Voltaiq data format—A standard data format for collection of battery data to enable big data comparisons and analyses across the battery lifecycle

Christianna N. Lininger*, Tony Thai, Taylor R. Juran, Eli S. Leland and Tal Z. Sholklapper

Voltaiq, Cupertino, CA, United States

Batteries have enabled modernization of society through portability of electricity. Batteries are also a crucial component to enabling clean technologies of the future such as grid storage and electrified transportation. Because of their ubiquity in modern society, global organizations develop and commercialize batteries for their electrified products. Across the field of battery development, in both commercial and academic settings, there is broad utility in standardization of data formats amongst disparate data sources, labs, equipment, organizations, industries, and lifecycle phases. Due to the way the nascent industry developed, there is a lack of standardization for how performance data is recorded, which is now hindering the industry's ability to learn from data and accelerate growth. Herein, we describe the different types of data, formats, conventions, and standardization for each phase in the battery lifecycle. Next, we provide a standard data format and conventions for the community to either utilize in their data collection practices or map their existing data into: the Voltaiq Data Format (VDF). This standard data format provides the flexibility needed to capture the variability in data formats and conventions along the battery lifecycle. The utility of this standard format aids in collaboration within and across organizations, accelerating innovation across the industry, and paves the way for the battery community to start utilizing the power of machine learning and data science.

KEYWORDS

lithium-ion battery, data, machine learning, electrification, standards, electric vehicle, best practices

Introduction

Conventions and standardization in science and engineering

Conventions enable groups of scientific researchers and engineers to rapidly communicate concepts, experimental results, and information with shared understanding. Some conventions are tradition, some are explicitly developed to further understanding or communication of concepts, while other conventions can be arbitrary. Examples of conventions in the field of electrochemistry are the signing of current to reflect the movement of electrons and the signing of the transfer coefficient in the exponential kinetic terms in the Butler-Volmer equation(Moran and Gileadi, 1989; Guidelli et al., 2014), the Standard Hydrogen Electrode has been chosen as the reference standard of choice for the definition of potential (Isse and Gennaro, 2010; Matsui et al., 2013). Another convention is the shorthand description of a Galvanic cell, where the anode chemistry is written out first on the left-hand side (Garnett and Treagust, 1992).

Standardization and the battery industry

The modern lithium-ion battery was developed in the 1970s–1980s and is attributed to John B. Goodenough, Stanley Whittingham, and Akira Yoshino (Tyutyunnik, 2021). The first reported commercialization of the lithium-ion battery is attributed to Sony in 1991 (Reddy et al., 2020). As such, the consumer electronics industry was the first to rapidly adopt and integrate the technology commercially, and was the primary commercial driver for the early development of lithium-ion batteries (Brodd, 1999). Other markets, such as automotive, have had less linear pathways toward electrification, in part due to the availability and cost of competing energy sources. The concept of the electrified vehicle (EV) has been around since the 1890s, however EVs have had varying degrees of commercial success (Santini, 2011). The 1996 EV-1 from General Motors being the most well-known recent commercialization attempt (Johnson, 1999) prior to the present-day commercialization of the electric vehicle that looks like it is here to stay (Global EV, 2019; Bartlett and Preston, 2022). Lithium-ion batteries did not come into commercialization in an organized manner; the technology is leveraged in various application spaces, and each application space has variable testing practices, testing hardware, and measurement practices. This has led to nascent state for standardization in the battery lifecycle across the commercial battery industry.

Announcements from leading automotive companies (Bartlett and Preston, 2022), emerging government regulations (European, 2015; Yang and Rutherford, 2019; California, 2022; Electric Vehicle Toolkit, 2022), the decreases in cost per kWh,

