Quantum Data Hub: A Collaborative Data and Analysis Platform for Quantum Material Science

Shweta Purawat[1][0000−0002−5183−2750], Subhasis Dasgupta[1][0000−0002−0754−0515], Luke Burbidge[1], Julia L. Zuo[2], Stephen D. Wilson[2][0000−0003−3733−930X], Amarnath Gupta[1][0000−0003−0897−120X], and Ilkay Altintas[1][0000−0002−2196−0305]

[1] University of California San Diego, La Jolla, CA, USA
shpurawat,sudasgupta,lburbidge,a1gupta,ialtintas@ucsd.edu
[2] University of California Santa Barbara, Santa Barbara, CA, USA
jlzu,stephendwilson@ucsb.edu

Abstract. Quantum materials research is a rapidly growing domain of materials research, seeking novel compounds whose electronic properties are born from the uniquely quantum aspects of their constituent electrons. The data from this rapidly evolving area of quantum materials requires a new community-driven approach for collaboration and sharing the data from the end-to-end quantum material process. This paper describes the quantum material science process in the NSF Quantum Foundry with an overarching example, and introduces the Quantum Data Hub, a platform to amplify the value of the Foundry data through data science and facilitation of: (*i*) storing and parsing the metadata that exposes programmatic access to the quantum material research lifecycle; (*ii*) FAIR data search and access interfaces; (*iii*) collaborative analysis using Jupyter Hub on top of scalable cyberinfrastructure resources; and (*iv*) web-based workflow management to log the metadata for the material synthesis and experimentation process.

Keywords: Quantum Material Science · FAIR · Data Management · Collaboration Platform · JupyterHub

1 Introduction

Quantum materials research is a rapidly growing domain of materials research, seeking for novel compounds whose electronic properties are born from the uniquely quantum aspects of their constituent electrons. Electronic states whose order can be defined locally, such as superconductivity and collective magnetism, emerge in quantum materials as well as electronic states forming non-local order, such as topologically nontrivial band structures and many-body entangled states. These and other states are sought to form the basis of the coming revolution in quantum-based electronics and can allow quantum information to be harnessed for next-generation computing and sensing applications.

Although there are material research data facilities built around generation, ingestion and sharing of data, the data from this rapidly evolving area of quantum materials require a community-driven approach for collaboration and sharing the

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

data from the end-to-end quantum material process. It is critical to establish a community network and a collaborative data management and analytical ecosystem to couple data to theory to materials development and to complement the growing number of theory-forward materials prediction databases in the field.

UC Santa Barbara's NSF Quantum Foundry [3], funded by the National Science Foundation, is a next generation materials foundry that develops materials and interfaces hosting the coherent quantum states needed to power the coming age of quantum-based electronics. Its mission is to develop materials hosting unprecedented quantum coherence, train the next generation quantum workforce, and to partner with industry to accelerate the development of quantum technologies.

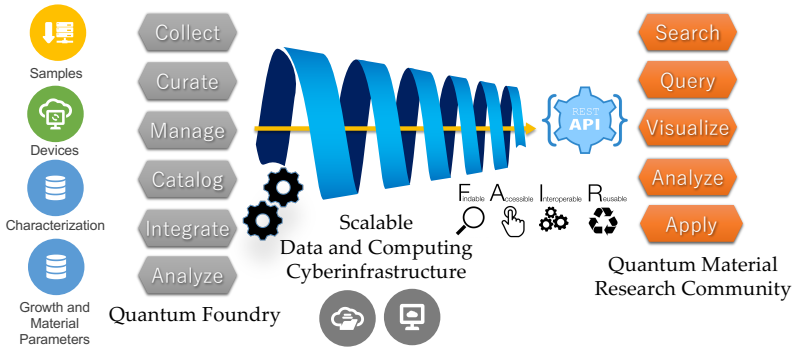


Fig. 1: The conceptual data pipeline for the Quantum Data Hub.

This paper describes the quantum material science process with an overarching example from the Quantum Foundry, and introduces the Quantum Data Hub (QDH), a platform to amplify the value of the Foundry data through data science. As depicted in Figure 1, the QDH collects, curates and manages the experimental and theoretical scientific data, often not in a searchable and queriable form until now. The data, which includes large amounts of physical objects (e.g. samples/devices), characterization data, and growth and material parameters serves as a backbone of our efforts, to aid the data-driven discovery and development of new materials with engineered functionalities. The data is made readily searchable for analysis using advanced cyberinfrastructure tools for automated workflows, machine learning, and statistical analysis. The QDH enables data cleaning to provide users with FAIR [19] views over it. The data collected is served through RESTful APIs that can serve the raw, cleaned and versions of analytical data products in a scalable fashion. This scalable approach allows for the simultaneous availability of such data to many processing modules.

