessibility, interoperability and reusability. Findability is supported through rich metadata indexed in a searchable front-end interface. Accessibility is ensured by providing data sets to researchers upon request, with access controlled through a licensing agreement. Interoperability is achieved by mapping metadata to a structured ontology,[10] which incorporates established chemical standards such as the Allotrope Foundation Ontology[11] (https://www.allotrope.org/ontologies). Reusability is enabled by providing detailed, standardized metadata and clear provenance, allowing datasets to be understood and applied beyond their original context. In addition to FAIR principles, reproducibility is a key strength of the system, made possible through the automation and standardization of both data generation and metadata capture. This ensures that the same workflows can be reliably implemented in other laboratories adopting similar infrastructure.

In this context, the HT-CHEMBORD (High-Throughput Chemistry Based Open Research Database) project provides an RDI for processing and sharing high-throughput chemical data. It is a collaborative project developed by Swiss Cat+ and Swiss Data Science Center (SDSC), with technical support from SWITCH,[12] the Swiss national foundation providing secure digital infrastructure for research and education. The platform is built on open-source technologies and is deployed using SWITCH's Kubernetes-as-a-Service,[13] enabling scalable and automated data processing. Each week, experimental metadata are converted to semantic metadata, Resource Description Framework (RDF),[14] using a general converter, and stored in a semantic database. These structured datasets can be queried directly by experienced users through SPARQL,[15] or accessed through a user-friendly web interface. The entire pipeline is automated using Argo Workflows,[16] with scheduled synchronizations and backup workflows to ensure data reliability and accessibility. This infrastructure aims to serve the entire chemistry community, by providing access to well-structured, high-throughput experimental data that can be browsed and downloaded by authorized users.

The experimental workflow architecture implemented on the Swiss Cat+ West hub platform is divided into two main parts: the synthesis platform for automated chemical reactions and the analytical platform equipped with instrumentation provided primarily by two major suppliers: Agilent[17] and Bruker.[18] This setup facilitates the generation of harmonized datasets between analytical techniques by reducing variability and promoting data standardization. Agilent and Bruker instruments are indicated with different dashed box styles: long dashes for Agilent and alternating dot-dash lines for Bruker, as shown in the workflow in Fig. 1. All intermediate and final data products are stored in structured formats depending on the analytical method and instrument: the Allotrope Simple Model[19]-JavaScript Object Notation (ASM-JSON), JSON or Extensible Markup Language (XML)[20] format. These formats support automated data integration, reproducibility, and downstream machine learning applications.

The proposed architecture is designed as a modular, end-to-end digital workflow, where each system component

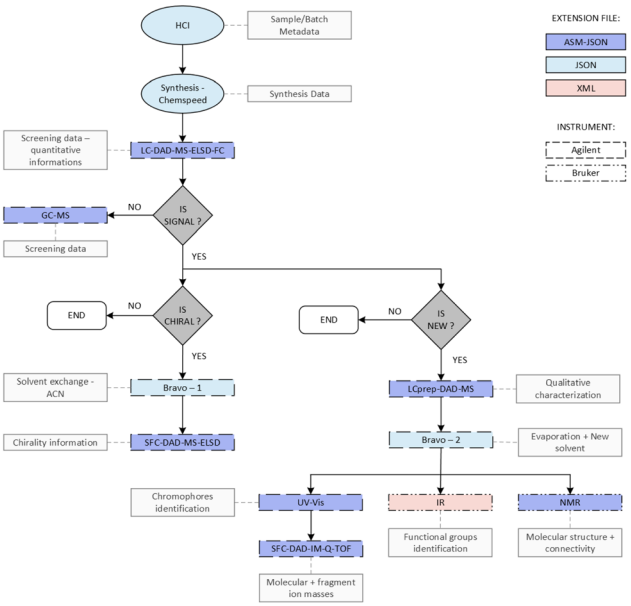


Fig. 1 Flowchart of the workflow architecture and standardized data output formats implemented at the Swiss Cat+ West hub – the process begins with Human–Computer Interface (HCI) guiding automated synthesis using Chemspeed systems, followed by a decision diagram based on signal detection, chirality, and novelty. Based on these criteria, samples undergo various screening and characterization steps using LC, GC, SFC, UV-Vis, FT-IR, and NMR techniques. Instrument-specific outputs are stored in structured formats: ASM-JSON (dark blue), JSON (light blue), or XML (pink), depending on the analytical method and hardware supplier. Each analysis step generates semantically annotated data, supporting downstream integration and interoperability across the platform.

communicates through standardized metadata schemes. It implements a fully digitized and reproducible platform for automated chemical discovery that captures the complete experimental context, including negative results, branching decisions, and intermediate steps such as solvent changes or evaporation. Through the development of HT-CHEMBORD, the Swiss Cat+ West hub addresses key challenges in standardization and reproducibility faced in modern chemical research. Beyond accelerating discovery, the system provides the groundwork for autonomous experimentation and predictive synthesis through data-driven approaches. The workflow begins with the digital initialization of the project through a Human–Computer Interface (HCI).[21] This interface enables structured input of sample and batch metadata, which are formatted and stored in a standardized JSON format. This metadata includes reaction conditions, reagents structures and batch identifiers, ensuring traceability and data integrity across all stages of experimentation.

Following metadata registration, compound synthesis is carried out using the Chemspeed automated platforms (2 Swing XL into Gloveboxes),[22] which enables parallel, programmable chemical synthesis under controlled conditions (*e.g.*, temperature, pressure, light frequency, shaking, stirring). These programmable parameters are essential to reproduce experimental conditions across different reaction campaigns. In

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

addition, the use of such parameters facilitates the establishment of structure–property relationships. Reaction conditions, yields, and other synthesis-related parameters are automatically logged using the ArkSuite software,[23] which generates structured synthesis data in JSON format. This file serves as the entry point for the subsequent analytical characterization pipeline.

Upon completion of synthesis, compounds (referenced as "`Products`" throughout our ontology) are subjected to a multistage analytical workflow designed for both fast reaction screening (screening path) and in-depth structural elucidation (characterization path), depending on the properties of each sample. The screening path is dedicated to the rapid assessment of reaction outcomes through known product identification, semi-quantification, yield analysis, and enantiomeric excess (ee) evaluation. In parallel, the characterization path supports the discovery of new molecules by leveraging detailed chromatographic and spectroscopic analyses. The first analytical step involves Liquid Chromatography coupled with Diode Array Detector, Mass Spectrometry, Evaporative Light Scattering Detector and Fraction Collector (LC-DAD-MS-ELSD-FC, Agilent), where compounds are screened to obtain quantitative information, using retention times of each detector. In the absence of a detectable signal, samples are redirected to Gas Chromatography coupled with Mass Spectrometry (GC-MS, Agilent) for complementary screening, typically for volatile or thermally stable species. All output data from these screening techniques are captured in ASM-JSON format to ensure consistency across analytical modules. If no signal is observed from either method, the process is terminated for the respective compound. The associated metadata, representing a failed detection event, is retained within the infrastructure for future analysis and machine learning training.

If a signal is detected, the next decision point concerns the chirality of the compound (screening path) and the novelty of its structure (characterization path). If the compound is identified as achiral (screening path), the analytical pipeline is considered complete. Otherwise, a solvent exchange to acetonitrile is performed prior to purification using Bravo instrument (Agilent). This solvent exchange ensures compatibility with subsequent chiral chromatography conditions. The purified sample is then analyzed by Supercritical Fluid Chromatography coupled with Diode Array Detector, Mass Spectrometry and Evaporative Light Scattering Detector (SFC-DAD-MS-ELSD, Agilent) to resolve enantiomers and characterize chirality. The method offers high-resolution separation based on stereochemistry, combined with orthogonal detection modes for confirmation and quantification.

