Concepts
doi.org/10.1002/batt.202100117

Chemistry Europe
European Chemical Societies Publishing

VIP *Very Important Paper*

Data Management Plans: the Importance of Data Management in the BIG-MAP Project**

Ivano E. Castelli,*[a] Daniel J. Arismendi-Arrieta,[b] Arghya Bhowmik,[a] Isidora Cekic-Laskovic,[c] Simon Clark,[d] Robert Dominko,[e] Eibar Flores,[a] Jackson Flowers,[f] Karina Ulvskov Frederiksen,[a] Jesper Friis,[g] Alexis Grimaud,[h, i] Karin Vels Hansen,[a] Laurence J. Hardwick,[j] Kersti Hermansson,[b] Lukas Königer,[k] Hanne Lauritzen,[a] Frédéric Le Cras,[l] Hongjiao Li,[m] Sandrine Lyonnard,[n] Henning Lorrmann,[o] Nicola Marzari,[p] Leszek Niedzicki,[q] Giovanni Pizzi,[p] Fuzhan Rahmanian,[f] Helge Stein,[f, r] Martin Uhrin,[a] Wolfgang Wenzel,[m] Martin Winter,[c] Christian Wölke,[c] and Tejs Vegge*[a]

[a] *Prof. I. E. Castelli, Prof. A. Bhowmik, Dr. E. Flores, K. Ulvskov Frederiksen, Dr. K. V. Hansen, Dr. H. Lauritzen, Dr. M. Uhrin, Prof. T. Vegge*
Department of Energy Conversion and Storage
Technical University of Denmark
2800 Kgs. Lyngby, Denmark
E-mail: ivca@dtu.dk
teve@dtu.dk

[b] *Dr. D. J. Arismendi-Arrieta, Prof. K. Hermansson*
Department of Chemistry-Ångström Laboratory
Uppsala University
Box 538, 75121, Uppsala, Sweden

[c] *Dr. I. Cekic-Laskovic, Prof. M. Winter, Dr. C. Wölke*
Helmholtz Institute Münster
IEK-12,
Forschungszentrum Jülich GmbH
48149 Münster, Germany

[d] *Dr. S. Clark*
SINTEF Industry,
New Energy Solutions
7034 Trondheim, Norway

[e] *Prof. R. Dominko*
National Institute of Chemistry
Hajdrihova 19, 1000 Ljubljana, Slovenia

[f] *J. Flowers, F. Rahmanian, Prof. H. Stein*
Helmholtz Institute Ulm (HIU)
Lise-Meitner Str. 16, 89081 Ulm, Germany

[g] *Dr. J. Friis*
SINTEF Industry,
Materials and Nanotechnology
7034 Trondheim, Norway

[h] *Dr. A. Grimaud*
Chimie du Solide et de l'Energie
Collège de France
UMR 8260, 75231 Paris Cedex 05, France

[i] *Dr. A. Grimaud*
Réseau sur le Stockage Electrochimique de l'Energie (RS2E),
CNRS FR3459
33 rue Saint Leu, 80039 Amiens Cedex, France

[j] *Prof. L. J. Hardwick*
Stephenson Institute for Renewable Energy,
Department of Chemistry
University of Liverpool
Liverpool, L69 7ZF UK

[k] *Dr. L. Königer*
Lab Automation and Bio-reactor Technology
Fraunhofer Institute for Silicate Research ISC
Neunerplatz 2, 97082 Würzburg, Germany

[l] *Dr. F. Le Cras*
University of Grenoble Alpes
CEA, LITEN, DEHT
38000 Grenoble, France

[m] *Dr. H. Li, Prof. W. Wenzel*
Institute of Nanotechnology (INT)
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz Platz-1, 76344, Eggenstein-Leopoldshafen, Germany

[n] *Dr. S. Lyonnard*
University of Grenoble Alpes
CEA, CNRS, IRIG, SyMMES
Grenoble 38000, France

[o] *Dr. H. Lorrmann*
Fraunhofer R&D Center Electromobility
Fraunhofer Institute for Silicate Research ISC
Neunerplatz 2, 97082 Würzburg, Germany

[p] *Prof. N. Marzari, Dr. G. Pizzi*
Theory and Simulation of Materials (THEOS), and
National Centre for Computational Design and Discovery of Novel Materials (MARVEL)
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland

[q] *Prof. L. Niedzicki*
Faculty of Chemistry
Warsaw University of Technology
Noakowskiego 3, 00-664 Warszawa, Poland

[r] *Prof. H. Stein*
Institute of Physical Chemistry (IPC)
Karlsruhe Institute of Technology (KIT)
Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany

[s] *Prof. M. Winter*
MEET Battery Research Center
University of Münster
48149 Münster, Germany

[**] *BIG-MAP: Battery Interface Genome – Materials Acceleration Platform. A previous version of this manuscript has been deposited on a preprint server (DOI: http://arxiv.org/abs/2106.01616)*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202100117

Open access to research data is increasingly important for accelerating research. Grant authorities therefore request detailed plans for how data is managed in the projects they finance. We have recently developed such a plan for the EU–H2020 BIG-MAP project—a cross-disciplinary project targeting disruptive battery-material discoveries. Essential for reaching the goal is extensive sharing of research data across scales, disciplines and stakeholders, not limited to BIG-MAP and the European BATTERY 2030+ initiative but within the entire battery community. The key challenges faced in developing the data management plan for such a large and complex project were to generate an overview of the enormous amount of data that will be produced, to build an understanding of the data flow within the project and to agree on a roadmap for making all data FAIR (findable, accessible, interoperable, reusable). This paper describes the process we followed and how we structured the plan.