Contributions. In this paper, we present the data and analysis components of the Quantum Data Hub to facilitate: (*i*) storing and parsing the metadata that exposes programmatic access to the quantum material research lifecycle involving experimentation and synthesis; (*ii*) FAIR data search and access interfaces with access control; (*iii*) collaborative analysis using Jupyter Hub on top of dynamic cyberinfrastructure resources; and (*iv*) web-based workflow management to log

[3] Quantum Foundry Website: https://quantumfoundry.ucsb.edu/

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

the metadata for the material synthesis and experimentation process. We also present a case study for powder synthesis and measurement process.

Outline. The rest of this paper is organized as follows. In Section 2, we describe the quantum material research process and a case study for powder synthesis. Section 3 introduces the Quantum Data Hub architecture and its main components. We review related work in Section 4 and conclude in Section 5.

2 Quantum Material Research Process and Data Model

2.1 Quantum Material Research Process

The past 15 years have witnessed a revolution in the computational modeling and theoretical prediction of quantum materials with tailored electronic properties. Experimental assessment of these predictions however proceeds at a much slower pace due to the bottleneck of the laborious and often iterative process of experimentally synthesizing newly predicted materials. Due to the difficulty in predicting and modeling inorganic reaction pathways, diffusion, and grain growth at elevated high temperatures, the materials growth synthesis process in the quantum materials domain remains dominated by chemically informed starting points followed by onerous trial and error iteration.

The research process itself starts with a prediction of a new material with desired functional properties. This is followed by developing a plan to synthesize the new compound, and starting points are typically chosen based off of a researcher's prior experience synthesizing related materials or via reported synthesis conditions of similar compounds. A starting point for the reaction conditions/processing space is chosen which involves the choice of the starting reagents, a thermal profile for reacting the reagents, and the choice of the correct processing space for the reaction to occur (e.g., what gas environment should be used; what type of furnace/heating source; what type of reaction vessel; etc.).

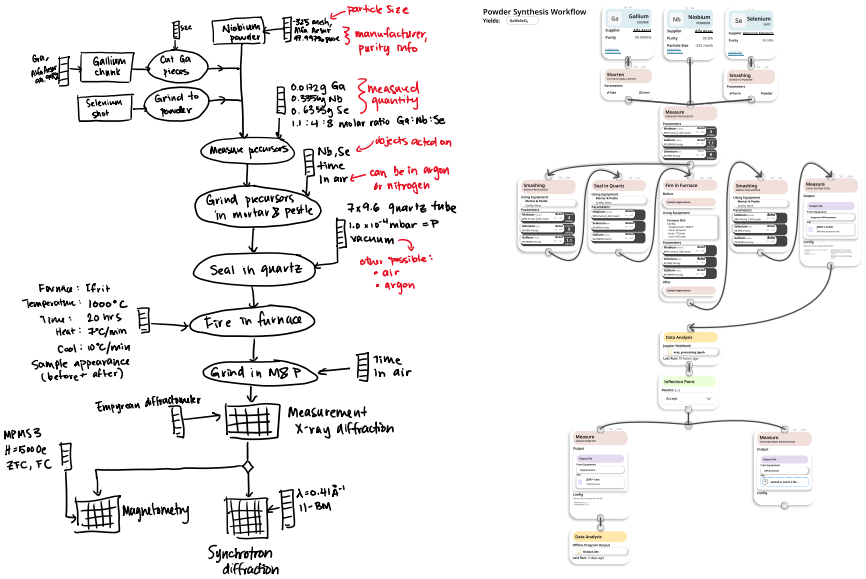
Once the initial conditions are chosen, the experiment is executed and then the product is analyzed via a number of experimental probes to ascertain what material was created. This typically involves x-ray structural analysis, various forms of chemical fingerprinting such as energy dispersive spectroscopy, and composition analysis via electron microscopy. Once the composition of the created material is ascertained, then the original conditions of the reaction are modified to push toward the reaction toward the desired result. In quantum materials research, the most common goal is to create a high purity, single phase sample of the desired compound for follow on study.

Once the desired compound is created with the requisite purity, then the electronic properties of the compound are explored via a number of complementary probes. This can include bulk measurements of electrical resistivity, the magnetic susceptibility, heat capacity, more advanced characterization with optical spectoscopy, angle-resolved photoemission, and scanning tunneling microscopy. Depending on the hypothesis being tested and the experiments needed, the form factor for a sample may need to be a macroscopic, single crystal of the new compound, rather than a multigrain powder (a collection of microscopic crystallites). Developing the necessary parameters for achieving crystal growth of a given compound then requires further experiment design and iterative growth/testing steps.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

We also note here that quantum materials are also heavily explored in thin film form, which entails additional complexities (substrate type, growth orientation, etc.) beyond the broad overview of "bulk" materials synthesis detailed above.

