Journal Pre-proof



Demonstrating Linked Battery Data to Accelerate Knowledge Flow in Battery Science

Philipp Dechentz , Elias Barbers , Simon Clark , Susanne Lehner , Brady Planden , Masaki Adachi , Dirk Uwe Sauer , David A. Howey , Sabine Paarmann

PII: S0013-4686(26)02284-X
DOI: https://doi.org/10.1016/j.electacta.2026.150404
Reference: EA 150404

To appear in: *Electrochimica Acta*

Received date: 30 September 2026
Revised date: 27 October 2026
Accepted date: 9 November 2026

Please cite this article as: Philipp Dechentz , Elias Barbers , Simon Clark , Susanne Lehner , Brady Planden , Masaki Adachi , Dirk Uwe Sauer , David A. Howey , Sabine Paarmann , Demonstrating Linked Battery Data to Accelerate Knowledge Flow in Battery Science, *Electrochimica Acta* (2026), doi: https://doi.org/10.1016/j.electacta.2026.150404

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 The Author(s). Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Journal Pre-proof

Demonstrating Linked Battery Data to Accelerate Knowledge Flow in Battery Science

Philipp Dechentz[1*], Elias Barbers[2*], Simon Clark[3], Susanne Lehner[4], Brady Planden[1], Masaki Adachi[1], Dirk Uwe Sauer[2,7,8], David A. Howey[1,6], Sabine Paarmann[5,7,8,z]

[1] Department of Engineering Science, University of Oxford, OX1 3PJ, Oxford, UK
[2] Helmholtz Institute Münster (HI MS), IMD-4, Forschungszentrum Jülich, Campus-Boulevard 89, 52066 Aachen, Germany
[3] Battery Technology, SINTEF Industry, Bratsbergvegen 5, Trondheim 7031, Norway
[4] Department Elektrotechnik-Elektronik-Informationstechnik (EEI), Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstraße 9, 91058, Erlangen, Germany
[5] Department of Mechanical Engineering, Imperial College London, UK.
[6] Faraday Institution, Quad One, Harwell Campus, Becquerel Ave, Didcot OX11 0RA, UK.
[7] Electrochemical Energy Conversion and Storage Systems, Institute for Power Electronics and Electrical Drives (ISEA), RWTH Aachen University, Campus-Boulevard 89, 52074, Aachen, Germany
[8] Center for Ageing, Reliability and Lifetime Prediction for Electrochemical and Power Electronic Systems (CARL), RWTH Aachen University, Campus-Boulevard 89, 52074, Aachen, Germany

[z] Corresponding authors philipp.dechent@gmail.com, sabine.paarmann@isea.rwth-aachen.de
*These authors contributed equally

Abstract

More than 20,000 papers mentioning 'lithium-ion battery' were indexed by OpenAlex for 2025 alone, far more than any researcher can read. This paper discusses strategies based on structured, semantic, and linked data as well as large language models (LLMs) to manage this information overload. Structured data follows a predefined, machine-readable format; semantic data includes metadata for context; linked data references other semantic data, forming a web of interconnected information. We use the battery-related ontology BattINFO, to standardise terms and enable automated data extraction and analysis. We compiled a corpus of more than 450,000 battery publications, converted manufacturer data sheets of more than 380 commercial cells into BattINFO-based JSON-LD, and linked each cell to the publications that tested it. Although only a first step, we show how this approach can support literature-based battery research, and we invite the community to contribute. We provide the structured data and the tools to access them as open source.

Keywords: Structured Data, Semantic Data, Ontology, Full-Text Search, Large Language Model

Journal Pre-proof

Journal Pre-proof

1. Introduction

Batteries are one piece of the puzzle that must be solved for the transition to climate neutrality and a sustainable future. Consequently, the battery research community is growing, and battery technology is developing rapidly. The number of research papers published each year is increasing; the OpenAlex Works API [1] returned 20,523 hits for “lithium-ion batteries”, restricting results to articles published in 2025 (filter=type:article,publication_year:2025). That is on average around 56 papers per day, including weekends. Reading literature should not be replaced by other means, but at this rate, it is impossible for researchers to manage and comprehend each article or even simply judge whether an article is relevant for them. This emphasises the need for alternative approaches to assessing the battery literature.

In other fields, such as bioinformatics, standardised structured data with ontologies (i.e., formal frameworks to ensure consistent meaning) have been established [2]. These provide examples of how structured data helps comprehension of large amounts of complex information. The Gene Ontology, developed in 2000, established a common vocabulary accessible for humans and machines [2]. It is used as the foundation for a comprehensive database that links genes and human diseases. This database structure enables easy use for a broad audience and is compatible with various analysis tools [3].

The concept of structured, semantic and linked data is equally valuable for other fields, such as battery research. This is closely related to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), which are designed to support knowledge discovery and innovation. They provide guidelines on data management with data being very broadly defined as any research output [4]. Adhering to these principles is much more challenging without a standard agreed on within the battery research community. Although we mainly create linked data, we use the term structured data more generally in this paper. Formally, we offer the following definitions:

(i) *Unstructured data* can include any file format such as text documents, pictures, web pages or social media posts. It does not conform to any standard or schema. While it often contains valuable information, it is difficult to process and to extract information from. Even if a text is written in clear English and its content is obvious to a human reader, this does not mean that useful information can be extracted from it automatically.

(ii) *Structured data* follows a predefined schema and is stored in a machine-readable format, e.g. an Excel sheet, a parquet file, a JSON file.

(iii) *Semantic data* is structured data that includes metadata providing context and meaning. This data uses a shared and machine-readable vocabulary.