1. Introduction

Consistent data management is of paramount importance for the acceleration of science and attracts steadily increasing attention from all entities throughout the entire research value chain, from the ones funding the research, to the researchers and the users of the data, which include industrial stakeholders. The attention is reasoned in a need for not only storing data related to published discoveries, but also to build new research on previously generated research data from any trustworthy source, which contributes to accelerating research. For us, the term research data includes all data generated during the research project, e.g., experiments and simulations, as well as any material underpinning the data such as laboratory notebooks, standards and protocols, pictures and videos, scripts and codes. All these should be open and accessible for reuse well beyond the duration of the specific project if there is no passable reason for restricting the openness.

With the integration of automated workflows[1–3] and Artificial Intelligence (AI) in our everyday research,[4,5] streamlined access to data becomes even more needed than before as AI models depend on the availability of large quantities of scientific data typically not found in a single publication, a single dataset or produced by an individual research group or institution. Next to this, strong voices in the scientific community have indicated that there is a significant crisis related to reproducibility of data and to establish procedures to enable that.[6] Curation, preservation, and sharing of data has thus become a fundamental part of our everyday life as researchers. The preparation of a detailed Data Management Plan (DMP) is one of the most significant initiatives that have been designed for increasing awareness on research data (store, share, reuse). The importance of a DMP is emphasized by the fact that its submission is now requested by many grant authorities and academic institutions, upon starting of a new project. An example is the European Commission, who requests submission of a DMP within the first six months for their ERC (European Research Council) and H2020 projects. The purpose of the DMP is to promote curation, storage, and sharing of the generated data by describing the research data generated within the project, by stating for whom these data might be useful and by unfolding how the project ensures that the storing, sharing, publication and preservation of the generated data comply with the FAIR (Findable, Accessible, Interoperable, Reusable) principles[7] to the greatest possible extent. Whereas this set compliance rules might be seen as a burden, there are measurable positive impacts for science as a whole as the fundamental way of pursuing research is rethought towards a sharing community. To enable this, in the last decade, several databases designed for curation and easy sharing of experimental and computational research data have been established. Examples from the world of materials research are: Inorganic Crystal Structure Database (ICSD),[8] Springer Materials,[9] the Cambridge Crystallographic Data Centre,[10] Materials Project,[11,12] Materials Cloud,[13] Open Quantum Materials Database (OQMD),[14] the AFLOW consortium,[15] the Open Materials Database,[16] and the Novel Materials Discovery Laboratory (NOMAD).[17]

This article describes how we developed a DMP for the H2020-LC-BAT-12 project Battery Interface Genome – Materials Acceleration Platform (BIG-MAP, www.BIG-MAP.eu). BIG-MAP[18] is a large-scale European research initiative, spanning 34 leading partners from academia, research organizations and large-scale R&D facilities as well as leading enterprises in the field. BIG-MAP is one of the six research projects belonging to the BATTERY 2030+ initiative,[19] aiming at discovering the battery of the future through disruptive development of new technologies. The goal of BIG-MAP is to achieve a five to ten fold acceleration of the discovery of new battery materials. To achieve this goal, it is necessary to develop a modular, closed-loop platform able to bridge physical insights and data-driven approaches. The BIG-MAP strategy is to cohesively integrate machine learning, computer simulations and AI-orchestrated experiments covering materials synthesis plus characterization and testing of materials and components, in order to accelerate the discovery and optimisation of battery materials. Optimal utilization of research data is crucial for achieving this goal and, in broader terms, to set a standard for future large-scale, data-centric projects aiming at accelerated discovery of materials for use in clean-energy technologies. Pivotal for this is the implementation of the BIG-MAP shared infrastructure, designed to support interoperability of the generated data, the BIG-MAP code repository[20] for shared development of computational tools and the BIG-MAP App Store[21] designed for easy public access to user-friendly versions of the developed tools. Tools will be made available in the GitHub organization and in the App Store. The tools will be able to access specific data needed for performing their task irrespectively of where the data is stored, e.g., the OPTIMADE API.[22–24] The data generated in the project will be archived in recognized repositories such as the

open-access repository Materials Cloud[13,25] for computational materials science recognized by Open Research Europe,[26] large-scale facility repositories for synchrotron data, and the repositories of project partners for electrochemical measurements. Gradually, as the BIG-MAP infrastructure develops, the data will also be made available for the project partners via this shared infrastructure.

This article comprises four main parts: i) an overview of BIG-MAP and its link with the BATTERY 2030+ family of European battery projects, ii) the strategy and process we have followed when working out the DMP, iii) an overall description of the research data generated in the project and iv) a description of how the project methodologies and open-data policy facilitate the storage of data of high FAIR-ness. The detailed description of the data generated within each work package (WP) is reported in the Supporting Information. Included in the Supporting Information, there is also a detailed description of the tools and precautions implemented in order to ensure high FAIR-ness of the research data. The DMP published here is a snapshot of a living document, which will be constantly updated for the whole duration of the project.