The second decision point addresses the novelty of the molecular structure (characterization path). If the compound is not novel and has been previously characterized in the internal database, it is excluded from further analysis to avoid redundancy. However, if the compound is considered new, it undergoes preparative purification *via* Preparative Liquid Chromatography coupled with Diode Array Detector and Mass Spectrometry (LCprep-DAD-MS, Agilent). Solvent from the purified fraction is evaporated and exchanged using the Bravo instrument (Agilent) to prepare the sample for advanced characterization. Structural information is subsequently acquired using Fourier Transform Infrared (FT-IR, Bruker) spectroscopy for functional group identification (data exported in XML format). Nuclear Magnetic Resonance (NMR, Bruker) spectroscopy is then performed for molecular structure and connectivity analysis (ASM-JSON format), often involving both 1D and 2D experiments. Additional characterization includes Ultraviolet-Visible (UV-Vis, Agilent) spectroscopy for chromophore identification. Supercritical Fluid Chromatography coupled with Diode Array Detector, Ion Mobility Spectrometry and Quadrupole Time-of-Flight Spectrometry (SFC-DAD-IM-Q-TOF, Agilent) is used for high-resolution mass spectrometry. This technique provides accurate mass and fragmentation data, along with ion mobility separation for conformer and isomer discrimination. Each of these datasets is formatted in ASM-JSON,[19] ensuring full interoperability with the broader Swiss Cat+ data infrastructure and enabling integration into machine learning pipelines, structural databases, and retro-synthetic planning tools.

2 Data infrastructure and workflow

As introduced earlier, a dedicated RDI is central for achieving FAIR principles, reproducibility, and AI-readiness in chemical research. This section details the complete data infrastructure and processing logic developed within Swiss Cat+ to achieve these goals. The diagrams in Fig. 2 and 3 provide an overview of the complete RDI, outlining key processes such as metadata extraction and normalization, RDF conversion and validation, automated orchestration *via* Kubernetes,[13] database interactions, and user-facing access. Together, these diagrams outline an integrated system that manages sample tracking, Matryoshka file handling, semantic mapping strategies, and backend services, forming the foundation for scalable, interoperable, and machine-actionable chemical research.

2.1 Ontology-driven semantic framework

A core objective of the Swiss Cat+ HT-CHEMBORD RDI is to achieve full semantic interoperability across the synthesis, screening, and characterization stages of high-throughput chemical experimentation. Each experiment generates multiple data files across different instruments and platforms, making consistent interpretation and integration a major challenge. To address this, Swiss Cat+ adopts an ontology-based approach to metadata modeling and data representation.

In computer science, an ontology refers to *"the specification of a conceptualization"*:[24] a formal model that defines the vocabulary such as concepts, the class hierarchy or taxonomy, and the interrelationships necessary to describe a domain of knowledge in a machine-readable manner. Originally rooted in philosophy as the study of *"the subject of existence"*,[24] ontologies in modern data science provide structured frameworks for encoding knowledge, ensuring that information is not only syntactically but also semantically interoperable across systems. This emphasis on semantic modeling reflects a broader transition in scientific data management: from siloed data capture to data-centric, ontology-driven science, where the standardization, linkage, and enrichment of experimental metadata

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

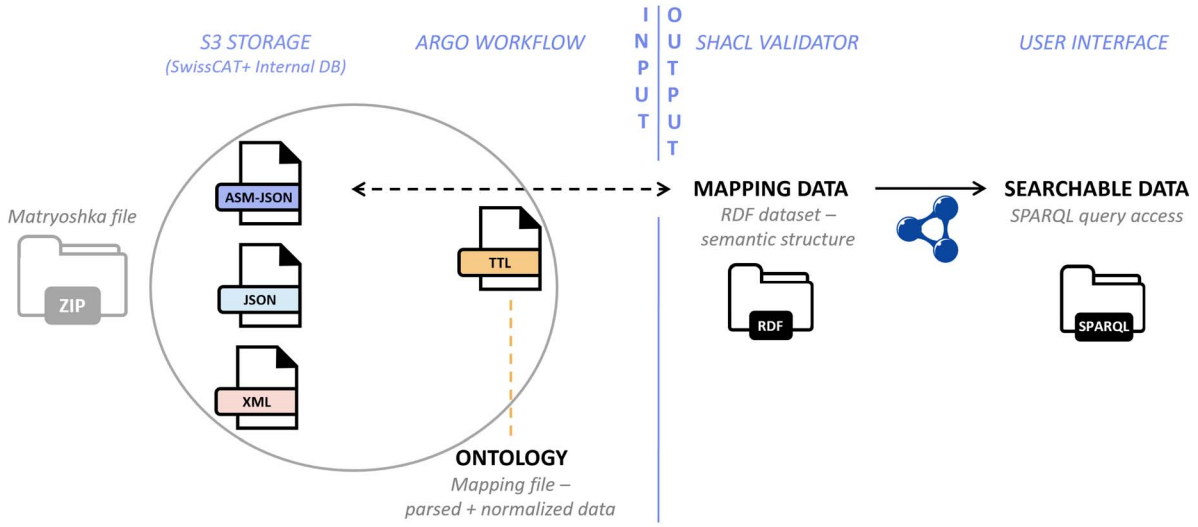


Fig. 2 End-to-end data handling pipeline in the Swiss Cat+ platform – the process begins with Matryoshka files (ZIP archives containing nested formats such as XML, JSON, and ASM-JSON) which are converted into semantically structured metadata. Data stored in an internal S3 (simple, storage, service) database is parsed and normalized through an argo-based workflow into turtle (TTL) files. These files are validated using a SHACL validator and converted into RDF graphs, which are then made accessible through a SPARQL query endpoint. The resulting structure supports user interface interaction and semantic search across experimental metadata.