(iv) *Linked data* is semantic data that refers to other semantic data and creates a web of interrelated data. An example would be a description of a cell in JavaScript Object Notation for Linked Data (JSON-LD) pointing to a JSON-LD description of the test data.

As the field of battery informatics is still young compared to other fields such as bioinformatics, the community lacks such unified procedures. This leaves the individual researcher with the challenge of identifying and filtering an overwhelming amount of data for relevant information. To address this and offer effective solutions, automation is necessary, propelled by advances in machine learning, large language models (LLMs) and structured battery data. The latter requires a clear vocabulary and a corresponding dictionary to “understand” the text of a specific field. There are different relevant standards and recommendations like the International Electrotechnical Commission IEC 60050, IUPAC Gold Book, IEEE Battery Terminology, etc.; however, these standards are focused on facilitating human

understanding with plain text descriptions. They do not provide the digital infrastructure necessary to integrate data into the Semantic Web, such as the Google Knowledge Graph and Wikidata.

Common dictionaries usually list the words of a language and give their meaning and synonyms. Ontologies add a level of complexity by capturing the relationships between the entries and aim to define a standard vocabulary without synonyms, as well as persistent identifiers (similar to DOIs for publications) for unambiguous reference. The Battery Interface Ontology (BattINFO) [5], [6], [7] is a semantic resource with essential terms and relationships to describe electrochemical systems, materials, methods, and data. As an example, Figure 1 shows the entry for 'RatedCapacity' of a battery where the capacity value is declared by the manufacturer. The entry includes a unique identifier link and an elucidation that defines the term. Furthermore, if available, a reference to an existing standard is provided. Finally, additional details can be provided in a general parsable comment.

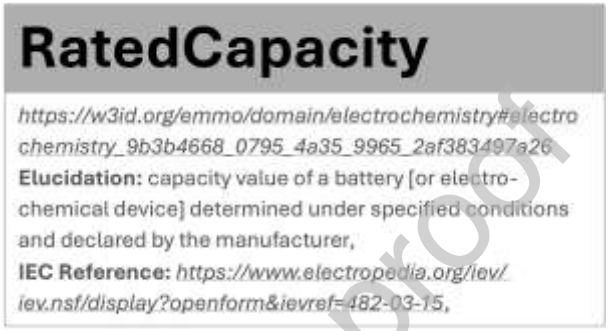


Figure 1: Data ontologies define consistent, standard frameworks for data, for example this entry for "RatedCapacity" in BattINFO. Each entry is also a node within the Elementary Multiperspective Material Ontology (EMMO) universe and in knowledge graphs, so machines can infer, for example, that it is a physical quantity with a given dimensionality that accepts a certain set of units.

Defining "RatedCapacity" may seem trivial to members of the battery community, but a common language is a prerequisite for automation. Once it is in place, subsequent steps can be accelerated and automated.

To identify relevant papers from the flood of publications, databases (e.g. Scopus, Crossref, Google Scholar, Semantic Scholar, OpenAlex) offer search engines that mostly scan the title, abstract and keywords for the search term. However, this does not enable a specific search for, e.g., the tested cells or cell chemistry, as this information is usually only stated in the experimental section. Using a text mining approach, El-Bousiydy et al. [8] compared searching only the abstract and the full text of the same data set and estimated that using only the abstract for search purposes recovers only 11% of the possible information. El-Bousiydy et al. also emphasise the problem of the lack of standardisation in data reporting. Likewise, Westergaard et al. [9] found that the text mining of full texts was consistently better than using only abstracts by analysing 15 million publications in the field of biomedicine. Thus, a full-text search is desirable. Tools like Google Scholar Labs aim to bridge this gap and are already good at the search within papers. However, they are restricted to publicly available papers and do not distinguish between publication with or without DOI. Alongside text mining from full texts, specific domain knowledge is required for meaningful data extraction. Consequently, battery-specific text mining tools have been developed [10], and LLMs combined with chain-of-thought reasoning have recently been used to mine the literature on solid-state batteries [11]. These tools can be used to make information from literature available for computational and simulation applications.

Further, LLMs can ease interaction with vast amounts of data and knowledge by using natural language to approximately retrieve the desired information. They are powerful and promising tools for accelerating research [12]; however, the alignment of human reasoning and knowledge and model

Journal Pre-proof

understanding and processing of information is challenging [13] and the tendency of LLMs to hallucinate [14] must be managed. Retrieval-augmented generation (RAG) is a promising approach as it provides the LLM with specific technical knowledge [15] and helps prevent it from hallucinating key facts. RAG combines the parametric memory (e.g. neural model weights) of an LLM generated during the training process with external information (sometimes called 'non-parametric memory'). This approach uses the capability of the LLM to "understand" natural language and its general, basic knowledge. However, instead of relying on the knowledge gained during training, the LLM is equipped with additional resources to deal with knowledge-intensive queries [16]. In this work, we do not focus on the development of LLMs, but instead use available tools and develop the data sources to augment LLMs with specific battery knowledge organised as structured data. The rapid progress in LLM development also means that any specific model is quickly outdated. Storing knowledge as structured data keeps it independent of the model, so that future generations of LLMs can reuse it.

The overall goal of this work is to advance battery research by implementing structured data methodologies that align with the FAIR principles. To this end, this work presents a methodology to create structured data and to reorganise the information presented in papers. This approach enables us to deal with the information from thousands of papers per year.

- We enhance findability through search and by linking publications and cell types. For this purpose, we use a full-text search of the battery literature to create structured data for batteries, comprising data files with information on commercial batteries which are additionally organised in a knowledge graph database for visualisation and comparison.
- We demonstrate how structured data can be complemented and accessed with the help of LLMs and thus improve accessibility and reusability via RAG for non-experts.
- We show how LLM agents can compile a source-linked literature wiki for a chosen research question

The resources and several examples are provided open source. While this work shows first results of this approach to demonstrate its potential, more efforts are necessary to make linked battery data a useful tool for more researchers and the battery community. So far, we have started the process of standardisation focusing on lithium-ion batteries, but the methodology is transferable to other chemistries and battery types.