and improvements to the underlying lithium-ion technology are consolidating the dominance of the lithium-ion battery in the ongoing electrification of the automotive industry (Berdichevsky and Yushin, 2020). As this transition accelerates, however, legacy OEMs and the vertical markets that support them are struggling to keep up. To meet the demands that automotive is placing on the battery field at large, innovations and a change in the status quo will be necessary.
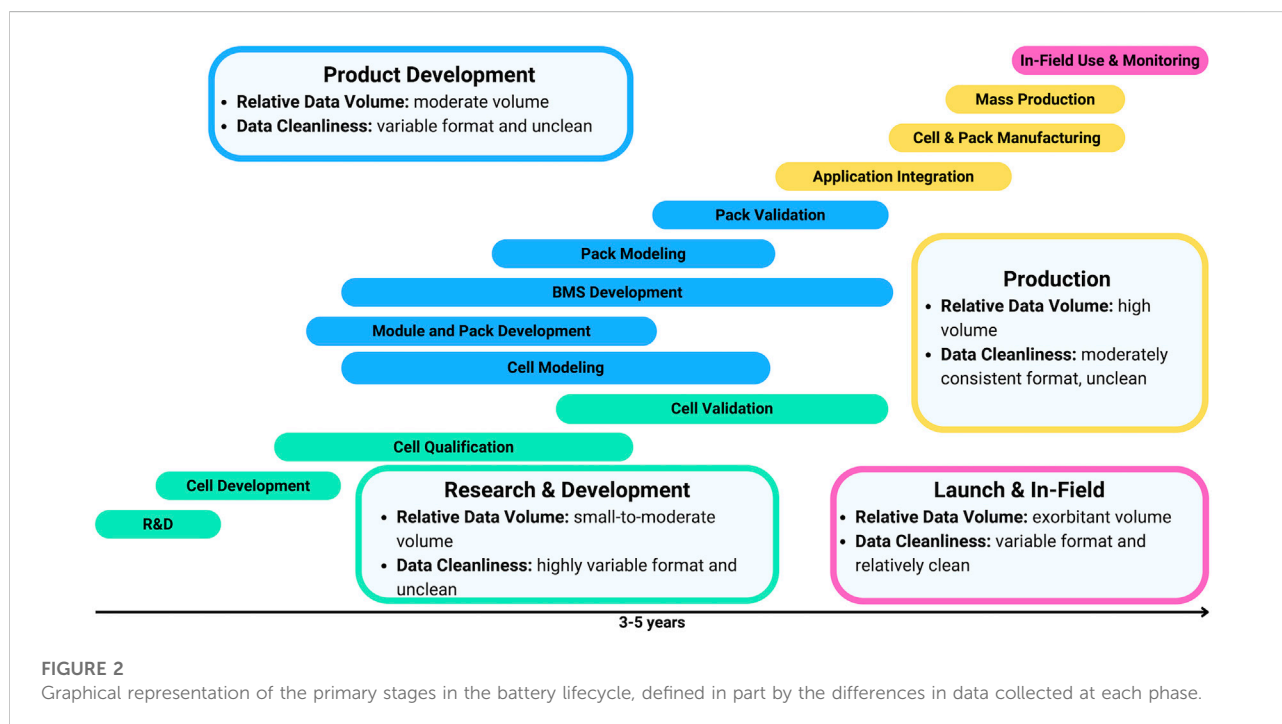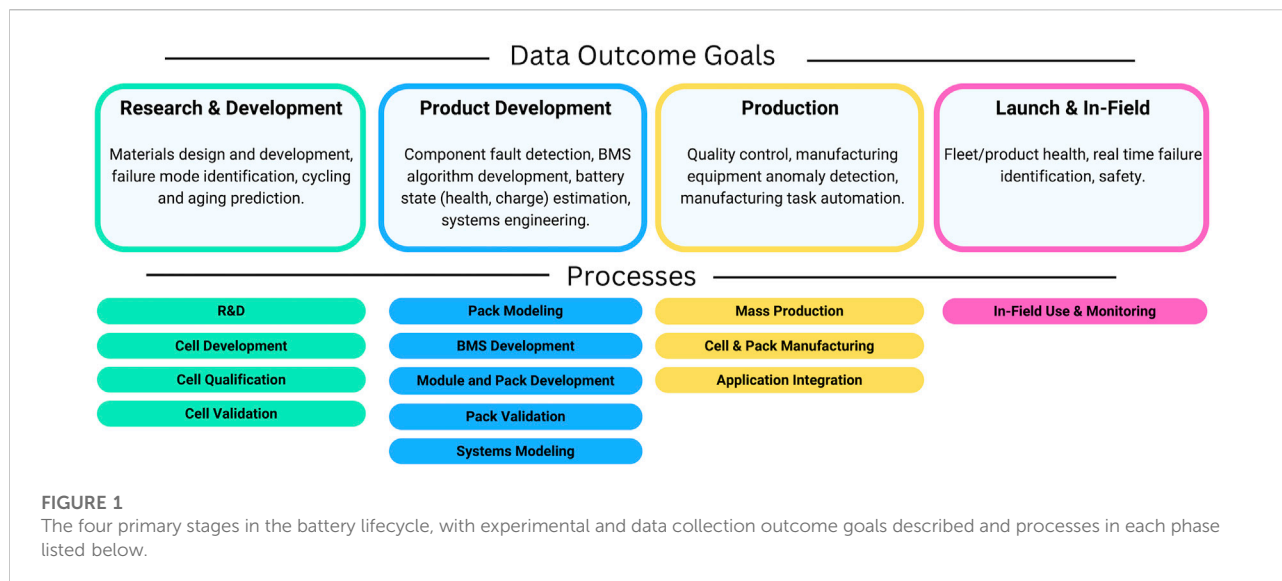
Due to the varying paths batteries have taken into the commercial realm, and the disparate application spaces that the technology is leveraged for, standardization is lacking across much of the battery industry. In this publication, we focus specifically on the standardization of data collection, data formats, conventions for definition of data traces. For batteries specifically, data standardization can enable traceability of materials, cells, systems, and defects across an entire lifecycle; data standardization can also enable meaningful collaboration and comparisons, ultimately accelerating innovation.