Once a high enough purity sample is created in the appropriate form factor (e.g., power, single crystal, thin film), the lifecycle of experimental exploration can be long. Measurement by multiple complementary probes is common, and many materials are tuned chemically following their initial measurement in order to test new hypothesis formed from the characterization data. Changing the composition of the starting material to address these hypothesis begins the iterative synthesis process again, which feeds into the *theory*, *synthesis*, *characterization* loop. The synthesis step in this loop is a major bottleneck for the field.



(a) (b)

Fig. 2: Powder synthesis and measurement process for the chemical $GaNb_4Se_8$: (a) An excerpt from a lab notebook outlining the process. (b)The Quantum Data Hub representation of the workflow for the process outlined in the lab notebook.

2.2 Example: Powder Synthesis and Measurement Process

An example use case of the synthesis and measurement of the chemical $GaNb_4Se_8$ is illustrated in Figure 2. The preparation of $GaNb_4Se_8$ involves several steps involving multiple instruments and processes before the figure of merit of the sample, its chemical purity, is confirmed using X-ray powder diffraction. This measurement step is common to many solid-state chemical reactions and serves as a node where the user then decides whether to proceed with other measurement processes, represented by magnetometry and synchrotron diffraction in Figure 2. The Quantum Data Hub (QDH) provides a number of features to make such an analysis simpler and streamlined within the quantum material research process.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

Ease of Integrated Data Access and Analysis. Throughout the synthesis and measurement workflow, all metadata and measurement results can be uploaded from any laboratory with a computer and an internet connection. The QDH provides a highly portable analysis environment for measurement data such as magnetometry. Typically, such magnetic data would be transferred from the instrument to the user's computer, where they perform analysis with a variety of methods including user-written Python scripts. On the QDH, the data and user-written Python analysis libraries can be uploaded and run anywhere, greatly streamlining the synthesis and measurement workflow. The QDH also allows the user to access information about their samples and data analysis online.

Capturing Reusable Synthesis Conditions. Quantum materials, as well as materials science, has historically relied on published journal articles for reported materials synthesis conditions. There have been a number of efforts to use published synthesis conditions to datamine and machine learn material synthesis to remove the synthesis bottleneck in materials research. However, researchers lose vital information in the publication process. Failed synthesis conditions provide invaluable data points for exploring a new chemical phase space but typically only successful synthesis conditions are published. The Quantum Data Hub retains failed synthesis conditions with query-able metadata associated with the material and its synthesis processes. By creating a database of materials and their synthesis conditions, the QDH can serve as a highly organized and useful dataset for data science efforts in quantum materials science.

Extracting Physical Properties. Other limitations in the field of quantum materials research include compilations of measured physical properties of candidate quantum materials. In fact, compiling data exhibiting the physical properties of novel materials presents many additional challenges including the same challenges of compiling published synthesis conditions. Measured properties are seldom published in raw data form. Instead, researchers present measured data in a variety of formats, including plots, tables, and other graphics. Additionally, data published in journals are often processed to exhibit certain features of the magnetic properties of a material, making it difficult for automated data extraction. The QDH automatically creates a searchable database of physical properties that are important to quantum materials as raw data, retaining a maximal amount of information and unpublished materials property measurements.

One such use case involves the magnetic properties of materials. Magnetic properties are very important to many quantum materials. Magnetometry measurements are common as an initial characterization method of a novel material. There are a variety of different descriptors of magnetic properties, including ordering temperatures, but in quantum materials research, often the most important magnetic features are qualitative, such as the general shape of the magnetization as a function of temperature or applied magnetic field. This is particularly true for quantum materials that are of interest to leverage exotic magnetic properties. Existing databases of experimental magnetic properties focus on limited descriptors such as ordering temperatures and common magnetic properties such as ferromagnetism and antiferromagnetism. By design, a lot of experimental information is lost in these databases. However, if researchers had access to raw

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

```
"processSteps": [
 {"@id": "_:b1",
 "processName": "http://sweetontology.net/procPhysical/Shorten",
 "processParameters": [
    {   "cuttingSizeValue": 20,
        "cuttingSizeUnit": "mm" } ],
  "http://rdf.data-vocabulary.org/#description":
                        "cut into small pieces",
  "http://www.loa-cnr.it/ontologies/
                        FunctionalParticipation#patient": 102,
  "next_steps": [ "_b4" ]},
  ...             ]
```

Fig. 3: Every process step, expressed as a semi-structured node, has its own ID. The `next_steps` element denotes a list of edges from the current node to other steps. The attributes of a data object may come from established ontologies.

magnetic data, initial assessments of the magnetic properties of a material are incredibly quick to a trained eye. The QDH will enable users to search through and quickly evaluate as-measured magnetic data in the form that they choose.