2. Methods and implementation

The following section explains the method of creating a structured dataset that is machine- and human-readable to improve findability and reusability of the data. The focus here is on data and information about specific cells. First, we introduce the ontology and semantic data treatment as the basis for a common understanding. We then describe the publication data extraction process and corresponding structure and storage according to the FAIR principles.

2.1 Semantic data treatment

Semantic data treatment creates a metadata markup with standardised and resolvable vocabulary terms that can provide links to a shared conceptualisation and data from other sources. This approach has already been applied with great success in the field of bioinformatics [2], [3], and more generally as an exchange format for the World Wide Web. Linked data is designed to be shared through a machine-readable extension of the World Wide Web called the Semantic Web. Our approach leverages similar semantic technologies to create battery linked data that is machine-readable and compliant with the FAIR data guidelines.

Ontologies are the foundation for semantic data markup. They provide a human- and machine-readable language that formalises knowledge using a common vocabulary. Although a foundational set of ontologies exists for describing simple resources, knowledge organisation, tabular data, or popular search engine queries, there has been a lack of comparable ontologies for describing batteries. The European Union project BIG-MAP [17] developed BattINFO [5], [6] to address this need. BattINFO is a collection of relevant domain ontologies from the top-level Elementary Multiperspective Material Ontology (EMMO) universe [18], [19]. It adheres to the World Wide Web Consortium (W3C) standards for ontologies and linked data, allowing data that has been marked up with BattINFO terms to be integrated into the Semantic Web. Battery 2030+ provides support for the continued development and maintenance of BattINFO as a resource for the battery research community [20].

In this work, we implement the mapping between battery data and its associated ontology terms using JSON-LD. JSON-LD is a serialisation format for the Resource Description Framework (RDF), the standard for expressing semantic data on the web. It includes the data itself and links to individual instances and context. JSON-LD is particularly advantageous because it is human-readable and can be parsed into a machine-readable knowledge graph. We use terms from BattINFO to create linked data descriptions of battery cells and test procedures, while relying on established RDF vocabularies like Schema.org for describing general information like manufacturers, product IDs, names, etc. This approach provides two main benefits: it ensures that the data from this work is easily reusable and interoperable, and it provides an opportunity to perform semantic queries using the RDF query language SPARQL.

2.2 Data curation and repository construction

Structured data makes the information easier to find and to reuse, both for humans and machines. We now show how we create structured data based on manufacturer cell data sheets. These structured data include not only typical properties such as capacity or voltage limits but also more detailed, machine-readable information. The additional information is extracted from the full texts of battery publications.

2.2.1 Create structured data

Numerous cells are commercially available and many of them have been tested in labs. The cell data sheets provided by the manufacturers vary widely in format and content, and the relevant information cannot be extracted automatically. Thus, as the first step, these data sheets are translated into machine-readable JSON-LD files based on the BattINFO ontology. In addition, data from Teichert et al. [21] were used. Our collection is limited to cells that have a manufacturer data sheet with the basic cell information but no further data on ageing and temperature-dependent discharge curves. Presently, the repository contains over 380 files, one for each cell type. These contain typical information such as nominal capacity, voltage and temperature limits as well as the reference to where each piece of information comes from. The data in these files is additionally linked and organised in a knowledge graph, as schematically depicted in Figure 2. In this form, the data are machine-readable and can be visualised.

Journal Pre-proof

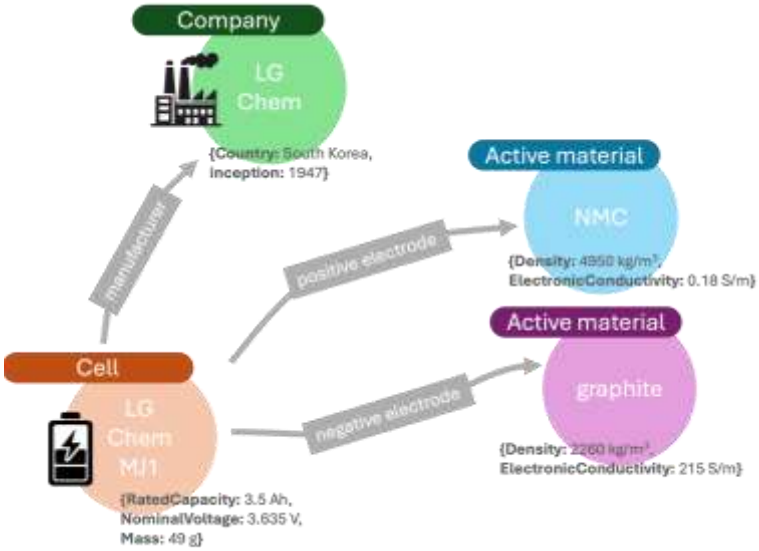


Figure 2: Complemented battery cell data sheets contain useful information that can be parsed and organised in a knowledge graph. This example shows the connections between a specific LG Chem INR18650 MJ1 cell, its manufacturer, and its active materials (nickel manganese cobalt oxide (NMC) and graphite).

Furthermore, owing to the interoperability of the format, additional user-requested information can be appended to the structured data. We included, for example, the DOIs of the papers that have tested the individual cells (see Section 3.1). An example in the Supplementary Information follows one cell from its data sheet to a linked publication, the experimental information extracted from it and the measurement data it refers to. To find and access the data, a database containing as many battery publications as possible was created. This step is explained in the following section.