Data science, machine learning, and battery data

Battery development accrues large volumes of data due to the extensive testing required to bring battery technology to market for an electrified product. In very early research stages, experimental protocols may be one-off, but broadly across the battery lifecycle, the experimental outcome goals and the types of experiments are well defined and are shown in Figure 1. The amount of testing required to realize a new battery technology in an electrified product can span years. At the time of writing, the industry standard is 9–18 months for consumer electronics and 3–7 years for automotive applications, as shown in Figure 2. These realities in battery development lend to large volumes of data from repeated tasks, a scenario where machine learning is effective at providing insight. Machine learning is also effective where insights are not obvious to humans or large volumes of data prohibit human insight.

Data standardization can pave the way for the use of data science and machine learning driven innovations in the battery development industry. The standardization of conventions and definitions for data are the first step to any data science or machine learning project, where "data cleaning" as it is referred to colloquially amongst data scientists, is known to comprise a substantial portion of the effort in any large-scale data science project (Chu et al., 2016; Kumar and Khosla, 2018; Petrova-Antonova and Tancheva, 2020; Wang and Wang, 2020; Ilyas and Rekatsinas, 2022).

A prerequisite for data science and machine learning applications is standardization and adherence to colloquial conventions, there are realities in battery research and development that make standardization difficult. There are

FIGURE 1
The four primary stages in the battery lifecycle, with experimental and data collection outcome goals described and processes in each phase listed below.

FIGURE 2
Graphical representation of the primary stages in the battery lifecycle, defined in part by the differences in data collected at each phase.

widely varying chemistries, measurement equipment, and end use cases for the battery that could cause a researcher to test or generate data with varying conventions and definitions. In the case of test protocol standardization, standards development organizations (SDOs) have developed evolving protocol recommendations for battery testing (United States Council for Automotive Research, 1996; Conover, 2016; Blair, 2021). Governmental standards development organizations such as the United States National Laboratories have produced the USABC Manual for automotive

battery testing (United States Council for Automotive Research, 1996), the Protocol for Uniformly Measuring and Expressing the Performance of Energy Storage Systems (Conover, 2016), the Global Overview of Energy Storage Performance Test Protocols (Blair, 2021), and the Battery Test Manual for Electric Vehicles (Christopherson, 2015a; Christopherson, 2015b), to name a few. Standards for testing have also been developed outside of the U.S., for example the Chinese GB/T32960.2–2016 regulation and European testing standards.

