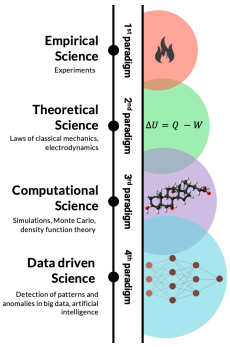
Accelerating the Adoption of Research Data Management Strategies

Johanne Medina[1], Abdul Wahab Ziaullah[2], Heesoo Park[3], Ivano E. Castelli[4], Arif Shaon[5], Halima Bensmail[6], Fedwa El-Mellouhi[2]

[1] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[2] Qatar Environment and Energy Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[3] Centre for Material Science and Nanotechnology, Department of Chemistry, University of Oslo, Oslo, Norway
[4] Department of Energy Conversion and Storage, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[5] Qatar National Library, Doha, Qatar
[6] Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar

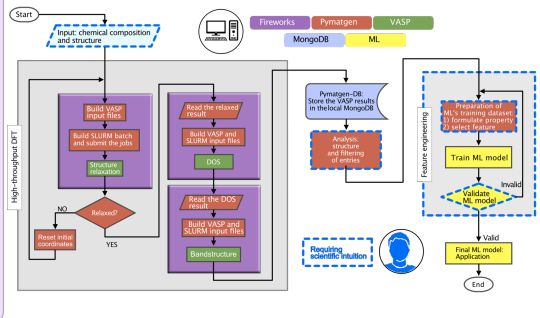
INTRODUCTION



The need for good research data management practices is becoming more recognized as a critical part of research. This is driven by the exponential rate at which data were obtained through high-throughput computations leading to the 5V challenge in Big Data, including volume, variety, velocity, veracity, and value. Poorly managed data often restricts its usability and accessibility, especially when various research groups and organizations collaborate on multidisciplinary research projects. This deters data users because they lack a basic understanding of the data's provenance and conditions of use. The material science community is no exception to the challenges of data deluge as it heralds its new paradigm of data-driven science. This paradigm uses artificial intelligence to accelerate materials discovery but requires massive datasets to perform effectively. Hence, there are efforts to standardize, curate, preserve, and disseminate these data in a way that is Findable, Accessible, Interoperable, and Reusable (FAIR). To understand the current state of data-driven material science, we surveyed researchers working on a small-scale research project and another within a large consortium. We analyze the current research status of each community and the challenges they face regarding the use and management of research data. This enables us to provide relevant recommendations to develop and/or procure an effective research data management system following the FAIR guiding principles. This work positions these recommendations on their urgency within the data-driven research life cycle.

Small-scale Group: AIPAM project

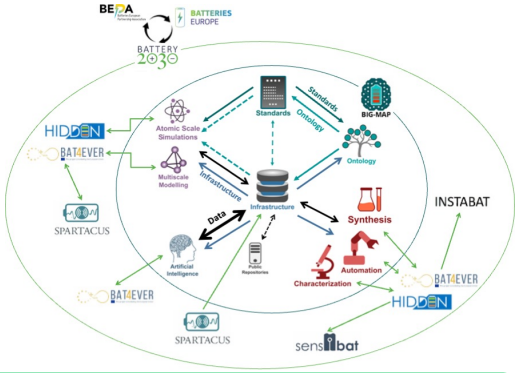
The Artificial Intelligence Platform (AIPAM) deals with high-throughput data-driven machine learning to discover novel materials. The data is generated by DFT computations.



Challenges:
- Difficulties in the induction of a new member to the team to replace / reproduce work of an existing team member who had left the team
- This is a time-consuming process and may require a steep learning curve
- Papers alone do not contain all information required to reproduce the results from previous computations. Most of the scientific journals are often concise due to page limits.
- One of the hurdles is the progression of pre-processing and post-processing tasks from high-throughput stage to machine learning stage.

Large Consortia: The BIG-MAP project

The Battery Interface Genome – Materials Acceleration Platforms (BIG-MAP) aims at accelerating the battery discovery through the development of a unique infrastructure and accelerated methodologies.



Challenges:
- Practical implementation of Data Management Plans (DMPs) is still in its early stages
- Very often, research data are stored in private repositories, lacking the basic FAIR principles
- Lack of common data ontology and the large number of different approaches adopted and repositories implemented make the link between data and projects difficult
- DMPs require the efforts of the entire community to define standards and ontologies which need to be chemistry and technology neutral
- Dedicated manpower is needed

RECOMMENDATIONS

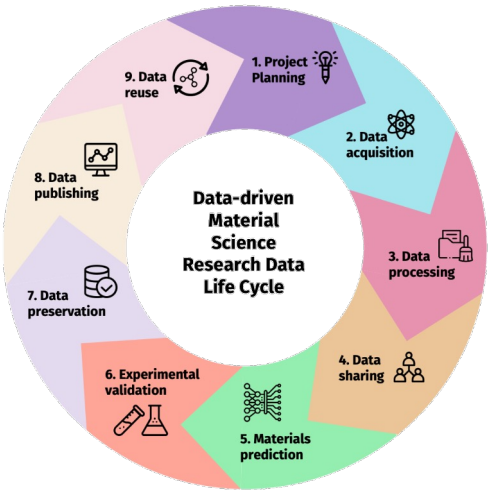
Materials Prediction
- Implement version control.
- Capture the environment with the use of containers, if possible.

Experimental Validation
- Use existing parsers where feasible and document their source.
- Document and version control any custom or purpose-built parsers.

Data Preservation
- Implement a suitable storage solution for both active and final data.
- Ensure regular backups of active data with an efficient data redundancy strategy. Understand the difference between backups and preservation.
- Preserve the final data in a trusted repository for the long term.

Data Publishing and Data Reuse
- Contribute to materials science consortia.
- Upload your data on established materials repositories.

Other recommendations
- Consider DMPs as part of research assessment.
- Recognize and reward FAIR data and stewardship.
- Provide sustainable funding.
- Organize and attend research data management training for researchers.
- Include data management in the curricula for degrees.



Project Planning
- Create and follow a detailed data management plan. Update the plan throughout the project lifecycle as needed. The plan should include important decisions about the source of the data to be collected for the project.
- Develop a file and folder naming standard for the project team to use.

Data Acquisition and Data Processing
- Use widely adopted data standards and associated file formats to ensure interoperability.
- Capture as detailed metadata as possible using existing metadata standards or ontologies. Support and develop domain specific ontologies, if needed.
- Write detailed documentation about all key aspects of data – e.g. provenance, storage, decisions, etc.
- Automate the generation of metadata and execution of data processing workflows, if possible.
- Store raw data independently, if feasible.

Data Sharing
- Adopt FAIR principles for sharing data. Create globally persistent and unique identifiers (PID) for datasets.
- Develop a user intuitive web portal with suitable Application Programming Interface (API) for automated data access, sharing, and discovery, if feasible.
- Display intelligent visualization.

ACKNOWLEDGMENTS: This work is supported by the Qatar National Research Fund (QNRF) through the Graduate Sponsorship Research Award Cycle 8 (GSRA8-L-2-0503-21026). IEC acknowledge support from the European Union's Horizon 2020 research and innovation program under grant agreement No957189 (BIG-MAP) and No 957213 (BATTERY 2030+).HP acknowledges the support of MoZEES, a Norwegian Centre for Environment-friendly Energy Research (FME),co-sponsored by the Research Council of Norway (project number 257653).