Paper database

As a key goal of this work is to enable an exhaustive full-text search of battery information, a database is required containing all (or as many as possible) publications on batteries. For this work, papers published until August 2026 were considered. The list of DOIs in the corpus, together with publication year and venue, is openly available on Zenodo [22]. Thus, we targeted a paper database that aspires to be exhaustive and is significantly larger than those used in previous studies [8], [10], [23], [24], [25].

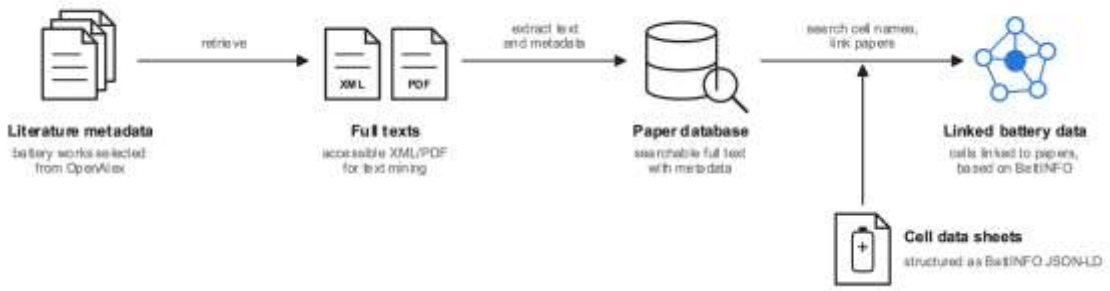


Figure 3: Information from articles was gathered and organised into a useful and searchable format by constructing a database from article text and metadata. Key structured data was then extracted and organised into a knowledge graph.

Journal Pre-proof

Figure 3 shows the flowchart of how the battery database is created. OpenAlex retrieval combined four battery-related primary-topic filters with searches for specific battery terms in titles and abstracts and for "battery" in titles. A ruleset then evaluated each work's type, title, available abstract, and primary topic, excluding specified non-research records (e.g. grants) and unrelated uses of "battery" (e.g. the legal term battery for physical force). Explicit battery terminology took precedence over topic and title matches. After resolving duplicate OpenAlex records by DOI, the August 2026 corpus contained 459,388 unique works. It must be mentioned that the dataset relies on underlying information from OpenAlex. Thus, it cannot contain publications not included in their data. To reduce the chance of errors during corpus definition and processing, we conducted spot checks on a range of around 100 papers known to the authors.

For the full-text search, the texts of more than 260,000 articles accessible to the authors for text and data mining were extracted from XML or PDF files into a database, which allows searching the entire text rather than only title, abstract and keywords. Full texts are used for non-commercial scientific research only and are not redistributed; only bibliographic metadata and derived facts, such as which cell was tested in which publication, are published.

Cell identification

To link publications to cells, each cell in the collection was searched for in the full texts by its part number. Because authors refer to the same cell in different ways, each cell was assigned a list of aliases, i.e. its part number and common variants of it, which are stored in the cell record. Matching followed a set of rules that balance missed publications against false links (Supplementary Information).

This approach can only find cells that are explicitly named within the publication. Publications that describe a cell generically (e.g. as a commercial 18650 NMC cell), cells missing from the alias list, and designations that differ from another designation by a single character are missed. Publications known to the authors that the search missed were added manually. Conversely, a cell named only in passing, for example in the reference list, is also linked to the citing publication. A link therefore means that a publication names a cell, not that it reports measurements on it (see Section 3.3).

3. Results: examples of how to use structured data

In the following, we show the potential of the proposed methods and the value of structured data for improving battery research. The following steps build on each other, yet each method offers added value on its own.

- Step 1: **Complement structured data and find information in publications via full-text search**
  With an exhaustive full-text search of the battery literature, structured data are created for batteries. This enhances findability through full-text searches and by linking the cell types with publications.
- Step 2: **Retrieve relevant information from structured data (stored as graph database)**
  Enhance accessibility and reusability with LLMs via RAG for non-experts.
- Step 3: **Agentic workflows for automated, overarching literature composition**
  An agentic workflow is implemented to condense relevant information from a subset of publications to enhance findability and accessibility.

3.1 Complement structured data and find information in publications via full-text search

Journal Pre-proof

"Which papers use cell type X?" Although this is a simple question, it can be hard to answer, as this information is not typically given in title, abstract or keywords. The retrieval of this information requires a full-text search. We did this search for all cells in our collection using all the papers in the database. To make this data available, we extended the structured cell data sheets created previously and added the references that include information about a cell in the respective data sheet to improve findability of related information.

As an example, the full-text search (Section 2.2) identified 115 papers that name the LG Chem INR18650 MJ1. This information was included in the JSON-LD file for the LG MJ1 (see file at https://github.com/phdechent/CellInfoRepository/blob/main/BatteryTypeJson/LG_Chem_INR18650_MJ1.json). This is only possible if the authors mention the specific cell type they have tested somewhere in the paper. During the process we realised that this is often not the case, and this hinders reproducibility checks, comparisons and transfer of findings. Thus, the method is only as good as the underlying information. Similarly, it is possible to search the full text of all publications for any information relevant to the researcher.

The collected data are available at https://github.com/phdechent/CellInfoRepository and are also presented at https://battdat.org. Researchers can use the "Cell Info Repository" tab on the website to filter for cell types and see where this cell type has been used in literature. This list is not yet complete, given the limitations explained above.