2. BIG-MAP and the BATTERY 2030+ Large-Scale Research Initiative

Today, energy production and transport are evolving fast to meet challenging environmental targets and growing demand. Energy storage is the Achilles heel of the accelerating efforts for sustainable energy production and use. The search for both low-cost and high-performance materials and devices cannot rely on incremental improvements of conventional technologies, but rather require the accelerated discovery of disruptive technologies and of ultra-performing storage materials. To answer this need, the European Union has funded several battery initiatives under the Horizon 2020 scheme. These projects are organized under the umbrella of the BATTERY 2030+ consortium.[19] This consortium is composed of a coordination and support action (CSA) and six research innovation projects, which address multiple key challenges related to batteries. Following a chemistry neutral approach, BATTERY 2030+ and its member projects aim at *"reinventing the way we invent batteries"* by generating a toolbox of versatile infrastructures, common data frameworks and transferable knowledge that can be translated into design principles for discovering and developing new battery materials, as described in the BATTERY 2030+ Roadmap.[27] The vision of BIG-MAP is fully aligned with the goals of BATTERY 2030+. BIG-MAP focuses specifically on the challenge of *accelerating the discovery of new battery materials and interfaces that can only be achieved by understanding and predicting how the battery interfaces evolve in time and space*, which is a theme shared by all projects under the BATTERY 2030+ umbrella. In BIG-MAP, this will be achieved by creating an autonomous infrastructure able to design, synthesize, and test new materials across all domains of the battery development cycle, as well as orchestrate experiments and simulations on-the-fly. The project will be a lever to create the infrastructural backbone of a versatile and chemistry-neutral battery Materials Acceleration Platform in Europe, focusing in particular on the role of the interface and on how to achieve a five to ten fold increase in the rate of discovery of novel battery materials, interfaces and cells. This will set the stage for the European battery research community to efficiently develop and proliferate new battery chemistries in the coming decade.

3. Our Strategy and Process for Working out an Operational and Consistent DMP

Our ambition has been to layout a DMP that serves multiple purposes:

- An operational tool promoting the interaction and flow of data between the 10 scientific WPs of the project.
- A practical guide for the project partners for how to fulfil the obligation set by the H2020 Open Research Data Pilot.
- A demonstrator for how a DMP can facilitate active interaction with related R&D projects, here with the BATTERY 2030+ projects as a case.

We have chosen to work out the DMP in a collaborative effort crossing the entire project. The template we used follows to a large extent the standard DMP template for H2020

Table 1. Example of types and formats of the data generated by WP2.

Datatype	Description	Data sets	Type	Format	Size
Electronic Structure: WFT, DFT, QMC	Structures, energy-related data, wave functions & electronic properties, ab-initio molecular dynamics (AIMD) trajectories, different types of spectra	Data generated by different tools: Engines (molecular): GAUSSIAN, ORCA, MOLPRO, TURBOMOLE, NWChem, QChem, ADF, PSI4, MRCC, NECI Engines (periodic): CP2K, VASP, QUANTUM ESPRESSO, Yambo, Castep, GPAW, QuantumATK, Crystal, NECI	Tarball files can be created from the calculation folder, including relevant inputs and output raw data	.tar.gz (an archive of input and output text, XML, netctdf, hdf5, or any other machine-readable file)	TB

projects. This consists of two sections: a data summary describing the generated data and the FAIR section describing the incentives taken for making the data FAIR. However, we have introduced some modifications to the template.

In order to arrive at a tangible description of the generated data, we have chosen to work out separate data summaries for each WP-summaries that can serve as read-alone documents. The summaries have been supplemented by a general description of the main data categories generated in the project, which gives an overview of the data. Standard tables describe the main properties of the data, while flow diagrams illustrate in/out data streams of each WP. An overview showing the main routes for data exchange – the data highways – has also been included. See examples of the tables and data flow diagrams in the next section.

For the FAIR section, the process has been somewhat different. Here, the starting point has been a draft for each WP worked out by the WP leaders and following the H2020 template. Subsequently, the drafts have been condensed into one unified description of the incentives that shall be taken for ensuring that the generated data is FAIR. The description is done at three levels: the agreed general principles to be followed, an elaboration of these when needed, and a listing of reasoned exceptions from the general principles.

The WP leaders managed the work related to their WPs with the help of key participants from the WP, and the task was solved concurrently for all WPs and under the supervision of the BIG-MAP data manager. The concurrent process is regarded as important as it allowed for regular coordination between the WPs, for efficient tightening of possible open ends, and for developing a shared understanding of the complex structure of the generated data that crosses many disciplines. The result is an operational plan with shared ownership by all WP leads. The plan concretely describes the exchange of data inside the project and the outflow of data to the research community. Moreover, the plan serves the purpose of guiding the project work towards trustworthy data that can be readily accessed and used also from outside the project.

The strategy and process used can readily be adapted to other projects and be used for mapping and promoting collaboration between related initiatives. The key issues here will be to agree on a unified way of describing the data in order to ensure transparency and promote collaborative work. Presently, we are assisting the BATTERY 2030+ community in implementing this concept. We hope that the present publication will encourage other projects (within and outside BATTERY 2030+) to follow our approach for the preparation of their DMP plans, as this would be the first step towards an interoperability (and ultimately standardization) of research data.

One of the tasks of BATTERY 2030+ is to standardize data and protocols. Workshops and a task force have been organized to achieve this. The first step towards this is to homogenize the DMPs of the projects under the BATTERY 2030+ umbrella. This will allow us to reach a higher level of data interoperability between the projects than without such homogenization. By pioneering the definition of DMPs in large (battery) projects, BIG-MAP has the ambition of proposing a DMP template for the other projects dealing with petabytes of data and with data stemming from a broad range of disciplines and sources. It is important to note that writing a DMP is a team effort, which involves the expertise of many scientists. Our approach has been to involve all WP leaders, each one responsible for the writing of their WP DMP followed by a homogenization of the WP DMPs. This has created a sense of ownership and awareness of both the DMP and the need for data sharing across the project.