In addition to test protocol standardization, another difficulty to standardization is that the battery data itself, and methods of accessing it, are beholden to the formats and conventions inherent to the brand of tester being used to collect the data. Today it is common to walk into a battery test lab and find two or more brands of testers. Each brand of test hardware that collects battery data has its own methods, conventions, and definitions for the collected data. For example, the definition of signed current for charging and discharging is inconsistent across tester companies, while some hardware companies do not sign current at all. In the calculation of capacity, some hardware companies calculate capacity per charge or discharge state, while others simply accumulate capacity over the entirety of the test.

The source of these variations in definition and convention can be better understood through the origin of the commercial hardware brands. The testers used broadly today in the research and development phase of the battery lifecycle originated regionally in small businesses with chemistry or application specificity. For example, Arbin, Maccor, and Bio-logic originated with research and development testing. Bitrode and Digatron started in the lead-acid industry. Neware, PNE and Toyo started as region-specific brands in Asia. Different tester types have differing capabilities—Bio-Logic, for example, is capable of extensive electrochemical testing methods that other tester brands are not capable of. However, due to the chemistry and region-specific development of the field of battery testing, the conventions and definitions for the common traces amongst the tester brands are lacking. In the past, this was less of an issue, as the scale of battery testing was smaller, and availability of hardware was more regionally dependent. A current hindrance to innovation, traceability, and meaningful comparison in the battery field is the standardization of the battery data itself, the format of the data, and the definitions for the traces within the data.

Publication overview

In the field of science and engineering, it is well accepted that standardization and conventions are powerful tools that we have as a community to enable rapid and clear communication of concepts and results, accelerate time to innovation, and enable establishment of best practices for a field of study. The realities of battery development (large volumes of data from repeated experiments that result in prohibitively large or complex sets of data) create a data space that is fit for the insights that machine learning and data science techniques can offer. Standardization in both experimental protocols as well as data formats are necessary prerequisites for broader collaboration and comparison. These standardization practices must be flexible enough to accommodate the some variability in data collection methods, data formats, and experimental protocols.

In this publication, we strive to provide standardization and conventions for the format and datafiles that result from battery testing. We provide an open-access publicly available common format, the Voltaiq Data Format (VDF), with defined conventions and standards that battery data collected across the lifecycle can be mapped to. Given that Voltaiq handles data across every step of the battery lifecycle today, and that Voltaiq has worked closely with dozens of customers and equipment vendors to standardize formats over the last decade, we believe we are uniquely qualified to present a format that is flexible enough to handle the variations throughout the development process while also maintaining the standards and conventions that enable us to communicate more effectively across functions. Additionally, this format has been used in industry for the better part of the past decade to collect data in research and development, product development, production, and in-field battery operation. As such, the data format has accommodated variability across each of these sub-processes in the battery lifecycle. In the following sections, we provide an overview of the battery lifecycles that Voltaiq Data Format has already been used in and discuss the variances in data formats, collection methods, and volumes at each phase. The goal of providing this standard data format is to provide a convention across industry and academia. If data is mapped to this format and the conventions within the format are followed, large volumes of battery data can be collected and leveraged. To our knowledge, this is the first release of a standard battery data format that has been used extensively in the commercial battery development space. This set of standardized data can be utilized for innovation, collaboration, meaningful comparison, and with data science and machine learning techniques to provide insight into some of the most challenging questions the community is addressing today.

Results

Types of battery data

We have categorized the types of data collected for batteries into three categories: time-series data, throughput data, and metadata.

- time-series data is any data that is collected as a function of time;
- throughput data is any data that is collected as a function of some unit of throughput of the battery;
- metadata is data that is collected about the battery or about the other data sources.

Time-series data is a sequence of data points ordered by time. For example, current and voltage measurements are two types of time-series data. Other examples of time-series data are discharge capacity, temperature, or power. Measurements taken with

auxiliary metrics such as thickness changes of the cell with cycling, or temperature measurements of the cell or the cell environment are also time-series data.