To illustrate the difference to general-purpose tools, we posed the same request to Google Scholar Labs and Google NotebookLM on 11 August 2026: "find papers in scientific literature that use the LG Chem INR18650 MJ1 cell". Every paper a tool returned was counted. Google Scholar Labs returned 50 papers and NotebookLM 8, compared with the 115 papers in which our search found the cell (Section 2.2). The papers that these tools found but our search did not were published after our full-text search was performed, were SSRN preprints or grey literature without a DOI, or were journal papers missing from our corpus. In this example, the limiting factor of our approach is therefore the coverage of the paper corpus rather than the retrieval step. Given the current pace of development, such tools are likely to become considerably more capable.

3.2 Retrieve relevant information from structured data (stored as graph database)

One application of structured data is in combination with an LLM to enable non-experts to query the data and receive human-friendly responses which could make the information much more accessible. For simple queries, a search term can be used for the full-text search in the database. For more complex queries, however, an LLM is a useful tool to translate search terms from natural language into specialised search query strings for the databases. For this purpose, we used RAG, a combination of an LLM and the structured data (cell data sheets) as external information (non-parametric memory), to obtain precise results and to reduce hallucination [14]. Our approach is similar to that of Buehler [26], who used an ontological knowledge graph as non-parametric memory. The relatedness of the stored data in the knowledge graph database provides advantages over regular retrieval augmentation approaches and improves LLM performance [26]. As the interaction with knowledge graphs is non-trivial and requires highly specific domain knowledge in formulating graph requests, this approach makes the knowledge graph database accessible for many more researchers and therefore increases its relevance.

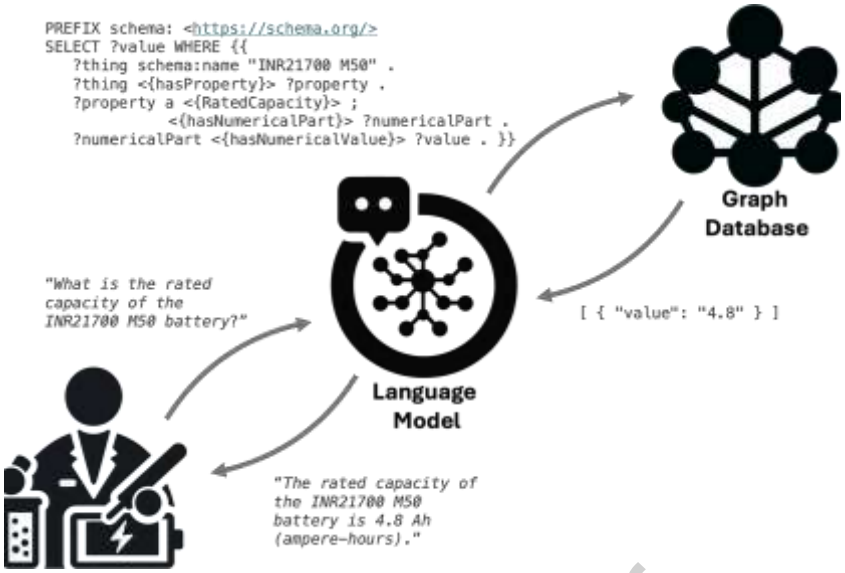


Figure 4: Battery-related example of a simple query being translated into SPARQL by the LLM and the answer being translated back into readable prose. Language models enable non-expert users to query complex structured databases and receive answers in human-readable form.

The schematic of this approach is shown in Figure 4. A researcher asks the LLM a question in natural language about the capacity of a specific cell. The LLM uses its general knowledge and a prompt that describes the knowledge graph to “understand” the query and to translate it into a search query for the graph database, which acts as external information. The database returns the result, which the LLM translates into natural language using only the retrieved facts. The RAG system can answer more complex questions than this simple example and additional examples are given in the Supplementary Information (Figures S1 and S2).

To evaluate RAG on battery data, we generated a set of 300 questions with reference answers from the cell data in the knowledge graph (available in the CellInfoRepository [27]) and posed them to three current LLMs with access to the knowledge graph database and to four LLMs without it. A second LLM graded all answers against the reference answers (details in the Supplementary Information). Figure 5 shows the outcome of all questions. With RAG, the LLMs answered 299 or all 300 questions correctly, compared with 34–48% without it. There are several methods to improve the quality of query generation of non-specialised LLMs for such questions, for example with additional pre-prompting or low-rank adaptation (LoRA), an inexpensive technique to fine-tune LLMs [28].

Journal Pre-proof

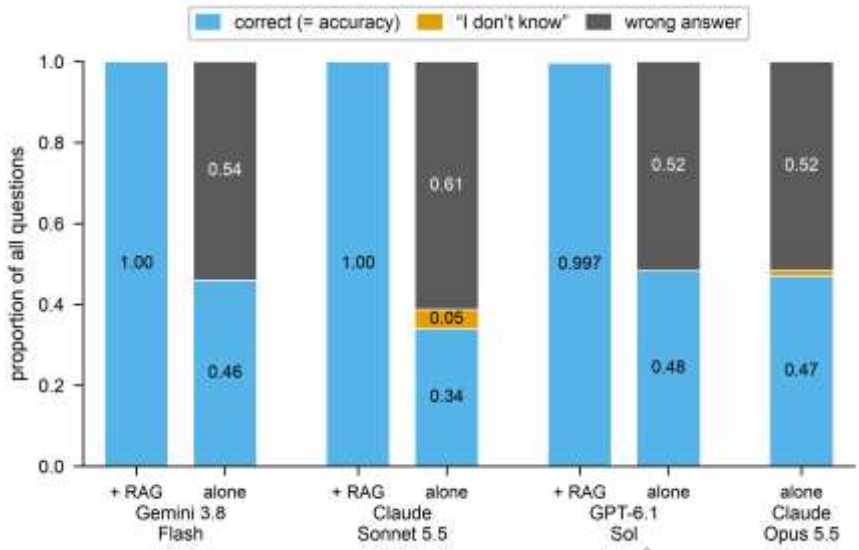


Figure 5: Adding structured data alongside a language model improves accuracy in response to battery queries. A set of 300 questions about batteries was posed to three LLMs with access to the battery graph data (+ RAG) and to four LLMs without it (alone). Outcome of all questions as proportions: correct answers (the accuracy), "I don't know" and wrong answers stated as fact.