4. The DMP Data Summary and the Highway of Data

The research data generated in BIG-MAP is essential for creating a modular platform that future large-scale battery projects can build upon. The data originate from multiple sources: from physical experiments, simulations, and artificial intelligence. The data comes in many different forms: data can be processed or unprocessed, and data can be structured (e.g., files holding labelled rows and columns with numbers) but also unstructured (e.g., models, metadata, code, documents in natural language). To summarize the data generated, tables containing multiple information, from description and type of data set to format and size, are reported for each WP (see Table 1). Efficient exploitation of such heterogeneous datasets calls for the project-wide guidelines for generating, storing and querying of data described in a DMP.

BIG-MAP is organized in one management work package (WP1) and ten scientific work packages, spanning from the generation of data describing battery materials by means of computational methods and experiments (WP2-WP6), to the design and development of data management tools facilitating data exchange within the WPs (WP7, WP8, WP9), and to the development of AI-accelerated materials discovery and deep learning models for inverse materials design (WP10-WP11), trained using the data generated in WP2-WP6. An outline of the generation and flow of research data within the project is shown in Figure 1.

The starting point of the work is the knowledge and data already generated by the project partners combined with any relevant open data, i.e., data available in the research literature and in public repositories, e.g. topical repositories such as Inorganic Crystal Structure Database (ICSD),[8] Materials Cloud,[13] Materials Project,[11,12] and the Computational Materials Repository,[28] as well as general purpose repositories such as GitHub, Zenodo, and FigShare.

Following the flow of the data within the project is crucial. This does not only allow keeping track of how the work progresses, but also minimizes redundant research activities and, when needed, enables requests for new data across the WPs. Figure 2 shows the data and information flow in the project in more details than Figure 1. The developed infrastructure (WP9) is the hub that will ensure interoperable communication and transfer of data across the project.

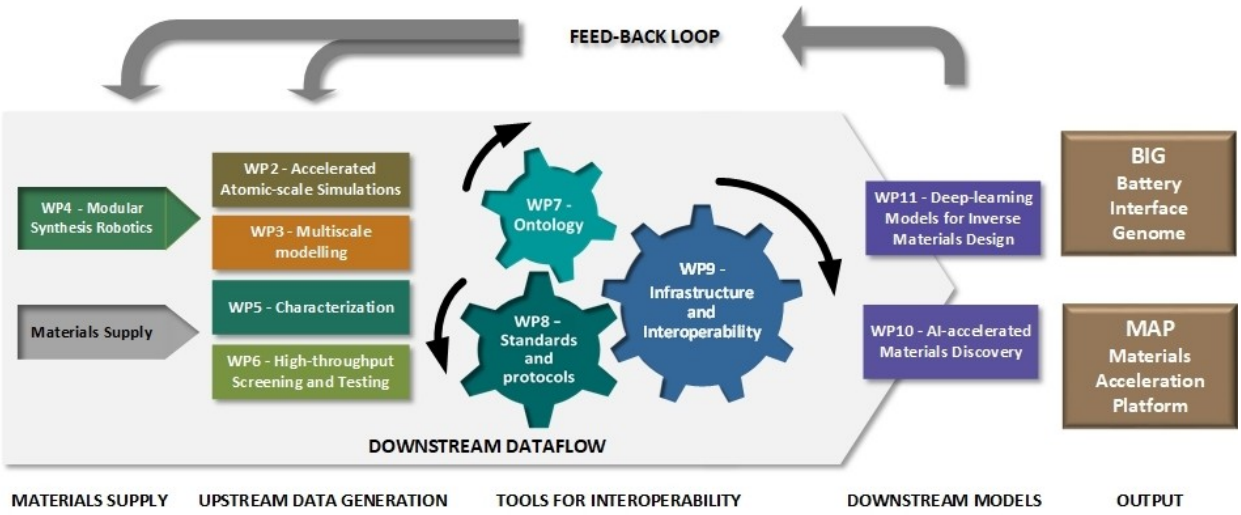

Figure 1. Schematic view of the types of data generated in the various WPs and the downstream data flow from WP2-WP6 via the tools developed in WP7-WP9 ensuring that the data can readily be used in WP10 and WP11.