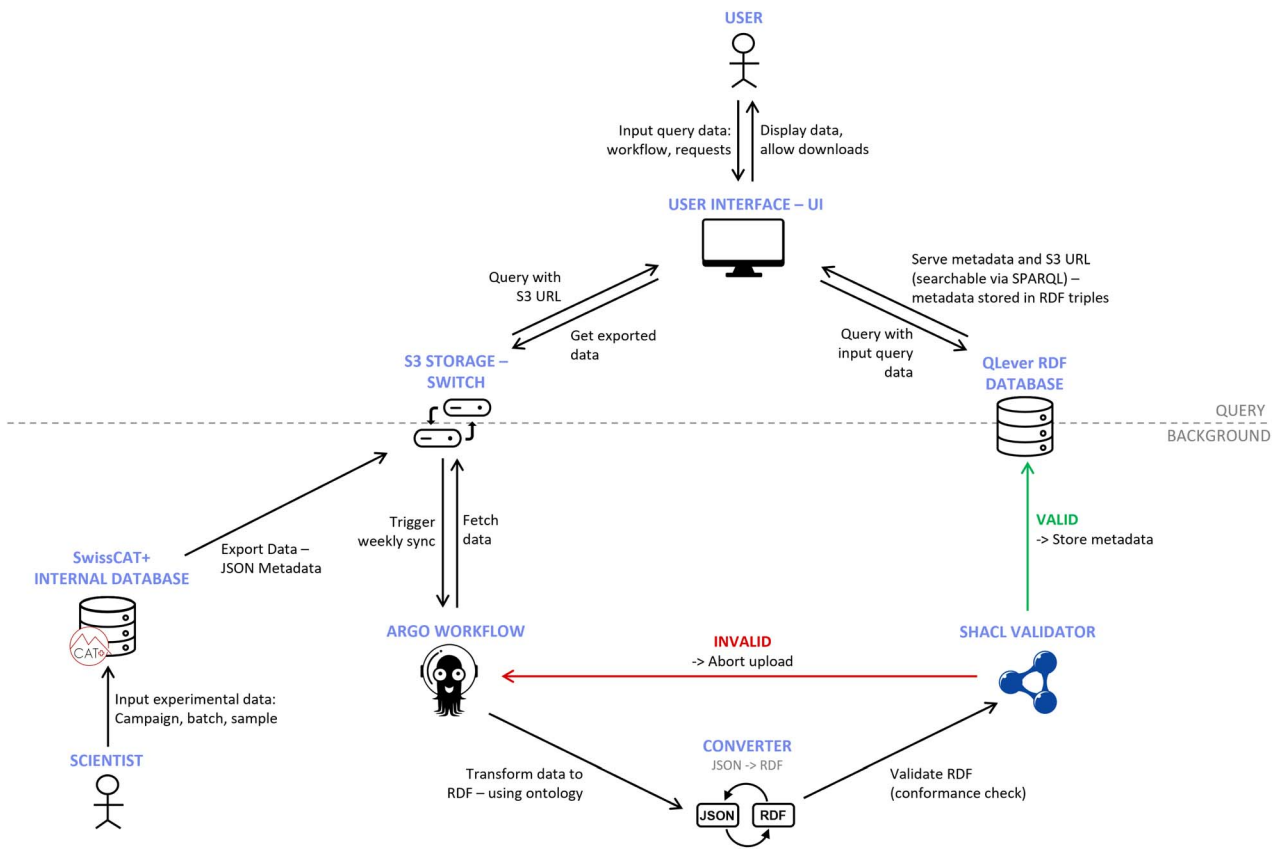


Fig. 3 Weekly synchronization infrastructure for semantic validation and RDF integration in an automated laboratory, Swiss Cat+ – the architecture presents an automated data pipeline, highlighting the separation between internal components (Swiss Cat+ storage) and external services (workflow management, RDF conversion, queryable RDF database and user-facing interface). Experimental metadata stored in the internal database is synchronized weekly to an S3 object store, then processed *via* an Argo Workflow that converts JSON to RDF. The resulting data is validated using a SHACL validator before being ingested into the QLever RDF database. The system also includes a feedback mechanism: RDF files that fail validation are redirected to the workflow for correction and reprocessing.

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

become prerequisites for reproducibility, integration, and AI-readiness.

To ensure that data generated across the experimental workflow can be integrated, interpreted, and reused, our RDI implements a consistent semantic layer that formally describes, links, and enables querying of all key entities – such as samples, conditions, instruments, and results – and provides the foundation on which the infrastructure components are developed. This semantic layer is realized by mapping metadata from all file types to an ontology represented using the Resource Description Framework (RDF).[14] RDF enables structured representation of data in the form of subject-predicate-object triples,[25–27] making relationships between data points explicit and machine-readable. This approach is particularly well-suited to scientific data because it supports linked data principles and semantic search, which are essential for integrating and querying across heterogeneous datasets.

The ontology-driven mapping process begins with the identification of terms in the metadata. Each term is systematically searched in the Allotrope Foundation Ontologies Dictionary (AFO), a curated semantic vocabulary developed by the Allotrope Foundation.[11] When a match is found, the term is directly adopted and integrated into the ontology with its formal definition. It is designated either as an object (class) or as a property (predicate), and where appropriate, a unit of measurement is assigned using existing unit ontologies (QUDT[28] – Quantities, Units, Dimensions, and DataTypes) or custom units defined in the Swiss Cat+ ontology. To support this process and help address the complexity of vendor-specific metadata formats, the Allotrope model provides a standardized structure to represent essential concepts, including samples, measurements, methods, instruments, and analytical outputs. These elements are semantically anchored to terms from the Allotrope Foundation Ontologies (AFO) *via* persistent Uniform Resource Identifiers (URIs).[29] This linkage ensures that each data element corresponds to a globally recognized concept, supporting both semantic harmonization and cross-platform integration.

In cases where a term from the metadata is not found in the Allotrope Dictionary,[11] a new term is constructed using the same principles. Its formal definition is developed based on the best available interpretation within the experimental context. The new term is then introduced using a dedicated namespace such as *cat:NewTerm*, and semantically categorized as either an object or a property. If a unit is relevant, it is defined and assigned accordingly. This approach ensures that even novel or dataset-specific terms are fully integrated into the semantic graph without compromising coherence or interoperability.

In addition to semantic annotation (*i.e.*, explicitly specifying each entity's labels, definitions, examples), the ontology also supports data validation through SHACL[30] (Shapes Constraint Language). SHACL provides a vocabulary and corresponding validation engine for checking whether a set of data (in our case, an RDF-serialized version of metadata representing an experiment), conforms to predefined constraints. These constraints are captured directly in the ontology using SHACL "shapes". Each object and property is associated with a shape that defines the expected structure of the data, for example, its data type, cardinality, or required units. When properly defined, such shapes enable automatic validation of incoming data before integration. This validation layer is particularly important in an open, evolving research infrastructure like Swiss Cat+, where data may originate from diverse instruments, workflows, or external collaborators. SHACL ensures that all integrated data remains structurally consistent, semantically meaningful, and ready for downstream analysis.

2.2 Sample tracking logic

The semantic model described above underpins a detailed system for sample tracking and analytical data integration. This section outlines how identifiers, formats, and ontology relations are applied across the experimental pipeline to maintain traceability and enable semantic reconstruction of experimental histories.

Several key objects act as anchors that tie together data from different stages of the experiment. A *Campaign* represents a high-level experimental objective. Within a campaign, individual batches group samples processed together, each assigned a `batchID`. Individual samples are identified by a unique `sampleID`, constructed from the `containerID` and its physical position (*e.g.*, `containerID-position`). The `sampleID` acts as a persistent, unique identifier that enables tracking from synthesis through to the final stages of characterization.

During the synthesis phase, samples are first generated and labeled using the `batchID` and `sampleID`. At this stage, the system only captures metadata related to sample handling and preparation. This metadata is sufficient to assign each sample a traceable identity and ensures its correct routing into the analytical pipeline. The compound resulting from synthesis is named a product in the Swiss Cat+ ontology. This product is then submitted to further analysis. The term Product is used to distinguish the known samples involved in the synthesis process from its chemical result, which may be either a known or a new compound.

Once the synthesis is complete, the samples (products) transition into the analytical phase. Each analytical platform, represented in the diagram as color-coded nodes (dark blue for ASM-JSON, light blue for JSON, and pink for XML), receives the input `sampleID`. In this phase, each analytical result is augmented by a `peakIdentifier`, which serves to link molecular signatures such as retention times, masses, and spectral features back to the originating sample. This dual system of `sampleID` and `peakIdentifier` ensures both sample-level continuity and molecular-level specificity across techniques. The `peakIdentifier` is an automatically generated Universally Unique Identifier (UUID)[31] produced by the Agilent acquisition software (OpenLab CDS) at peak detection, following the standard 36-character format `8-4-4-4-12` hexadecimal digits. It serves as a globally unique, stable key for unambiguous storage and cross-system tracking of results.

The diagram, Fig. 4, illustrates parallel and sequential analytical workflows. A sample may follow multiple characterization routes. Some routes run in parallel, such as spectroscopic and chromatographic analyses, while others occur in series, such as initial chiral separation (screening path)

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

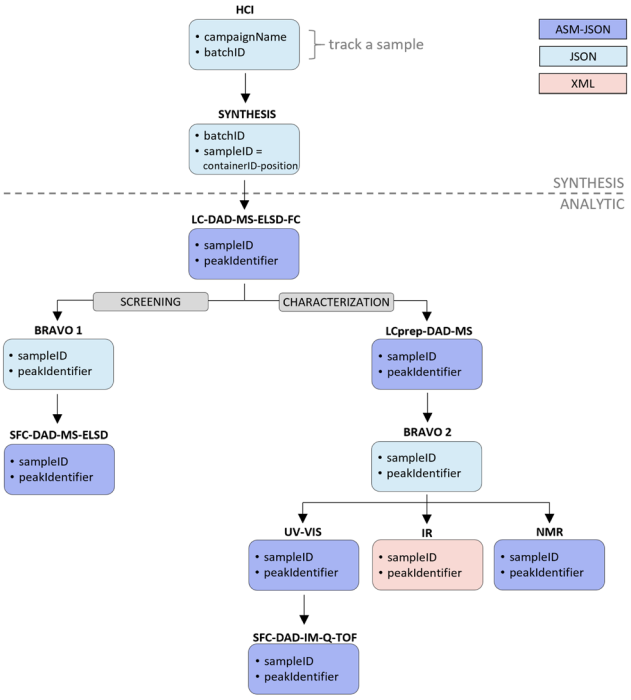


Fig. 4 Sample tracking and data standardization in the Swiss Cat+ platform – samples are tracked from Human–Computer Interface (HCI) and synthesis stages through successive analytical steps, including screening and characterization. Each sample is associated with a unique `sampleID`, derived from its batch and container position, and further linked to molecular-level data *via* `peakIdentifier`. This hierarchical system ensures continuity and traceability across all stages of experimentation. Structured metadata formats (ASM-JSON, JSON, XML) are used throughout the workflow to enable semantic annotation, data integration, and interoperability across heterogeneous platforms.