In the LLM-only scenario, knowledge comes from the model parameters learned during training and from the question itself. For example, "What is the capacity of the A123 23 Ah cell?" can be answered from the cell name itself, and the diameter of an 18650 cell can be correctly inferred to be 18 mm. Overall, this kind of LLM-only behaviour cannot be regarded as useful since it lacks trustworthiness.

RAG clearly outperformed the LLMs alone. These results show the potential of RAG for making information findable and accessible. This could develop into a natural-language interface to large datasets that still offers a verifiable ground truth.

3.3 Agentic workflows for automated, overarching literature composition

The linked cell data make it possible to collect, for any commercial cell, the publications that report experiments on it. We tested whether an LLM agent can turn such a collection into a structured, source-linked review. The workflow follows the "LLM-Wiki" pattern [29]: instead of answering each question from scratch, the agent builds and maintains a set of Markdown pages, one per paper and one per recurring concept, that link to each other and to their sources. The agent was Claude Code with the Claude Opus 5.5 model. The agent instructions, the question template and the verification scripts are available in the CellInfoRepository [27].

We selected all journal articles published in 2025 that were identified as using at least one of the 382 cells, are open access (OpenAlex status gold or diamond) and whose full text was available. This gave 108 candidate papers. Before answering any review question, the agent read each paper in full and checked whether the authors had collected new measurements on the target cell. Papers that only reused public or previously published data, or that reported no measurements, were skipped for this evaluation since the questions were focused on experimental layout.

48 of the 108 candidates (44 %) contained no new measurements: 36 reused existing datasets and 12 reported no data at all (simulation studies, reviews, or designs based on data-sheet values). 22 of the 48 concerned state-of-charge estimation. For well-known reference cells, the number of linked papers

therefore overstates the amount of independent experimental evidence. The distinction between testing a cell and reusing its data matters whenever the linked data are used to judge how well a cell has been characterised.

The remaining 60 papers cover 34 cells; 56 report only the authors' own experiments and 4 combine own and public data. They were reviewed with a template of eleven questions on experimental and modelling practice, with a focus on thermal and mechanical boundary conditions (Table 1). Answers follow a fixed format and are based only on explicit statements; when a paper does not report something, the agent writes "Not explicitly discussed" instead of inferring it. Every answer carries a quotation of at most 30 words from the paper, and a deterministic script confirmed that all quotations occur in the source texts. An example paper note and the concept page on immersion cooling are given in the Supplementary Information.

Table 1: Review questions answered for each paper.

#	Topic	What is extracted
Q1	Study purpose and approach	Aim; experimental, modelling or hybrid
Q2	Experimental protocol	Tests performed; own vs. reused data
Q3	Ambient boundary conditions	Ambient temperature, control method, stability
Q4	Mechanical boundary conditions	Constraint type, magnitude, fixture
Q5	Cell temperature measurement	Sensor type, number, placement, attachment, sampling
Q6	Temperature-dependent performance	Effect of temperature on capacity, impedance, power
Q7	Temperature-dependent ageing	Ageing compared across temperatures
Q8	Model purpose and type	Model class; thermal component
Q9	Parameterisation and validation	Parameter sources; calibration and validation data
Q10	Spatial resolution	Lumped, 1D, 2D or 3D; inhomogeneities represented
Q11	Applicability and limitations	Validated range; limitations stated by the authors

In a second step, the agent collected recurring topics into 33 concept pages, for example multi-temperature evaluation, cell-to-cell variation and immersion cooling. The page on cell-to-cell variation, for instance, contrasts two cell-batch datasets in which capacity depends on temperature in opposite ways, and notes that the cells differ in chemistry and format, so the results are not directly comparable. The page on immersion cooling shows that the studies judge their results against different thermal-uniformity limits (a temperature difference below 5 °C versus 2 °C), so a module that meets one limit fails the other. An index page lists all papers, cells and concepts, and the Markdown files can be browsed as a linked graph, e.g. in Obsidian [30].

The agent completed screening and extraction for all 108 candidates in 20 minutes. The workflow provides a structured first pass in which every statement is linked to a quotation, so that a reader can check it directly in the paper. The agent's output therefore remains a starting point that needs expert verification. Because the review presented in this work is restricted to gold and diamond open-access papers, the complete set of paper notes, the screening log and the concept pages can be shared openly [27]. The workflow itself is not tied to this selection and can be applied to any other set of papers e.g. linked to another cell.

4. Conclusion

This work demonstrates a proof of concept for advancing battery research through structured data methodologies. We have (i) compiled a corpus of more than 450,000 battery publications, enabling exhaustive full-text searches; (ii) used structured data from cell data sheets within a knowledge graph; (iii) paired LLMs with structured battery data to facilitate advanced queries by non-experts; and (iv) developed an agentic workflow to gain a cross-paper overview of specific research questions. These efforts illustrate how battery researchers can benefit from structured data that adhere to the FAIR principles.

However, this is just the beginning, and the approach has limitations. The corpus can only contain publications indexed by OpenAlex, and the full-text search is limited to English-language texts accessible to the authors. A publication is linked to a cell only if it names the cell, and a link does not mean that the publication reports new measurements on it (Section 3.3). Individual values in the cell files can be wrong, since the data sheets differ widely in format and content. Copyright restrictions prevent us from sharing full texts; the bibliographic corpus is openly available on Zenodo [22], and the cell database on GitHub [27].