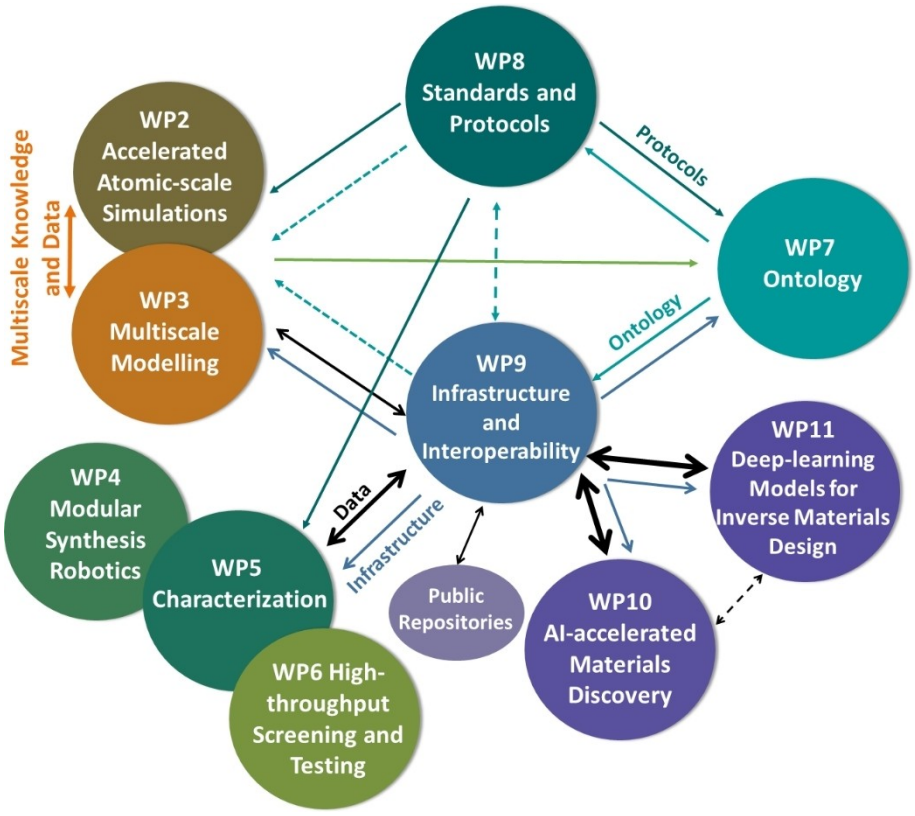

Figure 2. Highway of data: Overall information and data flow within BIG-MAP. The arrows show the main flows of information. The thickness of the arrows gives a visual indication of the amount of data transferred, which span from few KB and MB of text files and small datasets to TB of synchrotron data, as explained in more details in the data tables in the Supporting Information. The TB-PB dataset size puts serious constraints on where such data are stored, transferred and how the user can access them. The WPs have been grouped according to their affinity.

Included in the infrastructure will be a shared data storage and processing facility accessible for all BIG-MAP partners. The infrastructure will manage automated requests for data across the project, typically requests from AI-based models (WP10-11) for new data and the reporting back of the requested data to the models. The infrastructure will also include apps that can perform autonomously analyses of datasets on-the-fly, e.g., for spectroscopic or structural datasets. The tools supporting interoperability of the data and automation of the request-replay system are the battery interface ontology BattINFO[29]

(WP7) and the standards and protocols for data acquisition, for data processing and for reporting data and metadata (WP8). A typical automated workflow could be an acquisition request for new training data for the AI-models. The request will be sent to the infrastructure that in response to this, will search for existing data or request new data from the data-generating WPs (WP2-6) or automatically launch calculations that can generated the missing data. The ability of the system to perform such autonomous processes hinges on the built-in ability to describe the requested data and metadata explicitly. Here the ontology comes into play, as it allows for an unambiguous description of the data, e.g., a specific physio-chemical property of a specific part of the battery cell in a specific spatial domain. The ontology is implemented in the Web Ontology Language (OWL) [30] and defined using the top-level European Materials and Modelling Ontology (EMMO). It can be easily edited using open-source tools such as Protégé or EMMO-python. By adhering to the OWL standard and linking to other flexible domain ontologies of EMMO, the ontology can be easily integrated in the BIG-MAP infrastructure. Working examples on how to use an ontology in materials science can be found in the EMMO repository on GitHub.[31] Based upon this, the infrastructure can autonomously and accurately search for existing data, or alternatively identify the standards and protocols to be included in a request for new data to be forwarded to the data-generating WPs. All these workflows will be made available in the App Store.

WP2 and WP3 deal with AI-tools embedded atomistic and multiscale simulations of materials and interfaces. Overall, the synergy between these two WPs allows development of a multiscale knowledge of the battery using input from the experimental WPs, i.e., WP4, WP5, and WP6 (Figure 2) that cover respectively modular robotics experiments, characterization, and high-throughput synthesis and testing. The three experimental WPs work together on the automation of experiments following the agreed experimental protocols and standards that have been defined in WP8 and following the WP7 ontology (BattINFO), as well as on automated integration of the results in the simulation WPs which is essential for verifying the predictions and completing the models. Within the WP10 and WP11 a closed loop is created where all experimental and computational data are utilized for building a deep-learning model for spatio-temporal evolution of battery interfaces that will be used for effective exploration of the chemical and structural design space of batteries.

As the success of BIG-MAP hinges on easy and efficient sharing of data across the WPs and the scientific disciplines involved, it is crucial that all partners and WPs are fully aware of which data they shall deliver to other WPs and which data they shall receive from other WPs. Data format, data size and timing are essential information to convey between the involved partners in order to make this orchestra play. For this reason, we have chosen to highlight the data exchange within the project in the DMP. This has been done by including tables describing the inflow and outflow of data from each WP (see Table 2 and 3 for examples and find all tables in Supporting Information) and diagrams showing how each WP exchange data with the other WPs (see Supporting Information). Figure 2 represents an integration of the data-flow diagrams developed for each WP. The diagram shows only the most important data flows (the data highways), whereas the less significant but still important data flows have been omitted in order to arrive at a readable overview. Even though the data flow diagrams indicate that data is delivered and collected from a specific WP, all data passes through WP9, which ensures that data is transferred in a standardized manner between the WPs.

5. FAIR-ification of the BIG-MAP Data

In addition to the description of the generated research data given in the Data Summary section, the DMP shall also describe

Table 2. Example of types and formats of the data collected by WP2.

WP	What	To be used for	Suggested type	format	size
From WP5	Parameters from structural and chemical characterizations at the local scale, including e.g., spectra (vibrational, absorption, …)	Model refinement, cross-analysis, validation, and design of simulated experiments	Tarball files can be created with post-processed data and parameters of interest	.tar.gz (with databases and reproducible-data sheets)	MB

Table 3. Example of types and formats of the data delivered by WP2.