followed by further purification and reanalysis. Whenever a new product (characterization path) is synthesized or isolated, for instance post-separation or transformation, a new `sampleID` is assigned and linked to its origin. Each analytical block captures results in a standardized format. ASM-JSON is used for structured semantic output, especially where deep ontology alignment is required. Generic JSON is used to enable flexible scriptable data flows for the laboratory-developed instrument (Bravo, HCI). XML is used where instrumentation mandates legacy exports. Two examples in SI S1 and S2, (simple and complex cases) illustrate the tracking of the sample across the process pipeline.

All entities: *Campaigns*, Batches, Samples, Products, and Peaks are explicitly linked through defined ontology properties such as `hasCampaign`, `hasBatch`, `hasSample`, `hasProduct`, `producesProduct`, `preparesProduct` and `hasPeak`. These relations enable the full reconstruction of the experimental trajectory and support semantic search and integration.

2.3 Metadata converter software

While the semantic model defines how metadata should be structured and interpreted, an additional software component is needed to transform raw experimental data into this standardized format. To address this requirement, we developed a generic converter that transforms JSON metadata into RDF statements aligned with the ontology. This converter ensures that all experimental metadata generated throughout the platform is translated into a consistent, machine-readable, and semantically rich format.

To support high-throughput workflows and maintain reliability at scale, the converter was implemented in Rust,[32] a modern compiled language optimized for performance and memory safety. By translating directly into native machine code, Rust provides efficient execution and low overhead, while its strict safety guarantees help prevent runtime errors. These properties make Rust particularly well suited to production environments that require both speed and stability.

The main function of the converter is to map data from JSON metadata into RDF formats, such as Turtle or JSON-LD[33] (JSON-Linked Data), which are languages commonly used for representing structured linked data on the web. Although JSON is a popular format for exchanging data, it lacks semantic context and does not inherently define relationships between entities. In contrast, RDF provides a standardized framework for describing these relationships in a way that can be consistently interpreted by machines. Turtle and JSON-LD are simply two alternative serializations of RDF data. This transformation is enabled through the use of the Sofia crate,[34] a Rust software library that offers the necessary tools to construct RDF output from structured input data. By integrating Sofia into the converter, we ensure compatibility with RDF specifications, while supporting semantically representations of experimental metadata.

The converter is organised into two distinct modules to promote flexibility and reusability. The first module is specific to the Swiss Cat+ project and defines all RDF terms, ontologies, and mapping rules needed to express relationships between data elements according to Swiss Cat+ standards. The second module implements the core conversion logic that processes JSON metadata into RDF representations independently of any project-specific structures. This modular design allows the core converter to be reused in other contexts beyond Swiss Cat+, supporting JSON-to-RDF transformations for different projects simply by supplying an alternative set of semantic definitions. As a result, the tool can adapt to varied domains while maintaining a clear separation between generic processing functions and project-specific configurations.

2.4 Databases & servers

The Swiss Cat+ infrastructure combines scalable storage, automated transformation, and flexible querying to make experimental data both durable and accessible. Data files including metadata and raw experimental outputs in JSON format are deposited at regular intervals are deposited at regular intervals in an internal laboratory database. From there, they are transferred to a dedicated SWITCH S3 database. S3, or simple storage service,[35] is a widely adopted architecture that offers high availability and scalability, ensuring that the growing volume of data can be reliably retained over time.

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Once the metadata arrives in the object store, it is automatically converted to RDF and stored in a dedicated RDF triple store powered by the QLever engine.[36] QLever is optimized for fast, large-scale querying and enables users to interact with the data using SPARQL,[15] a query language specifically designed for RDF. SPARQL allows researchers to extract targeted information by defining patterns of relationships between entities, supporting complex exploration and analysis of experimental metadata. A worked example with six illustrative SPARQL queries is provided in SI S3.

To guarantee that RDF content remains correct and semantically consistent, every RDF file undergoes SHACL validation. SHACL, the Shapes Constraint Language, is a World Wide Web Consortium (W3C) standard for specifying structural constraints on RDF graphs. By validating metadata against SHACL shapes, the system ensures that all RDF content adheres to the expected structure and semantics. A dedicated SHACL server was deployed with its API (Application Programming Interface) to perform these checks automatically as part of the ingestion pipeline.

To maintain reproducibility and resilience, RDF metadata files are stored in two locations: in the QLever triple store for querying and alongside the raw files in S3. The raw files serve as the primary source of truth. Whenever needed, they can be reconverted into RDF Turtle[37] representations using templated workflows, ensuring that metadata can be regenerated consistently as standards evolve.

2.5 Workflow automation: Kubernetes and argo workflows

The deployment of the infrastructure leverages SWITCH Kubernetes-as-a-Service, a cloud-native platform that orchestrates and scales all workflow components reliably. This setup builds on modern containerization and orchestration technologies such as Docker,[38] which packages applications into portable containers, and Kubernetes, which manages and schedules these containers across the cluster. Together, these tools provide a solid foundation for reproducible and portable application execution across different computing environments.

Each software component, including the User Interface (UI) and the metadata converter, is encapsulated within Docker containers. These containers are lightweight, standalone environments that bundle application code together with system libraries and dependencies, ensuring that applications behave consistently regardless of where they are run, be it a developer's laptop, a test server, or a production node. This containerization ensures portability, scalability, and resilience across the infrastructure.

To manage these containers efficiently under dynamic or heavy workloads, Kubernetes acts as the orchestration layer. It automates deployment, scaling, and load balancing. For example, when user demand increases, Kubernetes automatically spawns additional instances (replicas) of the UI container and distributes traffic across them to maintain performance and prevent overloads. Kubernetes also ensures the reliable operation of backend components such as the metadata converter.

In the backend, essential data processing operations are fully automated through Argo Workflows, a Kubernetes-native engine designed to orchestrate complex, multi-step execution pipelines. In Swiss Cat+, Argo Workflows operate on both raw data and metadata to support consistent processing and integration. While raw data remains in its native formats to preserve compatibility with chemists' tools, associated metadata is transformed to enable semantic querying and integration. The complete metadata processing pipeline consists of several Argo Workflows. First, a scheduled Argo CronWorkflow[39] runs weekly to scan the S3-compatible object store for metadata files that have been uploaded or modified within the past seven days. This automated detection eliminates the need for manual tracking of incoming data. Once new or modified files are identified, Argo triggers a containerized instance of the metadata converter, which processes the input (JSON format), into RDF format. Each workflow step begins by launching a container configured with all required tools, ensuring that the code runs in a controlled, consistent environment. The actual transformation is executed through embedded bash scripts, that define the specific sequence of operations for data processing. After RDF transformation, a SHACL validation[40] step is performed to verify that the RDF documents conform to predefined structural and semantic constraints. This validation step ensures the integrity of the metadata quality before database ingestion. If the metadata passes validation, the pipeline proceeds to issue API requests to upload the RDF data into the QLever RDF database. This completes the full cycle of ingestion, transformation, validation, and integration, fully automated and executed without manual oversight. In cases where validation fails, the workflow halts the ingestion process and logs the issue, ensuring that only compliant metadata is stored.