Several next steps follow from these limitations: updating the corpus and the cell–publication links automatically to keep pace with the high volume of publications, linking measurement data such as cycling or impedance datasets to the cell files, and extending the cell database to further chemistries, such as sodium-ion cells. Open dataset registries such as the Battery Data Commons [31], which also indexes field data, and curated collections of public battery field data [32], can complement the linked publications as a source of measurement data for such overviews. Recognising this as a work in progress, we invite the battery community to participate, e.g. by contributing to the cell database. Creating such structured data demands initial effort and ongoing maintenance, but it can evolve into a collaborative endeavour that benefits the entire community [33].

Looking ahead, the integration of LLMs with structured battery data opens intriguing possibilities. As LLMs become more sophisticated, they can make complex, structured datasets more accessible to a broader range of researchers, including those without specialised expertise. Advances in RAG techniques are expected to enhance the accuracy and reliability of LLM responses, reducing issues like erroneous queries and hallucinations.

Acknowledgements

Funding statement

P. Dechent and S. Paarmann are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 506629963; 511349305

Funded by the Federal Ministry of Research, Technology and Space (BMFTR), Förderkennzeichen: 03XPB029

E. Barbers, S. Clark, D. Howey and B. Planden thank Horizon Europe and UK Research and Innovation (UKRI) for funding for Digital Solutions for Accelerated Battery Testing (Digibatt) 101103997 (UKRI ref. 10038031). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Declaration of competing interest

The authors declare no competing interest.

ORCID:
Philipp Dechent 0000-0003-3041-1436

Journal Pre-proof

Elias Barbers 0000-0002-5836-2210
Simon Clark 0000-0002-8758-6109
Susanne Lehner 0000-0001-7542-9686
Brady Planden 0000-0002-1082-9125
Masaki Adachi 0000-0003-2580-2280
Dirk Uwe Sauer 0000-0002-5622-3591
David Howey 0000-0002-0620-3955
Sabine Paarmann 0000-0003-1629-6830

CRediT Author Contributions Statement:
P.D.: conceptualisation, investigation, methodology, software, data curation, writing – original draft, visualisation.
E.B.: investigation, methodology, software, data curation, writing– review and editing.
S.C.: conceptualisation, methodology, validation, software, data curation, funding acquisition writing – original draft.
S.L.: validation, resources, writing– review and editing.
B.P.: methodology, software, writing– review and editing.
M.A.: writing– review and editing.
D.U.S.: supervision, Funding acquisition
D.H.: methodology, supervision, writing – original draft.
S.P.: conceptualisation, methodology, writing – original draft.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process
During the preparation of the manuscript, the authors used Anthropic Claude and OpenAI GPT to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Both chapter 3.2 and 3.3 describe an AI enabled workflow for knowledge retrieval. The icons in Figure 2, 3, 4, S1, S2 are generated with the assistants of AI image generation (OpenAI GPT).

Data availability:
The bibliographic corpus of battery publications (DOIs, OpenAlex IDs and metadata) and its selection rules are available on Zenodo [22]. The cell data in JSON-LD format, the code and the example use cases are available at https://github.com/phdechent/CellInfoRepository [27]. The version used in this article is archived on Zenodo (doi: [Zenodo DOI to be added]). The repository will move to an institutional host before publication.

References

[1] J. Priem, H. Piwowar, and R. Orr, 'OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts', 2022, *arXiv*: arXiv:2205.01833. doi: 10.48550/arXiv.2205.01833.
[2] M. Ashburner *et al.*, 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000, doi: 10.1038/75556.
[3] J. Piñero *et al.*, 'DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes', *Database*, vol. 2015, p. bav028, 2015, doi: 10.1093/database/bav028.
[4] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, vol. 3, no. 1, p. 160018, 2016, doi: 10.1038/sdata.2016.18.
[5] BIG-MAP, 'Battery Interface Ontology (BattINFO)', BIG-MAP. Accessed: Sep. 28, 2026. [Online]. Available: https://www.big-map.eu/key-findings/battinfo
[6] S. Clark *et al.*, 'Toward a Unified Description of Battery Data', *Advanced Energy Materials*, vol. 12, no. 17, p. 2102702, 2022, doi: 10.1002/aenm.202102702.
[7] EMMO Repository, *Battery Domain Ontology*. GitHub. Accessed: Sep. 28, 2026. [Online]. Available: https://github.com/emmo-repo/domain-battery