WP	What	Usable for	Suggested type	format	size
To WP5	Computational predictions for bulk/interfacial structures, "chemical environments", spectra, transport properties, or any experimental characterization requiring atomistic or electronic interpretation	Guiding the characterization effort and supporting the interpretation of the results	Tarball files can be created with post-processed data and parameters of interest	.tar.gz (with databases and reproducible-data sheets)	TB

how the project ensures that the generated research data is FAIR, and thereby giving the highest possible value for use within the project and reuse outside the project. BIG-MAP is part of the H2020 Open Research Data Pilot Programme of the European Commission. This programme provides Guidelines for FAIR Data Management, which serve therefore as guidelines for the handling of data within BIG-MAP. BIG-MAP will utilize the Open Research Data Pilot opening for restricting the access to specific data. For BIG-MAP this applies for specific data underpinning business secrets, data for which disclosure can obstruct protection of intellectual property, data collected from proprietary databases and some data generated by proprietary software. Any disclosure of BIG-MAP research data shall respect the Consortium Agreement.

Due to the complex relationships between the generated research data and the need for a high-level interoperability, it is seen as crucial to develop a project-intern data infrastructure and project-intern data-handling procedures that ensure high interoperability and thereby high data FAIR-ness for data throughout the entire data life cycle, i.e., already from the data acquisition. The exact procedure for handling the data and the incentives for FAIR-ification depend on the nature of the research data. The BIG-MAP data falls into two categories: basic data (tabulated data and images) and computational tools (code, software, scripts and apps).

The basic data will be documented, equipped with a unique identifier and transferred to the BIG-MAP shared data storage facility where the data will become available and usable for the consortium. Transfer to the shared storage facility shall happen as soon as practically possible after data generation. The basic data shall appear findable, accessible, interoperable and reusable for the project partners as soon as it becomes available via the consortium-wide storage facility. This serves as a test in a restricted forum of how FAIR the data is. Passing the test means that the datasets and their documentation are in form that allows them to be indexed in a data repository and subsequently published provided that there are no restrictions on the data openness. An exception from this standard data handling procedure applies for datasets generated at large-scale synchrotron or neutron facilities, as such datasets are typically in the TB scale. Due to their size, such data will remain in their raw form in the repositories of large-scale facilities, and only processed datasets of manageable size will be transferred from there and made directly available for a broader audience. The preferred site for publishing basic data will be repositories that are recognized within the battery research community, e.g., the ones from which BIG-MAP also harvests open data, and the institutional data repositories of the partners. All basic data shall, to the largest degree possible, be made available in open, non-proprietary formats. The software used should ideally be open source. If not possible, the software and tools needed for accessing the data shall be named. Published data shall be licensed in the least restrictive manner, e.g., Creative Commons Attribution (CC BY 4.0). Upon publication a persistent identifier (a DOI or handle) is assigned to the dataset.

In addition to the basic data, computational tools supporting the development of novel battery chemistries are highly valuable research outputs. The BIG-MAP GitHub organization serves as a shared space for developing the computational tools (code, software and scripts). GitHub version control, documentation and licensing will be utilized to their full extent. The preferred licences will be MIT, BSD, or GPL for software, codes and scripts, and Creative Commons Attribution (CC BY 4.0) for other content, e.g., the ontology. For further exposition of the tools of high usability for a broader audience, the GitHub organization has been linked with the public-facing BIG-MAP App Store where any user can access user-friendly versions of the tools, e.g., the "Quantum ESPRESSO AiiDAlab" app[32] for computing band structures and other structure properties. The App Store is equipped with a GUI (Graphic User Interface) that

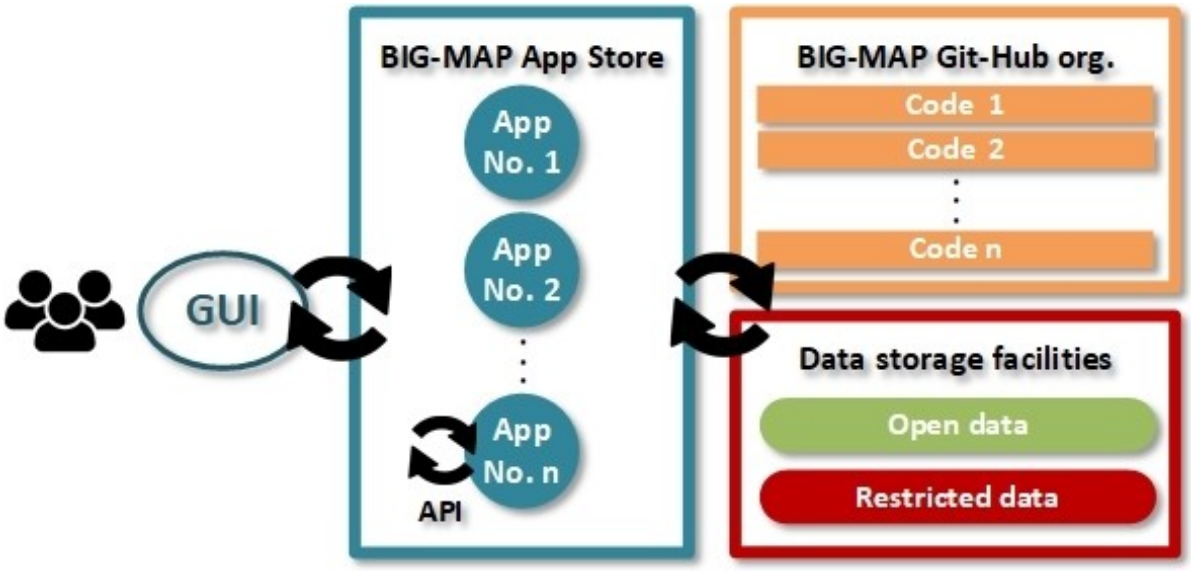

Figure 3. The BIG-MAP App Store: Conceptual view

allows search for content, i.e., apps, via standard web browsers, and the apps are equipped with APIs (Application Programming Interfaces) allowing software-independent communication and usage of the app, see Figure 3. The App Store is interlinked with both the BIG-MAP GitHub organization where the related source codes are deposited and with data storage facilities (the shared BIG-MAP data storage and public repositories) where the code/apps can access the data needed for performing their tasks. This architecture ensures maximum exploitation of all datasets, while complying with the legal constraints attached to the use of sensitive data.