To ensure operational resilience and recovery capabilities, the system includes two additional Argo Workflows specifically designed as backup and safeguard mechanisms. These auxiliary templates can be launched manually by system administrators at any point in time. The "upload-all" workflow is designed to reprocess and re-upload all data in the event of a failure in the S3 object store. The "restore-all" workflow addresses potential issues in the QLever database by re-uploading all previously converted RDF metadata in S3. Together, these workflows enhance system robustness and ensure data integrity throughout the platform.

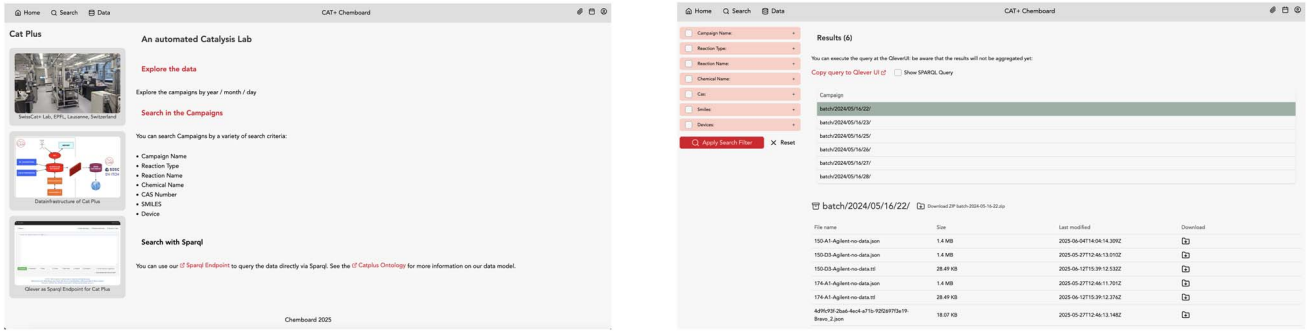
Altogether, this fully containerized and automated infrastructure manages the entire metadata lifecycle: from file ingestion and transformation to SHACL validation and API-driven storage. Each step is executed efficiently and reliably, without requiring manual intervention, ensuring that metadata remains consistent, traceable, and ready for downstream use. This approach decouples metadata processing from raw data storage, allowing the system to scale effectively while preserving clear links between metadata records and their corresponding raw data assets retained in S3.

2.6 User interface for serving the data

To bridge the gap between the system's semantic backbone and the practical needs of researchers with varying technical

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.



(a) User-Interface Landing Page. (b) User-Interface Searching Page.

Fig. 5 User interface of the Swiss Cat+ HT-CHEMBORD platform for accessing experimental data, with (a) landing page and (b) searching page – the interface is designed to facilitate intuitive navigation and semantic querying of experimental metadata. (a) Provides an overview of the system, entry points for search, and direct access to SPARQL documentation and ontology resources. Users can explore experimental campaigns by metadata fields such as campaign name, reaction type, chemical name, CAS number, SMILES, and device. (b) Enables the application of filters, visualization of campaign batches, and downloading of ZIP archives containing structured metadata and data files in JSON and TTL formats. Queries can also be forwarded to the QLever endpoint or viewed as raw SPARQL. A detailed user guide on how to use the interface is available in the SI S4, with Fig. S4.1 detailing the data page and Fig. S4.2 detailing the search page.

backgrounds, a dedicated user interface was developed. Most chemists and experimental researchers are not familiar with RDF, SPARQL, or linked data principles, yet still require practical, approachable tools to find, filter, and export information without learning specialized semantic web technologies. Additionally, many users want to integrate structured metadata into their established workflows without depending on informatics specialists. The interface was therefore designed to make the underlying data accessible and usable by a broad community of scientists. A concise, step-by-step user guide is provided in the SI S4, covering navigation, filtering, and downloads.

The interface supports intuitive browsing, filtering, and downloading of datasets. As illustrated in Fig. 5, users can explore experimental campaigns through an interactive dashboard that offers overviews, entry points for exploration, and direct access to SPARQL documentation and ontology resources (Fig. 5a). Search functionality allows queries by metadata fields such as campaign name, reaction type, chemical name, CAS number, SMILES string, or device. Users can apply semantic filters, visualize campaign batches, and download structured metadata and data files in both JSON and TTL formats (Fig. 5b). For advanced use cases, queries can also be forwarded to the QLever endpoint or inspected as raw SPARQL. By combining these features, the interface enables chemists, data scientists, and AI practitioners to access and reuse the data efficiently, without requiring specialized technical expertise.

2.7 Matryoshka file

To address the challenge of packaging complex, heterogeneous experimental records in a reproducible and interoperable format, the Matryoshka file concept is introduced, as illustrated in Fig. 2. The Matryoshka file is a structured ZIP[41] container designed to encapsulate the complete experimental workflow including metadata, processed data, and final results, within the Swiss Cat+ platform. Inspired by the concept of Russian nesting dolls, this file architecture reflects a layered, hierarchical organization of data. Each element of the workflow, starting from the initial HCI, passing through synthesis, and concluding with analytical results, is embedded in a logical and traceable structure. Each layer represents a stage of the experiment, and together they form a self-contained data entity.

This nested design allows the Matryoshka file to concatenate heterogeneous file formats (*e.g.*, ASM-JSON, XML, JSON, TTL) into a unified archive. The archive can be efficiently stored, exchanged, and interpreted across platforms. This is particularly important for analytical components, where data formats are often not standardized. By maintaining all these elements within a single ZIP file, the Matryoshka file acts as a bridge between different parts of the infrastructure: storage, database, user interface, and workflow orchestration, while minimizing file size and ensuring compatibility.

The ZIP file preserves both positive and negative results from the synthesis and analytical phases. This comprehensive approach not only supports experimental reproducibility, but also provides a reference point for future optimization. Failures are not discarded but conserved alongside successes to inform decision making, hypothesis refinement, and long-term research strategies. In addition, the structure of the Matryoshka file is designed to facilitate training data generation for predictive models. By systematically collecting data, it offers a robust foundation for machine learning applications[42] aimed at improving reaction outcomes or analytical accuracy. This design embodies the platform's broader goal of creating AI-ready, FAIR-compliant datasets that accelerate data-driven discovery in chemistry.

3 Implementation and discussion

The project was guided by the goal of designing solutions that are as chemistry-agnostic as possible, to maximize reusability and impact beyond a single domain. This guiding principle shaped both the converter architecture and the supporting

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

infrastructure, ensuring that core components do not depend on discipline-specific assumptions. As a result, many software aspects of the converter and infrastructure elements can be reused for setting up an RDI for another research community. If another discipline maps its metadata to a compatible semantic ontology to implement a domain-specific converter, then the rest of the infrastructure, including the automated workflows, QLever database, SHACL validation, and user interface, can be reused with minimal adaptation on a separate Kubernetes deployment.

This modular approach demonstrates technical feasibility and sets a precedent within the chemistry community for harmonized data collection and collaborative analysis. To our knowledge, the project is the first of its kind in the domain, offering a reference implementation for large-scale, semantically enriched research data infrastructures.

Finally, the developed ontology offers a foundational framework that can be readily adopted by other laboratories, including those that are not automated, to describe and standardize experimental workflows using a shared semantic vocabulary. By making these tools openly available, the project encourages broader adoption of FAIR principles and supports the development of interoperable datasets essential for advancing data-driven research.

3.1 Standardization of Matryoshka file

The goal of the Matryoshka file is not only to concatenate and preserve the complete set of metadata, experimental and analytical data, as discussed in the Subsection 2.7, but also to progressively reduce file format variability across the workflow. This is a significant challenge, as multiple instruments generate heterogeneous data formats. To address this, alignment with the Allotrope Foundation standards[11] is being pursued, particularly through the adoption of the ASM-JSON format, which offers a consistent, structured, and easily interpretable data representation.