The QDH allows searching through materials and their properties, enabling users to quickly assess the magnetic properties of materials with all the associated metadata of the measurement and material. This will accelerate the initial bottleneck of selecting materials candidates and synthesizing them as well as enabling data science initiatives in quantum materials research to connect materials descriptors such as chemistry or atomic structure to novel physical properties.

2.3 The Quantum Foundry Data Model

The Quantum Foundry Data Model (QFDM) builds on the premise that the data activities of the Foundry is centered around scientific processes and their products. The processes include the synthesis of new materials, taking a newly synthesized material through a series of instruments and computations to measure complex properties, recording these measurements and computational results, evaluation of these results, publishing the results in scientific venues, and possibly using the products of one synthesis process as the raw ingredients of another synthesis process. The data model captures the essential descriptions and order of these processes, as well as all artifacts produced at different stages of these processes.

Formally, the QFDM is a *federated heterogeneous data model*, which is a multi-part model, each part expressed with a different modeling language and implemented in a different store, yet schematically connected through explicit references (foreign keys). The data model is stored in a polystore based information management system called AWESOME [9] developed at UC San Diego.

1. **Process Model (Semistructured – DAG).** The objective for designing the process model is to enable new scientists find previous material synthesis experiments based on ingredients, instruments, experimental results and subprocesses that might possibly be reused. The process model takes the structure of a directed acyclic graph (DAG) where nodes represent subprocesses, and

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

edges designate a direct transition from one subprocess to the next. The nodes of the graph are typed, semistructured objects implemented as JSON-LD so that one element can reference another element within the same process or to an external object through a hyperlink. Figure 3 shows a node of the Process Model DAG. A schematic of the full process DAG is shown in Figure 2. Partially inspired by [17], the process DAG illustrates the following features.

- A process node may belong in one of many system-defined types, e.g., a mechanical process, a chemical process, a computational process, etc. Each process may have subcategories. For example, gas flow synthesis, spark plasma sintering and annealing are chemical processes.
- For each process type, there are a set of mandatory metadata attributes. For example, a mechanical or a chemical process must record the *environment* in which the process occurs. `grinding` might be performed in open air and another may require an inert gas environment at a prescribed pressure. Similarly, a measurement process must specify the measuring instrument and must point to the measurement settings.
- A node attribute may have external references. There are two kinds of references. A *URI reference* is used to point to other information objects like a PubChem entries, while a *data reference* points to a measurement item or a computational item within our system.

2. **Measurements Model (Relational/Semistructured).** Measurements are primarily outputs of different measuring devices or from computational processes. Figure 2 shows three measurement nodes, namely, X-ray diffraction, Synchrotron diffraction and magnetometry. These nodes point to data files whose formats may be relational or semistructured (XML). In either case, the measurement data has a "settings" component and a "measured values" component. Since one of our goals is to find materials synthesis processes that use similar measurement settings, we maintain both the original and flattened versions of the settings data. The "measured values" component is stored to be primarily consumed by analysis routines, and is transformed to a relational form for querying as well as to a form that the analysis routines expect. When measurements are produced from computational processes, the "settings" component contains the identity of the corresponding computational process and the parameters of its execution.
3. **Computation as Data (Semistructured/Vector).** The final component of the data model are computations that analyze data. These computations can be in a *black box* or *gray box* mode. A proprietary analysis software is considered a black box, while an accessible computation, performed through a Jupyter Notebook, is a gray box because parts of the notebook, including documentation, are expressed in an interpretable form (JSON) and can be analyzed algorithmically, while other parts, like the inner details of libraries called in a notebook cannot. As mentioned before, only the invocation information can be stored for black box computations. Gray box computations, provide significantly more details including (a) a vector of libraries used, (b) document vectors comprising all commentaries, and (c) a list of output items

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

that are stored externally. These three vectors are preserved and can be used in finding similar processes in downstream analysis.

The above description illustrates the cross-pointers between different parts of the data model, allowing us to query for materials synthesis processes through any of the stored parameters and then navigating to find all related information stored in other parts of the AWESOME system, which is designed to store relational, semistructured, graph and text-centric data.

The operations supported by the QFDM are developed based on the intended use. Typically, the actual data values, like an X-Ray diffraction measurement value at a specific angle, are not queried for. Rather, the whole measurement data is consumed by a computation or a visualization process. Similarly, an image produced through an experiment is not queried through content analysis.