Throughput data is calculated as a function of a defined measure of throughput; this unit of throughput varies based on the application space or battery development phase. For example, in the early cell development phase, performance metrics are calculated as a function of 'cycle' where cycle is the unit of throughput measure. In automotive applications, traditional full charge and subsequent discharge is not always a useful unit for throughput performance—instead, performance as a function of a unit of 'drive cycle' is more appropriate (Lawder et al., 2014; Tourani et al., 2014; Jafari et al., 2015; Baure and Dubarry, 2019). In grid storage applications, duty cycles are developed which match the energy provision the battery supplies for the grid (Rosewater and Ferreira, 2016; Moy et al., 2021).

Metadata gives context to the time-series and throughput data as the metadata is any data that describes the battery, time-series data, or the throughput data in more detail. For example, the weight, active material chemistries, manufacturing lot and date, or number of cells in series or parallel would all be metadata applicable to different phases of the battery lifecycle. Other metadata examples include the operating conditions during testing, tester hardware, software versioning, the test protocol, the operating system, the test operator, and so forth. These are all pieces of information that provide additional context to the time-series and throughput data.

Voltaiq data format

Working across much of the battery lifecycle, Voltaiq has developed and provided open-access to a data format for battery data collection which is structured enough to employ as a standard across the industry but flexible enough to support the variability inherent to the different phases of the battery lifecycle outlined in the previous section. Voltaiq Data Format (VDF) is purposefully similar to the data format a battery tester may write data, however, VDF provides one standard set of definitions and conventions for the data and the data format, therefore data collected with varying conventions and formats may be mapped to VDF. As such, VDF is designed to capture data from research and development all the way through production. In-field data capture is an emerging space, we propose here guidelines on how to modify Voltaiq Data Formation for in-field data. Additionally, we open-source our guidelines and invite the community to evolve the guidelines into standards that can be adopted and accepted as best-practices across the community.

Voltaiq Data Format files have two main sections, a header and a body. The header consists of metadata and the body consists of the time-series data. The throughput data can be calculated from the time-series data with the conventions outlined in the following sections. Example files can be seen in the Supplementary Information Section S1.1.

The voltaiq data format requires the following formatting

- Datafiles are in CSV format with a "Tab" delimiter.
- Each datafile represents one and only one test—datafiles will not contain data from multiple tests. However one test can be written in multiple datafiles, if needed, due to file size limitations.
- All datafiles have a unique file name.
- Certain metadata and time-series entries are required, as they comprise the bare minimum entries needed to calculate the full set of performance data. Other fields are optional or recommended as specified below.
- The metadata section and the time-series section must be separated by the string "[DATA START]".
- Every time-series data column must have units defined.
- The number of data columns must be consistent throughout a file and match the number of Data Header columns.

Metadata header

Each datafile should begin with a specially formatted Metadata Header which can be any number of lines, in which each line contains a single "key:value" pair representing one piece of metadata (with a ":" delimiter). There are a set of required fields, but any quantity of metadata fields can be included in the header, up to 1024 "key:value" pairs. Metadata associated with the testing conditions is recommended to be included in the Voltaiq Data Format datafile header, while a broader set of metadata is recommended to be captured separately as discussed in the subsequent section "Metadata File". A list of recommended metadata for the test file header is included in Table 1. The termination of the header is indicated by a line containing only the string "[DATA START]". Example files can be seen in the Supplementary Information Section S1.1.

Large format energy storage systems are comprised of multiple cells electrically connected in series and parallel configurations to create a battery that can meet a performance requirement that a single cell cannot meet alone. There are metadata headers, time-series data, and metadata entries that are recommended specifically for these large format energy storage systems composed of more than one cell, as shown in Figure 4. For these systems, it is important to specify the hierarchical relationship between the different electrochemical components (cell, module, pack); for example, a pack could have four modules, with each module consisting of 10 cells. In the above example, the convention in VDF is that the highest hierarchical level is the pack, while the lowest hierarchical level is the cell. Additionally, it is also important to specify the electrical connection between components; for example, a module of 10 cells could consist of two series connections of five parallel-connected cells.

TABLE 1 Exemplary Metadata header entries.