One of the key ongoing efforts is the conversion of legacy file formats. For example, XML files produced by FT-IR instruments, devices manufactured by Bruker, are being converted into ASM-JSON. A converter is being developed to map Bruker's native data structure to the Allotrope Ontology Dictionary, ensuring semantic consistency and reusability of the data. Similarly, NMR data, originally exported in Joint Committee on Atomic and Molecular Physical Data-DX (JCAMP-DX)[43] format, has already been successfully converted to ASM-JSON using publicly available templates[19] provided by the Allotrope Foundation. This conversion step is integrated into the internal database system, so that raw NMR files are automatically standardized prior to storage. The same approach has been applied to other spectroscopic data, UV-Vis, such as files produced in Comma-Separated Values (CSV)[44] format. These files are also mapped and converted into ASM-JSON using corresponding Allotrope templates.

By following this direction, the long-term objective is to create a fully standardized Matryoshka file. The file will be composed exclusively of ASM-JSON files, covering the entire experimental pipeline, from HCI and synthesis to the analytical platform. Such a harmonized structure will improve interoperability, simplify data parsing, and support advanced use cases such as automated reasoning, cross-experiment comparisons, and machine learning integration.

Currently, the Matryoshka structure operates at the batch level, as outlined in Subsection 2.7. It aggregates data from multiple samples, which may correspond to different types of reactions. However, the concept is evolving toward a more modular and hierarchical architecture. In future iterations, each individual sample will be encapsulated as a smaller Matryoshka file nested within the overarching batch-level Matryoshka. This more granular structure will enable faster and more targeted queries, while also facilitating access to data at varying levels of granularity: batch, reaction, or sample, according to the specific demands of complex experimental workflows.

3.2 Challenges and lessons learnt

A central challenge in building the Swiss Cat+ metadata system was the need to identify and track unique objects across multiple experimental campaigns. Chemicals, for example, are first introduced in the HCI phase and later observed in the synthesis phase. Similarly, experimental devices are referenced in synthesis metadata and again during agilent analysis. Without persistent identifiers, relationships between these observations would be fragmented, undermining reproducibility and interoperability. Establishing persistent and unique identifiers was critical to maintaining traceability across these steps.

The system employs a hierarchy of identifiers to connect metadata across the experimental workflow. The Batch identifier links the HCI metadata, where experimental intent is defined, to the synthesis metadata, where that intent is realized. The Product identifier connects the synthesis output to the Agilent analysis data, which characterizes its chemical structure. The Peak identifier bridges the Agilent data with further analysis from the Bravo system and spectroscopy tools such as UV-Vis, NMR, and IR. To guarantee global uniqueness and prevent duplication or ambiguity, the use of URIs or UUIDs is essential. These identifiers serve as unambiguous references to objects within and across datasets, enabling reliable data integration and machine-readability. These globally unique identifiers act as unambiguous references to each object, supporting reliable data integration, reproducibility, and machine-readability across the entire platform.

3.3 Limitations

Despite significant progress, there were limitations in harmonizing the formats of all data files. File formats produced by laboratory instruments, particularly NMR spectrometers, are often proprietary and not designed for ease of parsing. These constraints can limit interoperability, delay integration into the standardized pipeline, and complicate efforts to make data fully FAIR-compliant. The lack of standardization in such files makes

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

automated conversion and metadata extraction extremely difficult.

In the current implementation, specific challenges remain for NMR data in JCAMP-DX format and UV-Vis spectroscopic data exported in CSV format. These formats lack consistent structure and often require manual interpretation or format-specific knowledge, as discussed in the Subsection 3.1. To integrate these data types into the standardized ASM-JSON framework, custom Python scripts are being developed to extract and structure relevant metadata according to the Allotrope ontology. This task is complex and time-intensive, as it involves parsing free-text fields, handling inconsistent formatting across instruments, and ensuring semantic alignment with the target schema. While progress continues, full harmonization remains a work in progress and will require sustained development and community collaboration to achieve complete coverage of all experimental data types.

3.4 Future enhancements

To further strengthen interoperability, scalability, and support for AI-driven research, several enhancements are planned for the Swiss Cat+ RDI. A summary of these proposed developments is provided in Fig. 6. One potential direction is to manage new processed data using the Zarr format. If the original data was manipulated in data analysis ("processed") then it could be stored again onto the Swiss Cat+ database, alongside the original raw data. In this design, the processed data would be stored, alongside its metadata, in the Zarr format.[45] Zarr supports metadata embedding and chunked data storage, making it suitable for cloud-based analysis.

An additional improvement to the metadata model would be to incorporate labels for the peaks identified in a product, especially if this is a new compound. At the end of the analysis pipeline, the Swiss Cat+ system identifies known substances or discovers new chemical entities. Similarly, the detected peaks correspond to known molecular structures. Currently, however, this identification information is not fed back into the raw ASM-JSON data and the metadata database. Including these labels would greatly enhance the utility of the data for AI applications, as labeled data is essential for training supervised learning models. Making such annotations available would significantly enrich the resource for the broader research community.

To improve practical reusability for chemists while avoiding exposure to linked-data internals, the CHEMBORD interface