At this point, QFDM operations are being designed to support **sample-based queries** and **process-based queries**. In one type of sample-based query, the user knows (or can query the system to determine) the sample, and retrieves a data product derived from the sample by measurement or computation. A more general type of sample-based query locates samples for which general process parameters are specified. For example, "which samples were subjected to the X-Ray diffraction method but did not get characterized based on magentometry"? Yet as third category of sample-based queries would be on the characterization and experimental settings of the synthesized material. For example, in the use case described in Section 2.2 the query can be stated as "Find all samples for which, tin (Sn) is a component material, Single Crystal Neutron Diffraction and Electron Paramagnetic Resonance were measured and Field-dependent magnetization data were collected at a temperature below 5K with magnetic field below 10T". In contrast to sample-based queries, a process-based query retrieves a subgraph of the process DAG based on query conditions. For example, "What mechanical and chemical process steps are executed for synthesizing materials for which electron probe microanalysis are conducted? In which of these process steps do we need to use high-pressure inert atmosphere"? The resulting subgraphs may be edited and extended to create a new synthesis process.

The full extent of the QFDM is designed but it is currently under development. Next, we provide a summary of the progress as a part of the QDH architecture.

3 Quantum Data Hub Architecture

The Quantum Data Hub (QDH) platform provides scientists and researchers a unique combination of virtually unlimited storage space combined with a powerful data analysis environment. To this end, the QDH platform exploits developments in storage and database systems to provide large scale storage capabilities. QDH also leverages developments in computing to create a secure layered architecture. This coupled compute-data structure gives each scientists an agile workshop to build their ideas by leveraging advanced data science and artificial intelligence (AI) libraries to transform large quantum material datasets.

The QDH architecture, shown in Figure 4, consists of 5 main subsystems: (a) a User Authentication and Authorization; (b) a Cloud-based Object Storage back-end and an associated QDH application programming interfaces (APIs); (c)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

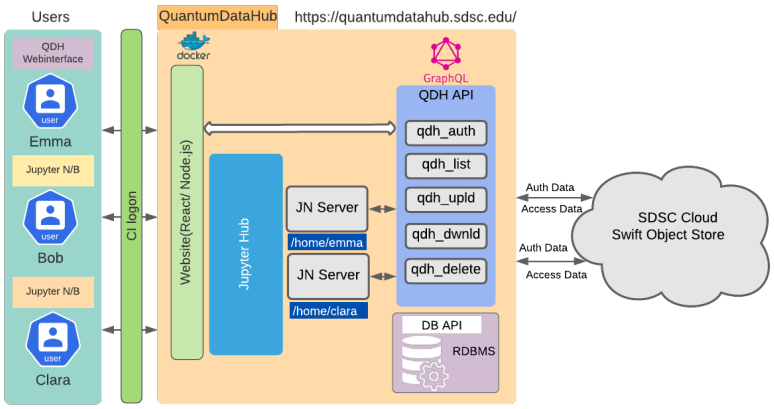


Fig. 4: System architecture of the Quantum Data Hub Platform.

a web-accessible user front-end; (d) a Jupyterhub based analysis environment; and (e) a database system that implements the data model described in Section 2.3. The QDH[4] platform enables multiple users to log in, record process and associated metadata, and upload data products related to the material synthesis process. A common usecase involves the material research scientists to log their sequence of steps, various equipment settings and corresponding outcomes of material synthesis process in their logbooks manually as shown in Figure 2a. The front-end web interface was designed to empower researchers and students to electronically capture a material synthesis process in form of scientific workflow DAG (e.g., the DAG shown in Figure 2b). The associated data is saved as metadata in a relational database whereas the data and computational products are saved in the cloud object store.

The platform is designed for collaborative research and the environment enables multiple researchers to simultaneously perform complex data analysis using the QDH, and yet store their data securely with required permission management to a unified shared Object Storage system. In order to discover relevant quantum datasets and experiments from unified Cloud Object store, researches can leverage the web-based search interfaces which uses DB queries with metadata attributes in a FAIR-compatible [19] way (e.g., Findable, Accessible, Interoperable, and Reusable). Researchers can connect to the unified Swift Cloud Object store using our QDH APIs and ingest data for analysis. The QDH APIs provide seamless integration of distinct users to shared storage space with precise access control capabilities. The platform leverages JupyterHub [14] as a data analysis platform. The JupyterHub enables each user to launch a dedicated Jupyter Notebook inside a sandboxed Docker [16] container that proxies it back to the user's web browser. The QDH APIs provide simple commands to perform data access operations on the unified Cloud Object storage such as, upload, download, update and delete through the Jupyter notebooks. Integrating the JupyterHub in the QDH amplifies the value of the Quantum Foundry data through

[4] The URL for Quantum Data Hub: https://quantumdatahub.sdsc.edu

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

data science. Its *workshop-in-a-browser* structure enables users to perform data analysis on the quantum datasets with low overhead.