Journal Pre-proof

[8] H. El-Bousiydy *et al.*, 'What Can Text Mining Tell Us About Lithium-Ion Battery Researchers' Habits?', *Batteries & Supercaps*, vol. 4, no. 5, pp. 758–766, 2021, doi: 10.1002/batt.202000288.
[9] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, 'A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts', *PLOS Computational Biology*, vol. 14, no. 2, p. e1005962, 2018, doi: 10.1371/journal.pcbi.1005962.
[10] S. Huang and J. M. Cole, 'BatteryDataExtractor: battery-aware text-mining software embedded with BERT models', *Chemical Science*, vol. 13, no. 39, pp. 11487–11495, 2022, doi: 10.1039/D2SC04322J.
[11] X. Zhou, X. Han, C. Li, D. Guo, and M. Ouyang, 'Research on solid-state battery information mining based on large language models and chain-of-thought reasoning', *Journal of Energy Storage*, vol. 182, p. 125065, 2026, doi: 10.1016/j.est.2026.125065.
[12] G. Lei, R. Docherty, and S. J. Cooper, 'Materials science in the era of large language models: a perspective', *Digital Discovery*, vol. 3, no. 7, pp. 1257–1272, 2024, doi: 10.1039/D4DD00074A.
[13] S. Zhang, L. Pan, J. Zhao, and W. Y. Wang, 'The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models', 2023, *arXiv*: arXiv:2305.13669. doi: 10.48550/arXiv.2305.13669.
[14] Y. Zhang *et al.*, 'Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models', 2023, *arXiv*: arXiv:2309.01219. doi: 10.48550/arXiv.2309.01219.
[15] Y. Gao *et al.*, 'Retrieval-Augmented Generation for Large Language Models: A Survey', 2023, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.
[16] P. Lewis *et al.*, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 9459–9474.
[17] BATTERY 2030+, 'BIG-MAP – Battery Interface Genome – Materials Acceleration Platform', BATTERY 2030+. Accessed: Sep. 28, 2026. [Online]. Available: https://battery2030.eu/battery2030/projects/big-map/
[18] S. Clark *et al.*, 'Semantic Resources for Managing Knowledge in Battery Research', *ChemSusChem*, vol. 18, no. 16, p. e202500458, 2025, doi: 10.1002/cssc.202500458.
[19] EMMO Repository, *EMMO – Elementary Multiperspective Material Ontology*. GitHub. Accessed: Sep. 28, 2026. [Online]. Available: https://github.com/emmo-repo/EMMO
[20] J. Amici, 'A Roadmap for Transforming Research to Invent the Batteries of the Future Designed within the European Large Scale Research Initiative BATTERY 2030+', *Advanced Energy Materials*, vol. 12, no. 17, p. 2102785, 2022, doi: 10.1002/aenm.202102785.
[21] O. Teichert, S. Link, J. Schneider, S. Wolff, and M. Lienkamp, 'Techno-economic cell selection for battery-electric long-haul trucks', *eTransportation*, vol. 16, p. 100225, 2023, doi: 10.1016/j.etran.2022.100225.
[22] P. Dechent, 'Battery literature corpus – OpenAlex, version 2026-08-30'. Zenodo, 2026. doi: 10.5281/zenodo.23021689.
[23] L. Montanelli, V. Venugopal, E. A. Olivetti, and M. I. Latypov, 'High-Throughput Extraction of Phase–Property Relationships from Literature Using Natural Language Processing and Large Language Models', *Integrating Materials and Manufacturing Innovation*, vol. 13, no. 2, pp. 396–405, 2024, doi: 10.1007/s40192-024-00344-8.
[24] S. Zhao *et al.*, 'Potential to transform words to watts with large language models in battery research', *Cell Reports Physical Science*, vol. 5, no. 3, p. 101844, 2024, doi: 10.1016/j.xcrp.2024.101844.
[25] L. Foppiano, S. Dieb, M. Zain, K. Kasama, K. Sodeyama, and M. Tanifuji, 'Construction of a Battery Research Knowledge Graph using a Global Open Catalog', 2026, *arXiv*: arXiv:2604.20241. doi: 10.48550/arXiv.2604.20241.
[26] M. J. Buehler, 'Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design', *ACS Engineering Au*, vol. 4, no. 2, pp. 241–277, 2024, doi: 10.1021/acsengineeringau.3c00058.
[27] P. Dechent, *Cell Info Repository*. GitHub. Accessed: Sep. 28, 2026. [Online]. Available: https://github.com/phdechent/CellInfoRepository

Journal Pre-proof

[28] E. J. Hu *et al.*, 'LoRA: Low-Rank Adaptation of Large Language Models', in *International Conference on Learning Representations (ICLR)*, 2022.
[29] A. Karpathy, *LLM Wiki*. (2026). GitHub. Accessed: Sep. 27, 2026. [Online]. Available: https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f
[30] Dynalist Inc., 'Obsidian', Dynalist Inc. Accessed: Sep. 27, 2026. [Online]. Available: https://obsidian.md/
[31] M. Hassini *et al.*, 'The Battery Data Commons registry', *Joule*, p. 102712, 2026, doi: 10.1016/j.joule.2026.102712.
[32] R. Masse, 'Public Battery Field Data', 2026, *arXiv*: arXiv:2610.05811. doi: 10.48550/arXiv.2610.05811.
[33] L. Ward *et al.*, 'Principles of the Battery Data Genome', *Joule*, vol. 6, no. 10, pp. 2253–2271, 2022, doi: 10.1016/j.joule.2022.08.008.
[34] N. P. Williams, D. Trimble, and S. M. O'Shaughnessy, 'Thermal performance of lithium-ion battery tabs under liquid immersion cooling conditions', *Future Batteries*, vol. 5, p. 100037, 2025, doi: 10.1016/j.fub.2025.100037.
[35] L. D. Tai, K. S. Garud, and M.-Y. Lee, 'Experimental Study on Thermal Management of 5S7P Battery Module with Immersion Cooling Under High Charging/Discharging C-Rates', *Batteries*, vol. 11, no. 2, p. 59, Feb. 2025, doi: 10.3390/batteries11020059.
[36] H. Hong, X. Song, L. d'Apolito, L. Sun, and S. Shen, 'Thermal analysis of nanofluid submerged battery tab under fast charging conditions', *Case Studies in Thermal Engineering*, vol. 73, p. 106474, Sep. 2025, doi: 10.1016/j.csite.2025.106474.
[37] S. Qaadan, A. Alshare, A. Popp, M. Tiemann, U. Spaeth, and B. Schmuelling, 'Data on battery health and performance: Analysing Samsung INR21700-50E cells with advanced feature engineering', *Data in Brief*, vol. 59, p. 111346, 2025, doi: 10.1016/j.dib.2025.111346.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Ethics declaration

The author(s) declare(s) that the study does not involve humans nor animal subjects.

Journal Pre-proof