All research data, irrespective of their form (basic data or computational tools), will be documented to the extent needed for relevant users to efficiently reuse the data. The documentation shall hold information on why, how and when the data was generated and who generated the data. For experimental data the documentation could take the following form: the specific aspect of the battery interface described by the data, the method used to generate the data with reference to standard protocols and standard operating procedures, and information on data fidelity and provenance. Essential for documentation is the battery interface ontology (BattINFO) and the BIG-MAP standards and protocols. BattINFO provides explicit naming of the key aspects related to the battery materials in general and especially the battery interfaces, see Figure 4, whereas the standards and protocols allow for precise references to how the data was generated and processed.

The documentation to follow the research data shall be in the form of metadata and keywords, preferably supplemented with written guidelines, e.g., Readme text files, to improve the reusability of the data. Metadata and keywords shall be machine-readable. The metadata shall, as far as possible, be generated through automatic routines. The automatically generated metadata can be manually complemented with custom metadata, if needed for efficient query and reuse of the data. In addition, automated metadata extraction shall be used both on dataset already published and as a benchmark for the metadata autonomously generated when the experiment or the simulation is performed. Tools such as the Battery Data Toolkit from the Materials Data Facility group have been implemented for this purpose,[33,34] and provide an important path to extract metadata. Data and metadata need to be stored together or connected via persistent links and indexed with machine-readable search keywords.

All research data shall be available for at least 10 years after the project's end date.

The BIG-MAP management structures have placed the responsibility for handling the research data securely and in compliance with the DMP per default with the partner who generated the data, but other models for sharing the responsibility can be agreed if more convenient and more partners are involved. A data manager whose task it is to guide the partners in any aspect related to data management has been appointed. All partners are responsible for complying the European Code of Conduct for Research Integrity[35] when performing their tasks in the project. More details regarding our approach to FAIR data are provided in the Supporting Information.

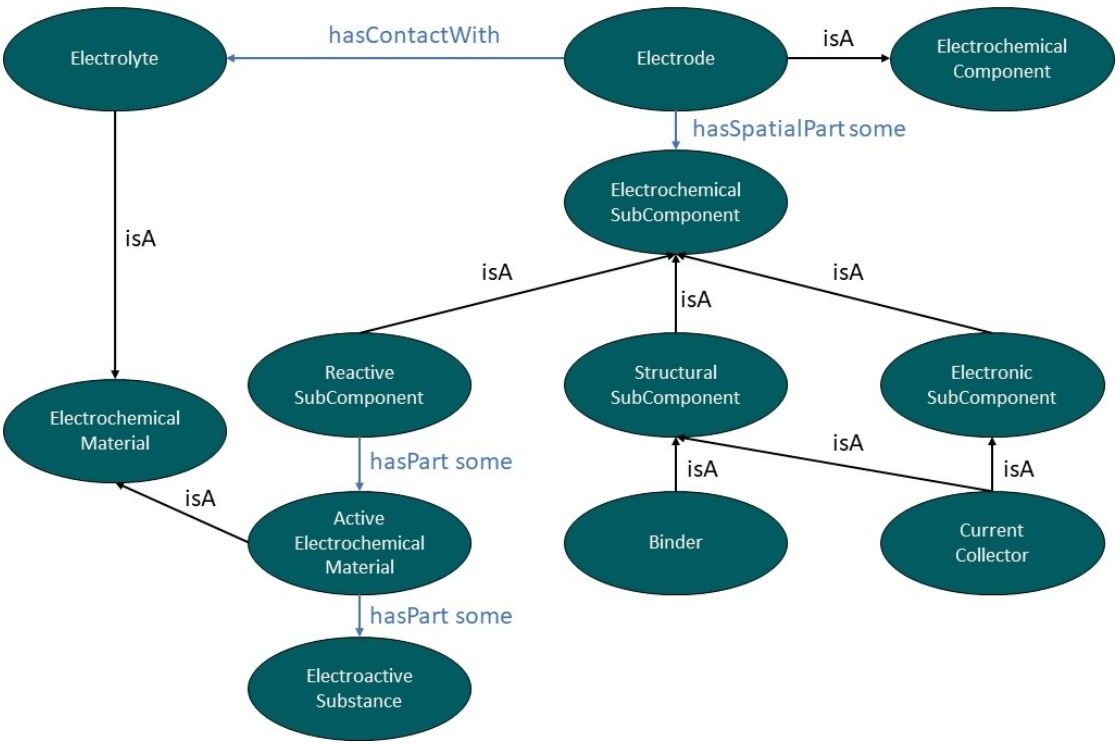

Figure 4. Example of one branch of the BattINFO ontology show the classes and relationships that describe an electrode.

6. Conclusions

In this work, we have highlighted the importance of data management in large and multi-domain data-centric research projects, where interoperability of data is fundamental for connecting many different tasks reliant on effective interactions between WPs and scientific disciplines.

Our take-home message is that a well worked-through DMP, holding concrete information on how to structure and handle the data, will serve as a valuable tool for guiding the project work towards trustworthy data that can be readily accessed and used also from outside the project. Another important point is that it takes efforts to concisely describe the data generated in large multi-disciplinary projects like BIG-MAP, however the effort pays back as a clear picture of how the many different datasets generated can and shall play together for maximizing the project output. This is not only valuable for managing the project, but also for providing a map that will give the project members an understanding of how their data fits into to the overall puzzle – an understanding that is essential for targeting the data generation to the precise need of the destination WPs.