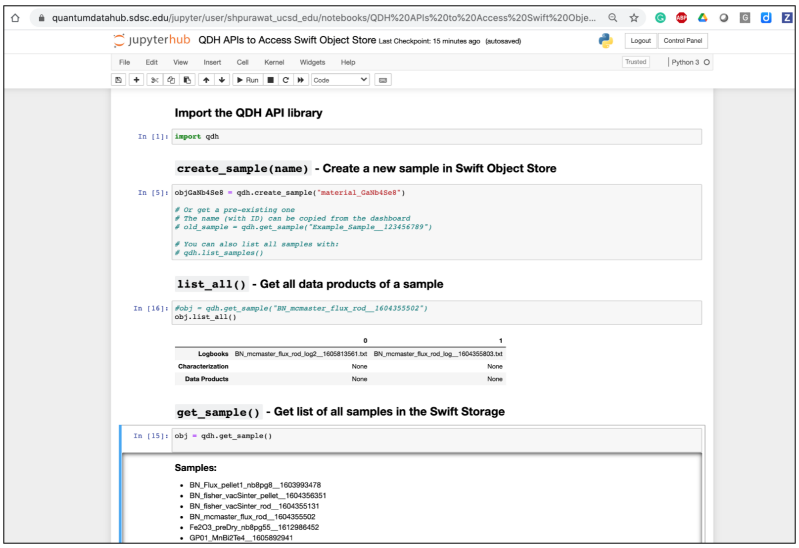


Fig. 5: QDH API calls to access Swift Storage from Jupyter Notebook

In the rest of this section, we summarize the main components of the QDH.

Authentication. The QDH platform authenticates users through CILogon [5]. CILogon leverages the OAuth 2.0 standard for token-based authentication to the cyberinfrastructure. Users can gain access to the QDH with their existing university credentials, or other preferred identity providers in few steps. The QDH platform uses a single-sign-on authentication paradigm to allow for navigation and access across its subsystems, such as the QDH front-end, JupyterHub and the Cloud Object Store. Once users' credentials are established at entrypoint to the QDH, users can log their experiments, work in Jupyter notebooks to analysis large amounts of data, leverage the APIs to store and access data products using multiple interfaces, and exploit other platform capabilities for their use-cases.

Authorization. The access control model in this system is based on a hierarchical group-based model and represented by the access control triple `<subject, object, permission>`. The subject of the access control model is an individual group, and objects are artifacts, results, or documents produced by the user. Users may have read, write or update permission on the document. A user may be a member of multiple groups. The fine-grained object-level access control will be the future work[11].

Cloud Object Store. We are using the Cloud storage at the San Diego Supercomputer Center using Swift (`https://www.sdsc.edu/support/cloud_storage_account.html`) to store and archive data products associated with quantum material synthesis process. Swift stores unstructured data in a scalable way to support growth of data over the time, and reliably maintains redundant

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

copies of data, performs error checking, and provides an economical option for research and academic projects.

QDH API. The Python-based QDH API library provides a high-level interface for researchers to interact with the Swift Object storage to store and retrieve data. The API commands use GraphQL queries to call Swift APIs, and can be called from a Jupyter notebook as well as to support the data operation behind the web-interface to perform create, upload, download, delete operations on data objects stored in Swift. Some of the QDH API commands include:

- `qdh.create_sample("sampleName")` - Creates an object with the given sampleName by a user as a top-level object to stores its associated data products.
- `qdh.list_samples()` - Lists all the samples stored in the Swift storage.
- `obj.list_all()` - Lists all the data products associated with a sample, e.g., the data products and files under categories Logbooks and Characterization.
- `sampleName.upload_logbook()` - Submits a logbook into Swift.
- `uploaded_file.delete()` - Removes a file from Swift.
- `uploaded_file.download()` - Creates a local copy of a file.

Data Analysis Platform. JupyterHub provides a platform where multiple users can access a Jupyter Notebook environment to perform data analysis. The Jupyterhub in QDH enables users to perform data analysis, access, upload, and share data. Once the user logs in the QDH, with a click of a button they can start a dedicated Jupyter Notebook single-user application that proxies back to the user's browser. Users can develop or upload their algorithms in Jupyter notebooks and can use simple QDH API commands from Jupyter notebook to access quantum data objects in Swift. Figure 5 shows simple API calls from Jupyter Notebook to create, upload, download, delete operations on this data.

Web Interface. The QDH front-end leverages modern advancements in web development practices in order to provide a feature-rich application for researchers on all platforms. It was engineered with the following goals:

- Create, modify, and collaborate on quantum material sample projects.
- Upload, download, and edit data files and Jupyter Notebooks on Swift.
- Launch JupyterHub to edit and run Jupyter Notebooks.
- Edit procedures for each sample.