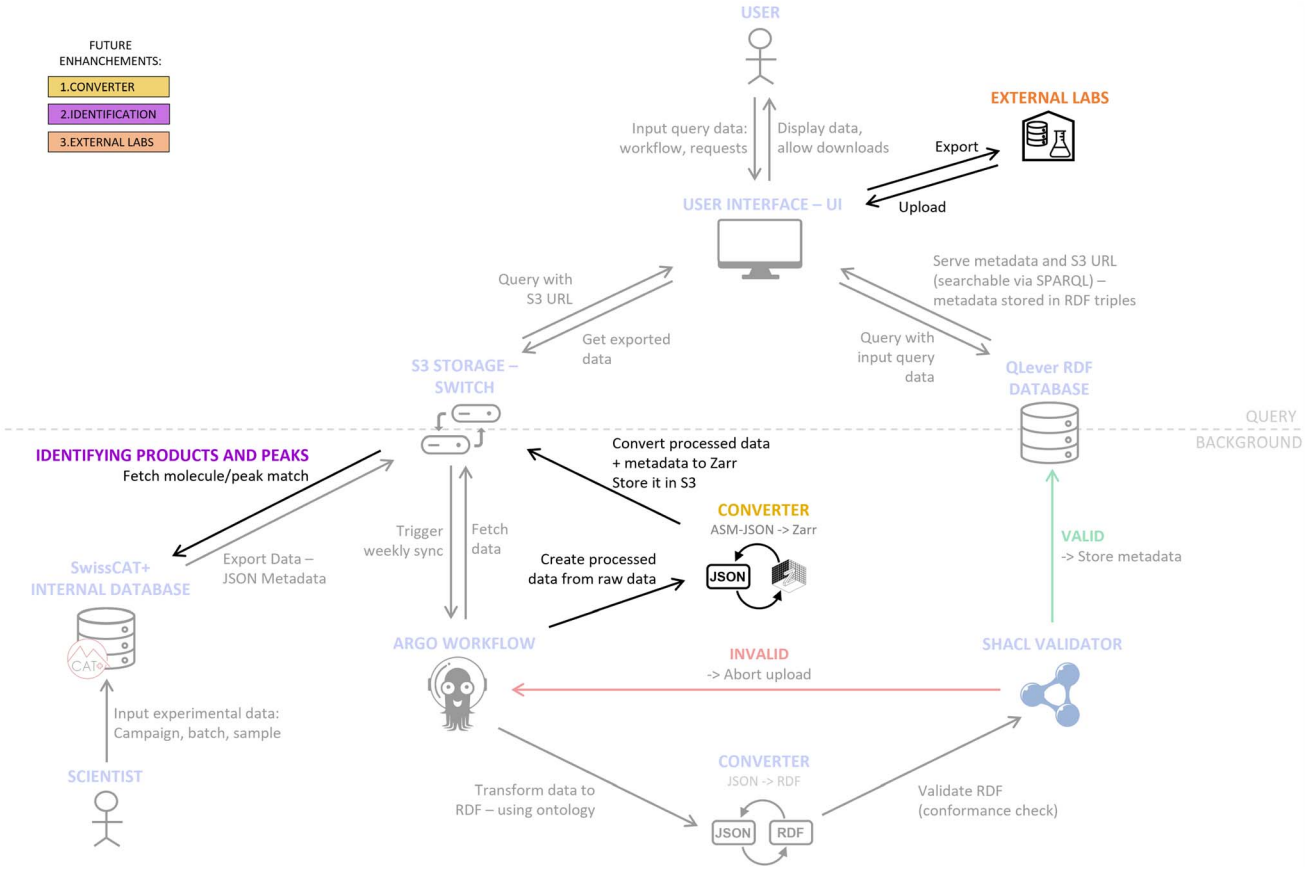


Fig. 6 Integrated data management and future enhancement workflow for the Swiss Cat+ RDI, including (1) converter (yellow), (2) identification (purple), and (3) external labs (orange) – planned developments focus on three main areas: (1) save processed data and its metadata in Zarr format (2) enriching RDF data with semantic labels for products and peaks to enhance reusability and support AI applications; and (3) enabling external laboratory contributions through structured metadata entry, automatic RDF generation, and SHACL validation to ensure semantic integrity and data quality.

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

will be evaluated on a large, heterogeneous test set. This aligns with the first project phase, which builds the end-to-end data pipeline and data model. The goal is to compile large, high-quality datasets for large-scale analysis and algorithm training. The test will span multiple campaign types and involve a broad cohort of practicing chemists. This evaluation goes well beyond the current single-molecule demonstration. It will be used to systematically identify usability issues, prioritize points to improve or change, and guide a targeted second iteration under realistic data volume and diversity. In parallel, a downloadable report folder is being developed within the laboratory (as part of a separate project) and is intended to be linkable from CHEMBORD. For each campaign, this folder will provide per-instrument data files in CSV format (*e.g.*, LC, SFC, IR, UV-Vis, NMR). It will also include a report that summarizes campaign context (including synthesis information), presents consolidated analytical plots, and reports key outcomes such as isolated yield and when applicable, enantiomeric excess. This resource will enable users to download a ready-to-use, human-readable package directly from the CHEMBORD UI.

In addition, RO-Crate (Research Object Crate)[46] packaging could be proposed as a unified request option within the user interface, allowing users to conveniently export or share complete data and metadata packages. If this feature is widely adopted, it could later be integrated as a permanent backend function. This addition would further enhance interoperability and reproducibility by providing a standardized, machine-readable data exchange format consistent with FAIR principles.

Finally, the most important enhancement involves enabling users to contribute their own data. The ontology enables the expansion of the system into a collaborative data hub. New laboratories-automated or not-can contribute their experimental data to the shared S3 storage. The UI could provide metadata entry forms specific to each data file type. The forms would assist users in entering the required metadata and automatically generate RDF representations in the background. All user-submitted metadata would undergo SHACL validation before being added to the database to ensure the metadata is conform to the ontology definitions. While metadata quality and interoperability can be assessed by the current RDI, an additional system would need to be developed by the chemist experts for evaluating incoming data quality and interoperability, so as to ensure that the overall database maintains quality standards. This will help ensure that the overall database continues to meet high standards of reliability and scientific utility.

4 Summary

The Swiss Cat+ RDI came into being from a fruitful collaboration of chemistry domain experts and data infrastructure and research software engineers. The chemistry team focused on data format standardization and comprehensive mapping of the field's concepts into an ontology. The engineering team contributed semantic expertise and implemented the infrastructure components required to serve the data and metadata to the broader research community.

The data format standardization is a clear illustration of the need for simple and interoperable data formats for doing data science, as well as the considerable effort required for making changes in data practices. The established RDI shows how open-source technologies combined with a strong infrastructure provider such as SWITCH empower research communities towards data FAIRness. By systematically capturing both positive and negative results in machine-readable formats, this infrastructure also lays the groundwork for reproducible AI models and collaborative discovery across laboratories. Together, these advances provide a reference implementation that can inspire and guide future initiatives aimed at building open, interoperable research data ecosystems across disciplines.

Author contributions

Conceptualization: J.-C. C., A. G., L. V., C. M.-D., R. F., S. M., P. M., O. R. G. Funding acquisition: P. M., O. R. G. Investigation: J.-C. C., A. G., L. V., C. M.-D., R. F., S. M. Methodology: J.-C. C., A. G., L. V., C. M.-D., R. F., S. M. Project administration: J.-C. C., L. V. Resources: A. G., J.-C. C. Software: J.-C. C., L. V., C. M.-D., R. F., S. M. Supervision: J.-C. C., L. V. Validation: J.-C. C., L. V., O. R. G. Visualization: A. G. Writing (original draft): A. G., L. V. Writing (review and editing): J.-C. C., A. G., L. V., C. M.-D., R. F., P. M., O. R. G.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data supporting this study are openly available. The curated dataset used in this work is archived at DOI: **https://doi.org/10.5281/zenodo.17260302**. Data can also be browsed or downloaded *via* the CHEMBORD platform at **https://catplus.swisscustodian.ch/chembord/**(DOI: **https://doi.org/10.5281/zenodo.17226128**). The source code is hosted on GitHub: documentation repository, *catplus-docs*: **https://github.com/swisscatplus/catplus-docs** (DOI: **https://doi.org/10.5281/zenodo.17225715**), front-end repository, *catplus-chembord*: **https://github.com/swisscatplus/catplus-chembord** (DOI: **https://doi.org/10.5281/zenodo.17225943**), converter repository, *catplus-converters*: **https://github.com/swisscatplus/catplus-converters** (DOI: **https://doi.org/10.5281/zenodo.17226021**), ontology repository, *catplus-ontology*: **https://github.com/swisscatplus/catplus-ontology** (DOI: **https://doi.org/10.5281/zenodo.17226026**). Ontology documentation is accessible at (index of all terms used with their definitions, as well as a graph representation of all interlinked objects): **https://sdsc-ordes.github.io/catplus-ontology/** (DOI: **https://doi.org/10.5281/zenodo.17226173**). Kubernetes deployment manifests (*catplus-manifests*) are available upon request (contact authors for access).

Supplementary information: further details on two applied workflow examples illustrating sample tracking across the Swiss

 © 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Cat+ platform (Fig. S1–S2), a set of examples of SPARQL queries used by CHEMBORD for metadata retrieval (Fig. S3), and a step-by-step procedure for navigating the CHEMBORD UI (Fig. S4.1–S4.2), are provided in the supplementary information (SI). See DOI: https://doi.org/10.1039/d5dd00297d.

Acknowledgements

This project was supported by the Open Research Data Program of the ETH Board.

References

1 P. Laveille, P. Miéville, S. Chatterjee, E. Clerc, J.-C. Cousty, F. de Nanteuil, E. Lam, E. Mariano, A. Ramirez, U. Randrianarisoa, K. Villat, C. Copéret and N. Cramer, Swiss CAT+, a Data-driven Infrastructure for Accelerated Catalysts Discovery and Optimization, *Chimia*, 2023, **77**(3), 154–158, DOI: 10.2533/chimia.2023.154.

2 P. Miéville and F. de. Nanteuil, Modern Automation in Organic Synthesis Laboratories, in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*, Elsevier, 2024, DOI: 10.1016/B978-0-323-96025-0.00047-8.

3 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021, **596**, 583–589, DOI: 10.1038/s41586-021-03819-2.

4 W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke and B. A. Grzybowski, Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling, *J. Am. Chem. Soc.*, 2022, **144**, 4819–4827, DOI: 10.1021/jacs.1c12005.