Metadata header	Required	Definition	Convention
Start Time	Required	UTC timestamp for the start of this test. Either of these formats is allowed	The VDF convention is that epoch is specified in milliseconds; e.g. record 1577836800000 for 12 AM 1 Jan 2020 GMT. If using the "Date and Time" Unit, record with ISO 8601 standard: ("yyyy-MM-dd'T'H:mm:ss'Z'")
		Integer (epoch) or String (Date and Time)	
		Absolute (date) time datapoint	
Timezone	Required	The timezone where the test is being run. Either of these formats are allowed	
		International timezone format ("America/New_York")	
		UTC offset ("-4:00")	
Channel Number	Recommended	Which equipment channel number (or similar identifier) was used to run this test	
Tester ID	Recommended	An identifying string unique to the tester running the test across the organization	Examples for this might be appending on a lab location to a tester name: "SF_Maccor1" or "sf_xyz123" where xyz123 is the serial number for the tester
Device ID	Recommended	An identifying string unique to the device/cell/pack being tested across the organization	
Format Version	Recommended	A version string that changes when the file format (required column list, etc.) changes	
Metadata Filenames	Recommended	A list of string filenames for any additional files containing relevant metadata for the test	
Procedure Name	Optional	Unique name of the tester protocol used to run the test	
Description	Optional	Non-unique description of the test	
Nominal Capacity	Optional	The nominal capacity for the device being tested, specified in "# Ah"	Units must always be included alongside measurement metadata. Format is "{#} Ah", where "#" is replaced with the measurement. Supported units are specified in Supplementary Information
Project	Optional	The project that the device being tested belongs to	
Server Version	Optional	Version of Tester Server Software	
Client Version	Optional	Version of Tester Client Software	
Test Protocol ID	Optional	File name of the test protocol	
Tester Calibration Date	Optional	UTC date for the tester calibration date for the test hardware. Either of these formats is allowed	The VDF convention is that epoch is specified in milliseconds; e.g. record 1577836800000 for 12 AM 1 Jan 2020 GMT. If using the "Date and Time" Unit, record with ISO 8601 standard: ("yyyy-MM-dd'T'H:mm:ss'Z'")
		Integer (epoch) or String (Date and Time)	
		Absolute (date) time datapoint	
Comment	Optional	Any comment on the test file	
Pack Hierarchy Levels	Required for large format storage data	The number of hierarchy levels that exist within the pack	The number of distinct electrochemical entities in the storage system, for example a pack with modules composed of cells would have '3': cell, module, pack
Pack Hierarchy Naming Schematic	Required for large format storage data	Format in a high to low hierarchy	The naming for each distinct electrochemical entry from high to low hierarchy, comma separated. For example: "pack, module, cell"
Number of Components per Hierarchy Level	Required for large format storage data	Format in a high to low hierarchy	The number of each entity, comma separated. For example, 1 pack, 9 modules per pack, 12 cells per module "1, 9, 12"

TABLE 1 (*Continued*) Exemplary Metadata header entries.

Metadata header	Required	Definition	Convention
Connection Type	Required for large format storage data	A string (ie. "Series" or "Parallel"). For multi configurations, _s_p is acceptable. Format in a high to low hierarchy	The connectivity for each electrochemical entity should be specified in high to low hierarchy order. For instance 1, 3s3p, 4s3p
{Pack} Voltage Trace	Required for large format storage data	The name for voltage data traces for the pack level of storage system hierarchy. A string formatted in a high to low hierarchy, separated by commas	For example Pack_Voltage, Mod01_Cell05_Voltage, Cell18_Voltage
{Pack} Current Trace	Required for large format storage data	The name for current data traces within a given level of storage system hierarchy. A string formatted in a high to low hierarchy, separated by commas	For example Pack_Current, Mod01_Cell05_Current, Cell18_Current
{Pack} Temperature Trace	Recommended for large format storage data	The name for temperature data traces within a given level of storage system hierarchy. A string formatted in a high to low hierarchy, separated by commas	For example Pack_Temperature, Mod01_Cell05_Temperature, Cell18_Temperature Sensor numbers can be augmented to the end of the trace name. Temperature can be concatenated to "Temp"