The front-end is written using React — an open-source JavaScript state management library that allows for the rapid creation of reusable stateful components, with each encapsulating its own logic. By composing a web front-end with a component-oriented model, development iteration cycles are sped up, and any user-experience issues may be triaged efficiently.

Once authenticated, a dashboard that includes data and notebooks available to the user is presented. The dashboard dynamically caches data retrieved from a GraphQL microservice that aggregates data from the Cloud Object store and Database. We leveraged technical advancements in Kepler scientific workflow management system, provenance and team science [3] [15] [4] to design QDH Procedure Editor (QDHPE). Users are able to create the procedure associated with each sample using the QDH Procedure Editor (QDHPE), designed to streamline the input of metadata by users. Each vertex displayed in the QDHPE stores associated mutable metadata for each procedure step. For each vertex,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

users can add custom sets of parameters, and reuse these modifications. The inspector pane for each vertex provides input suggestions based on linked URI references. For instance, a chemical sample vertex may provide a URI reference to a research chemical catalog entry. From there, associated properties (i.e. form, purity) will be offered as an auto-complete suggestion. After each change by a user, the QDHPE dynamically translates the graph representation into the QFDM (see Sec 2.3).

4 Related Work

Quantum materials research is a sub-branch of material science research that sits at the convergence point of Quantum Physics and Material Science. Approaching Material Science from a Quantum Physics perspective necessitates a fundamental shift in the process of innovation. The QDH was designed to support such an innovation process in a malleable, scalable, adaptable and collaborative fashion.

In recent times, there has been growing interest in the development of storage and computing platforms that are problem domain sensitive. Some of the storage platforms that allow users to share data for research purposes are CKAN [1], Seedme [7], NoMaD [10], OQMD [13]. These platforms solve a very pertinent problem of data sharing and tackle the problems related to scientific data itself. In contrast, QDH is a unified system that combines computation and data storage technologies to enable quantum material researchers to perform complex analytical tasks. Further, there have been numerous developments in the application of core computer science tools to benefit material science research in recent times. This includes the development of platforms such as MaterialCloud [18], NOMAD [10], AFLOW [8], Material Project [12], CMR [6]. The QDH removes the need to build and maintain separate systems and allows for streamlined research.

Although there are other material science data and collaboration platforms as described above, to the best of our knowledge, The QDH is the only material data platform dedicated to quantum material research for collaborative research that enables multiple researchers to capture a material synthesis process in form of scientific workflow as a Directed Acyclic Graph (DAG). In this DAG, each node saves associated metadata, data, and computational products of the respective synthesis steps and users can perform complex data analysis in JupyterHub using the data porting capability of the system. It offers a searchable database with user-provided metadata that scientists can query to find datasets relevant to their problem domain and combine it with reproducible analysis in a unified platform to accelerate experimentation and streamline the innovation process for synthesis, discovery, storage, and analysis of quantum materials.

5 Conclusions ad Future Work

This paper presented a new material data and analysis platform for ingestion, management and analysis of quantum material data to couple theory, experimentation and synthesis of quantum materials, built as a part of UC Santa Barbara's NSF Quantum Foundry. Quantum Foundry is the only resource funded by NSF for design and development of materials related to quantum information. The

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

open exchange of data and its organization together with a built in analytical platform within one environment accelerates quantum material design and development as well as enabling new forms of training in this field. It also enables validation, reuse and repurposing of data and analytical products within the Foundry as well as sharing the built in know how with the rest of the world.

In this first description of the QDH, our objective was to describe the vision and progress towards this new resource as an example and in relationship to other related work in scientific computing and material science. While the QDH is fully functional and accessible to a limited group of researchers, the development is ongoing towards the full vision presented in Section 2.

As a part of the future work, we would like to link the generated data and insights with other material science data platforms through ontologies and knowledge graphs developed, e.g., the material commons by [2]. In addition, the presented QDH is being extended to enable multiple notebook based analysis process with seamlessly as an analytical workflow with self-reporting capabilities. Future work also includes an evaluation of system performance related to data ingestion, querying efficiency, analytical scalability and collaboration capabilities.

Acknowledgments

This work was supported by the National Science Foundation (NSF) through Enabling Quantum Leap: Convergent Accelerated Discovery Foundries for Quantum Materials Science, Engineering and Information (Q-AMASE-i): Quantum Foundry at UC Santa Barbara (DMR-1906325).