5 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, Negative Data in Data Sets for Machine Learning Training, *Org. Lett.*, 2023, **25**, 2945–2947, DOI: 10.1021/acs.orglett.3c01282.

6 T. Taniike and K. Takahashi, The value of negative results in data-driven catalysis research, *Nat. Catal.*, 2023, **6**, 108–111, DOI: 10.1038/s41929-023-00920-9.

7 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, The Open Reaction Database, *J. Am. Chem. Soc.*, 2021, **143**(45), 18820–18826, DOI: 10.1021/jacs.1c09820.

8 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 2016, **3**, 160018, DOI: 10.1038/sdata.2016.18.

9 P.-C. Huang, C.-L. Lin, P. Tremouilhac, N. Jung and S. Bräse, Using the Chemotion repository to deposit and access FAIR research data for chemistry experiments, *Nat. Protoc.*, 2025, **20**(5), 1097–1098, DOI: 10.1038/s41596-024-01074-z.

10 N. F. Noy, Semantic integration: a survey of ontology-based approaches, *ACM SIGMOD Rec.*, 2004, **33**(4), 65–70, DOI: 10.1145/1041410.1041421.

11 H. Oberkampf, H. Krieg, C. Senger, T. Weber, C. Wolfgang, *Allotrope Data Format – Semantic Data Management in Life Sciences*, Semantic Web Applications and Tools for Healthcare and Life Sciences, 2018, DOI: 10.6084/m9.figshare.7346489.v1.

12 Switch Kubernetes-as-a-Service (KaaS), 2025, https://www.switch.ch/en/kaas, accessed 24.06.2025.

13 Kubernetes – Production-Grade Container Orchestration, https://kubernetes.io/, accessed 24.06.2025.

14 World Wide Web Consortium (W3C), G. Klyne and J. Carroll, Resource description framework (RDF): concepts and abstract synta, 2023, https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, accessed 15.06.2025.

15 World Wide Web Consortium (W3C), *SPARQL 1.1 Query Language*, 2023, https://www.w3.org/TR/sparql11-query/, accessed 15.06.2025.

16 GitHub, Workflow Engine for Kubernetes – What is Argo Workflows?, https://github.com/argoproj/argo-workflows, accessed 17.06.2025.

17 Agilent Instruments – LC, GC, SFC, UV-Vis, https://www.agilent.com/, accessed 24.06.2025.

18 Bruker Instruments – FT-IR, NMR, https://www.bruker.com/en.html, accessed 24.06.2025.

19 GitLab – Allotrope Foundation Members, Allotrope Simple Models, https://gitlab.com/allotrope-public/asm/-/tree/main?ref_type=heads, accessed 15.06.2025.

20 World Wide Web Consortium (W3C), T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler and F. Yergeau, *Extensible Markup Language (XML) 1.0*, 5th edn, 2008, https://www.w3.org/TR/xml/, accessed 15.06.2025.

21 T. F. Waddell, B. Zhang and S. S. Sundar, in *Human–Computer Interaction*, John Wiley & Sons, Ltd, 2015, DOI: 10.1002/9781118540190.wbeic182.

22 Chemspeed Insstruments – Swing XL, https://www.chemspeed.com/example-solutions/swing/, accessed 24.06.2025.

23 Arksuite – Workflow management software for automated systems, https://www.chemspeed.com/software-and-digitalization/arksuite/, accessed 24.06.2025.

24 R. Kumbhar, in *8 – Modern knowledge organisation systems and interoperability*, Chandos Publishing, 2012, DOI: 10.1016/B978-1-84334-660-9.50008-4.

25 World Wide Web Consortium (W3C) and D. Beckett, *RDF 1.1 N-Triples - A line-based syntax for an RDF graph*, 2014, https://www.w3.org/TR/n-triples/, accessed 25.06.2025.

26 World Wide Web Consortium (W3C), G. Schreiber and Y. Raimond, *RDF 1.1 Primer*, 2014, https://www.w3.org/TR/rdf-primer/, accessed 07.07.2025.

27 Z. G. Ives, A. Y. Halevy, P. Mork and I. Tatarinov, Piazza: mediation and integration infrastructure for Semantic Web data, *J. Web Semant.*, 2004, DOI: 10.1016/j.websem.2003.11.003.

Open Access Article. Published on 03 November 2025. Downloaded on 1/12/2026 5:54:15 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

28 GitHub, *QUDT – Quantities, Units, Dimensions and DataTypes*, 2022, **https://github.com/qudt/qudt-public-repo**, accessed 15.06.2025.
29 E. F. Duranceau, Naming and describing networked electronic resources: The role of uniform resource identifiers, *Ser. Rev.*, 1994, **20**(4), 31–51, DOI: **10.1016/0098-7913(94)90018-3**.
30 World Wide Web Consortium (W3C), *Shapes Constraint Language (SHACL)*, 2017, **https://www.w3.org/TR/shacl/**, accessed 20.06.2025.
31 P. Leach, Microsoft, M. Mealling, Refactored Networks, LLC, R. Salz and DataPower Technology, Inc., *A Universally Unique IDentifier (UUID) URN Namespace*, 2005, DOI: **10.17487/RFC4122**, accessed 30.09.2025.
32 GitHub, *The Rust Programming Language*, **https://github.com/rust-lang/book**, accessed 25.06.2025.
33 World Wide Web Consortium (W3C), M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, P.-A. Champin and N. Lindström, *JSON-LD 1.1 – A JSON-based Serialization for Linked Data*, 2020, **https://www.w3.org/TR/json-ld11/**, accessed 15.06.2025.
34 Sofia crates – A Rust toolokit for RDF and Linked Data (LD), **https://crates.io/crates/sophia**, accessed 25.06.2025.
35 M. Mesnier, G. Ganger and E. Riedel, Object-based storage, *IEEE Commun. Mag.*, 2003, **41**, 84–90, DOI: **10.1109/MCOM.2003.1222722**.
36 H. Bast and B. Buchhold, in *QLever: A Query Engine for Efficient SPARQL+Text Search*, Association for Computing Machinery, New York, NY, USA, 2017, DOI: **10.1145/3132847.3132921**.
37 World Wide Web Consortium (W3C), D. Beckett, T. Berners-Lee, E. Prud'hommeaux and G. Carothers, *RDF 1.2 Turtle*, 2025, **https://www.w3.org/TR/rdf12-turtle/**, accessed 17.06.2025.
38 Docker: Accelerated Container Application Development, **https://www.docker.com/**, accessed 24.06.2025.
39 Argo Cron Workflows, **https://argo-workflows.readthedocs.io/en/latest/cron-workflows/**, accessed 25.06.2025.
40 S. D. S. C. GitHub, *A Github action performing SHACL validation on turtle files in Github repositories*, **https://github.com/sdsc-ordes/shacl-validation-action**, accessed 25.06.2025.
41 World Wide Web Consortium (W3C), *Lightweight Packaging Format (LPF)*, 2020, **https://www.w3.org/TR/lpf/**, accessed 15.06.2025.
42 A. Toniato, A. C. Vaucher, T. Laino and M. Graziani, Negative chemical data boosts language models in reaction outcome prediction, *Sci. Adv.*, 2025, **11**, DOI: **10.1126/sciadv.adt5578**.
43 IUPAC – International Union of Pure and Applied Chemistry, JCAMP-DX, **https://iupac.org/what-we-do/digital-standards/jcamp-dx/**, accessed 15.06.2025.
44 World Wide Web Consortium (W3C), *CSV on the Web: A Primer*, 2016, **https://www.w3.org/TR/tabular-data-primer/**, accessed 15.06.2025.
45 Zarr – chunked, compressed, N-dimensional arrays, **https://zarr.dev**, accessed 17.06.2025.
46 University of Technology Sydney, *The University of Manchester UK and RO-Crate contributors, Research Object Crate (RO-Crate)*, 2021, **https://www.researchobject.org/ro-crate/**, accessed 12.10.2025.

 © 2025 The Author(s). Published by the Royal Society of Chemistry