Time-series data

After the Metadata Header and the "[DATA START]" line, the remainder of the file should contain a data header followed by the time-series data in columnar, tab-separated format. The minimum time-series data required in the file is Test Time, Current, and Voltage; from these three traces, all other data can be calculated or inferred alongside the metadata. For example, the capacity can be calculated as the integral of the current over time, the power can be calculated as the voltage multiplied by the current. Where possible, we recommend always including the measured value. The complete list of time-series data traces are included in Table 2. For any additional time-series data that are not explicitly included in Table 2, the columns should be included in the file as auxiliary traces, named "Aux. {Trace Name}", where {Trace Name} is replaced with a descriptive name for the time-series data. Example files can be seen in the Supplementary Information Section S1.2.

The data header for the time series data consists of a tab separated row with entries according to the "Trace Name" column in Table 2. Below the Trace Name row is a second tab separated row with the units for the trace in that column. The third row is then the raw data. The accepted traces are listed in Table 2, the accepted units are listed in the Supplementary Information Section S2.1.

In the case of the product development phase, modules and packs are designed and developed into large format storage systems as shown in Figure 4. Data is collected on the pack, module, and cell levels and clear designation of the origin of the measurement data needs to be specified in the data. The origin for all of the individual measurements in

Table 2 should be specified in augmented trace names. For instance, Voltages coming from cells 01, 05, and 11 would be labeled as below:

"Cell01_Voltage".

"Cell05_Voltage".

"Cell11_Voltage".

If there was an energy storage system with more than one module in the configuration, the module would need to be identified as well, labeled as below:

"Mod01_Cell01_Voltage".

"Mod01_Cell05_Voltage".

"Mod02_Cell11_Voltage".

In general, each measurement should be labeled with a mapping to the full configuration in the energy storage system. The convention is to list the hierarchy from high to low, as shown in the example above.

Also in the case of the product development phase, data originating from the BMS may be recorded. This data is to be treated as Auxiliary data and collected according to columns named "Aux. {BMS Trace Name}", where {BMS Trace Name} should be the name of the data field as written by the BMS system. We recommend that where users can specify their BMS trace names, they include the configuration mappings written above which clearly indicate the identification of the trace source, for example "Mod01_Cell01_Voltage".

Voltaiq metadata file

In many battery testing scenarios, the amount of metadata that needs to be collected is on a larger scale than what is reasonable to include within the header of the time-series data

TABLE 2 Time Series data trace definitions, logical requirements, conventions, and supported units.

Trace name	Required	Definition	Logical requirements	Convention
Test Time	Yes	Float Time that has elapsed since the start of the test	Sequential values within the test may not decrease (i.e. values should be in ascending order)	
Current	Yes	Float Instantaneous value of current		The sign convention is positive for charge current and negative for discharge current
Voltage	Yes	Float Instantaneous value of voltage		
Datapoint Number	Recommended	Integer Datapoint number of the test	Must start at 1 and strictly increase without duplicates (e.g., 1,2,3,4 …)	If a test spans multiple datafiles, the datapoint number is representative of the number belonging to the test and does not start back to 1 at the start of a new datafile
Cycle Number	Recommended	Long Cycle number of the test	Must start with 1 and strictly increase by 1 each cycle, without duplicates. There cannot be any gaps or jumps	If a cycle column is not observed, the default algorithm will look for the first charge datapoint after any discharge datapoint as the boundary for a new cycle
Timestamp	Recommended	Integer (epoch) or String (Date and Time) Absolute (date) time datapoint	Sequential values may not decrease (i.e. values should be in ascending order)	VDF convention is that epoch is specified in milliseconds; e.g. record 1577836800000 for 12 AM 1 Jan 2020 GMT. If using the "Date and Time" Unit, record with ISO 8601 standard: ("yyyy-MM-dd'T'H:mm:ss'Z'")
Step Type Index	Recommended	Long Test procedure step type number		When a test procedure is set up, step types are specified and numbered. A constant charge at a specified value of current until a voltage limit is reached is an example of a step type. This step type is numerically indexed. Each step type index should belong to one and only one step type
Step Time	Recommended	Float Elapsed time since the start of the current step		This value resets to zero anytime the Step Index changes Sequential values within a step may not decrease (i.e. values should be in ascending order)
Charge Capacity	Recommended	Float. Cumulative charge capacity of the current cycle	Values must be greater than or equal to zero	This value must reset to zero at the beginning of every cycle (including beginning of the test), and should not reset at step changes within a cycle The sign convention for charge capacity is always positively signed Over a cycle, charge capacity accumulates. Sequential values within a cycle may not decrease (i.e. values should be in ascending order)
Discharge Capacity	Recommended	Float Cumulative discharge capacity of the current cycle	Values must be greater than or equal to zero	This value must reset to zero at the beginning of every cycle (including beginning of the test), and should not reset at other step changes within a cycle The sign convention for discharge capacity is always positively signed. Over a cycle, discharge capacity accumulates. Sequential values within a cycle may not decrease (i.e. values should be in ascending order)