References

1. ckan, `https://ckan.org/`
2. Aaegesen, L.K., Adams, J.F., et al.: Prisms - an integrated open source framework for accelerating predictive structural materials science. JOM. (2018) October, p **1-17.** (2018)
3. Altintas, I., Purawat, S., Crawl, D., Singh, A., Marcus, K.: Toward a methodology and framework for workflow-driven team science. Computing in Science Engineering **21**(4), 37–48 (2019). https://doi.org/10.1109/MCSE.2019.2919688
4. Altintas, I., Wang, J., Crawl, D., Li, W.: Challenges and approaches for distributed workflow-driven analysis of large-scale biological data: Vision paper. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. p. 73–78. EDBT-ICDT '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2320765.2320791, `https://doi.org/10.1145/2320765.2320791`
5. Basney, J., Flanagan, H., Fleury, T., Gaynor, J., Koranda, S., Oshrin, B.: CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations. In: Proceedings of International Symposium on Grids Clouds 2019 — PoS(ISGC2019). vol. 351, p. 031 (2019). https://doi.org/10.22323/1.351.0031
6. Bligaard, T., DuÅak, M., Greeley, J., Nestorov, S., HummelshÃ¸j, J.S., Landis, D.D., NÃ¸rskov, J.K., Jacobsen, K.W.: The computational materials repository. Computing in Science Engineering **14**(06), 51–57 (nov 2012). https://doi.org/10.1109/MCSE.2012.16

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52

7. Chourasia, A., Nadeau, D., Norman, M.: Seedme: Data sharing building blocks. In: Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact. PEARC17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3093338.3104153, https://doi.org/10.1145/3093338.3104153
8. Curtarolo, S., Setyawan, W., Hart, G.L.W., Jahnatek, M., Chepulskii, R.V., Taylor, R.H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M.J., Stokes, H.T., Demchenko, D.O., Morgan, D.: AFLOW: An automatic framework for high-throughput materials discovery **58**, 218–226. https://doi.org/https://doi.org/10.1016/j.commatsci.2012.02.005, https://www.sciencedirect.com/science/article/pii/S0927025612000717
9. Dasgupta, S., Coakley, K., Gupta, A.: Analytics-driven data ingestion and derivation in the AWESOME polystore. In: IEEE International Conference on Big Data, Washington DC, USA. pp. 2555–2564. IEEE Computer Society (Dec 2016)
10. Draxl, C., Scheffler, M.: Nomad: The fair concept for big data-driven materials science. MRS Bulletin **43**(9), 676–682 (2018). https://doi.org/10.1557/mrs.2018.208
11. Gupta, M., Patwa, F., Sandhu, R.: An attribute-based access control model for secure big data processing in hadoop ecosystem. In: Proceedings of the Third ACM Workshop on Attribute-Based Access Control. pp. 13–24 (2018)
12. Jain, A., Montoya, J., Dwaraknath, S., Zimmermann, N.E.R., Dagdelen, J., Horton, M., Huck, P., Winston, D., Cholia, S., Ong, S.P., Persson, K.: The materials project: Accelerating materials design through theory-driven data and tools. In: Andreoni, W., Yip, S. (eds.) Handbook of Materials Modeling : Methods: Theory and Modeling, pp. 1–34. Springer International Publishing. https://doi.org/10.1007/978-3-319-42913-$7_6$0 − 1, https://doi.org/10.1007/978-3-319-42913-7_60-1
13. Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., RÃŒhl, S., Wolverton, C.: The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies **1**(1), 1–15. https://doi.org/10.1038/npjcompumats.2015.10, https://www.nature.com/articles/npjcompumats201510
14. Kluyver, T., Ragan-Kelley, B., Pérez, F., et al.: Jupyter notebooks - a publishing format for reproducible computational workflows. In: Loizides, F., Scmidt, B. (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90. IOS Press, Netherlands (2016), https://eprints.soton.ac.uk/403913/
15. LudÃ€scher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system: Research articles **18**(10), 1039–1065
16. Merkel, D.: Docker: lightweight linux containers for consistent development and deployment. Linux journal **2014**(239), 2 (2014)
17. Russ, T.A., Ramakrishnan, C., Hovy, E.H., Bota, M., Burns, G.A.: Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. BMC bioinformatics **12**(1), 1–15 (2011)
18. Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A.V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S.P., Zoupanos, S., Adorf, C.S., Andersen, C.W., SchÃŒtt, O., Pignedoli, C.A., Passerone, D., VandeVondele, J., Schulthess, T.C., Smit, B., Pizzi, G., Marzari, N.: Materials cloud, a platform for open computational science **7**(1), 299. https://doi.org/10.1038/s41597-020-00637-5, https://www.nature.com/articles/s41597-020-00637-5
19. Wilkinson, M.D., Dumontier, M., , et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 2052–4463 (2016). https://doi.org/https://doi.org/10.1038/sdata.2016.18

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_52