We have also learned that the process used for working out the DMP is important. By working out the plan in a collaborative effort crossing the entire project, the DMP will appear not only as a document describing the generated data and how these are stored, documented and shared, but will also appear as an effective tool for planning and execution of the project and for keeping track of the work done by various partners in different WPs and their role within the overall project. The collaborative process itself is also important, as it raises the awareness in the team members of consistent data management and facilitates the project development of methodologies that support FAIR data handling. The DMP is thus a living document, which will evolve during the project and be constantly updated. In this respect, the DMP will benefit from a deeper use of the developed ontology, from the general implementation, which would help in better organizing the data tables into more operational categories, to the tools, such as Protégé,[36] which will contribute to visualize the connections between data in a more dynamic way. With these updates, we aim at increasing the connectivity between the data, not only of the BIG-MAP project, but, more ambitiously, among the whole BATTERY 2030+ community, establishing for the first time a common platform for sharing battery data.

Thanks to its connection with the BATTERY 2030+ initiative, BIG-MAP is defining new standards for data management across the whole battery community. The DMP described here could be used as a template by other initiatives aiming at a deeper integration of data and projects under a unified umbrella, which will ultimately contribute to accelerate battery discovery.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grants agreement No 957189 (BIG-MAP) and No 957213 (BATTERY 2030+). Jackson Flowers, Fuzhan Rahmanian and Helge Stein acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2154 – Project number 390874152. Sandrine Lyonnard would like to acknowledge Poul Norby, Aleksandar Matic, Marnix Wagemaker, Ennio Capria, Duncan Atkins, and Stéphanie Belin for the fruitful discussions during the preparation of the section describing WP5.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: batteries · data curation · data management plan · databases · FAIR data

[1] F. T. Bölle, N. R. Mathiesen, A. J. Nielsen, T. Vegge, J. M. Garcia-Lastra, I. E. Castelli, *Batteries & Supercaps* **2020**, *3*, 488–498.
[2] L. Kahle, A. Marcolongo, N. Marzari, *Energy Environ. Sci.* **2020**, *13*, 928–948.
[3] S. M. Blau, H. D. Patel, E. W. C. Spotte-Smith, X. Xie, S. Dwaraknath, K. A. Persson, *Chem. Sci.* **2021**, *12*, 4931–4939.
[4] A. Bhowmik, I. E. Castelli, J. M. Garcia-Lastra, P. B. Jørgensen, O. Winther, T. Vegge, *Energy Storage Mater.* **2019**, *21*, 446–456.
[5] F. T. Bölle, A. Bhowmik, T. Vegge, J. Maria García Lastra, I. E. Castelli, *Batteries & Supercaps* **2021**, 4, 1516–1524.
[6] M. Baker, *Nature* **2016**, *533*, 452–454.
[7] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018.
[8] Inorganic Crystal Structure Database, https://icsd.fiz-karlsruhe.de.
[9] Springer Materials, https://materials.springer.com.
[10] The Cambridge Crystallographic Date Centre, https://www.ccdc.cam.ac.uk.
[11] The Materials Project, https://materialsproject.org.
[12] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[13] Materials Cloud, https://www.materialscloud.org.
[14] The Open Quantum Materials Database, https://oqmd.org.
[15] Automatic – FLOW for Materials Discovery, http://aflowlib.org.
[16] The Open Materials Database, http://openmaterialsdb.se.
[17] NOMAD Centre of Excellence, https://nomad-coe.eu.
[18] The Battery Interface Genome – Materials Acceleration Platform (BIG-MAP) project, https://www.big-map.eu.
[19] BATTERY 2030+, https://battery2030.eu.
[20] BIG-MAP GitHub Organization, https://github.com/BIG-MAP.
[21] BIG-MAP App Store, https://big-map.github.io/big-map-registry.
[22] C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, Á. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M.

Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. Mostofi, K. A. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu, X. Yang, *Sci. Data* **2021**, *8*, 217.
[23] OPTIMADE – Open Databases Integration for Materials Design, https://www.optimade.org.
[24] OPTIMADE Web Client, https://big-map.github.io/big-map-registry/apps/optimade-web.html.
[25] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, N. Marzari, *Sci. Data* **2020**, *7*, 299.
[26] Open Research Europe, https://open-research-europe.ec.europa.eu.
[27] BATTERY 2030+, Inventing the sustainable batteries of the future, https://battery2030.eu/research/roadmap.
[28] The Computational Materials Repository (CMR), https://cmr.fysik.dtu.dk.
[29] BattINFO – Battery Interface Ontology, https://www.big-map.eu/https://github.com/BIG-MAP/BattINFO.
[30] Web Ontology Language, https://www.w3.org/OWL.
[31] European Materials Modelling Council (EMMC), European Materials and Modelling Ontology (EMMO), https://github.com/emmo-repo.
[32] Quantum ESPRESSO AiiDAlab App, https://big-map.github.io/big-map-registry/apps/aiidalab-qe.html.
[33] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, *JOM* **2016**, *68*, 2045–2052.
[34] Materials Data Facility - Battery Data Toolkit, https://github.com/materials-data-facility/battery-data-toolkit.
[35] The European Code of Conduct for Research Integrity, https://allea.org/code-of-conduct.
[36] Protégé – A free, open-source ontology editor and framework for building intelligent systems, https://protege.stanford.edu.

Manuscript received: May 28, 2021
Revised manuscript received: August 27, 2021
Accepted manuscript online: August 28, 2021
Version of record online: September 24, 2021