TABLE 2 (*Continued*) Time Series data trace definitions, logical requirements, conventions, and supported units.

Trace name	Required	Definition	Logical requirements	Convention
Charge Energy	Recommended	Float.Cumulative charge energy of the current cycle	Values must be greater than or equal to zero	This value must reset to zero at the beginning of every cycle (including beginning of the test), and should not reset at other step changes within a cycle
				The sign convention for charge energy is always positively signed
				Over a cycle, charge energy accumulates. Sequential values within a cycle may not decrease (i.e. values should be in ascending order)
Discharge Energy	Recommended	Float	Values must be greater than or equal to zero	This value must reset to zero at the beginning of every cycle (including beginning of the test), and should not reset at other step changes within a cycle.The sign convention for discharge energy is always positively signed.Over a cycle, discharge energy accumulates. Sequential values within a cycle may not decrease (i.e. values should be in ascending order)
		Cumulative discharge energy of the current cycle		
Power	Recommended	Float.Instantaneous value of Power		The sign of power should align with the defined sign convention of current
Aux. {Identifier}	Optional	Float		Any time series data not included in previous columns can be recorded as Auxiliary traces, named appropriately in {Identifier} An example would be "Aux. Chamber Temperature"

testing file. This is specifically relevant to the research and development and production phases. In these scenarios, we recommend a metadata file in CSV format with metadata collected on a per-device basis. This metadata file should be linked to any testing file for that device through the metadata header as outlined in Table 1. We have listed additional examples of metadata that would be included in the metadata file in Tables 3, 4. The format should be as follows:

- Each row represents a separate device.
- Each column is labeled with the name of the metadata entry at the top of the column.
- Following each metadata entry column, a separate column is included explicitly specifying the units of measurement for the subsequent metadata value.

Example files can be seen in the Supplementary Information Section S1.2.

In-field data collection

Due to the emergence of in-field battery data collection, we do not propose a full set of standards for in-field data collection. We do however provide recommendations on how to augment

the above standards for in-field data. Additionally, we open-source our guidelines and invite the community to evolve the guidelines into standards that can be adopted and accepted as best-practices across the community.

The collection of the time-series data should match the conventions outlined for large scale energy storage systems in the product development phase in the "Time-Series Data" section. Namely, each battery time-series measurement should be labeled with a mapping to the full configuration in the energy storage system. The convention is to list the hierarchy from high to low, i.e. pack number > module number > cell number. The following additional guidelines for time-series in-field data are below".

- Datafiles are in CSV format with a "Tab" delimiter.
- Each datafile represents a continuous operating period—datafiles will not contain data from multiple operating periods. However one operating period can be written in multiple datafiles, if needed, due to file size limitations.
- All datafiles have a unique file name.
- Every time-series data column must have units defined in the datafile or in an accompanied datafile that is linked to the time-series data through filename or metadata header.

TABLE 3 Metadata to include in designated metadata file by battery lifecycle phase.

Research & development	Product development, production & manufacturing
Area	Number of Cells in Series
Cell Mass	Number of Cells in Parallel
Active Material Mass	Pack configuration
Nominal Capacity	Battery Size Factor (BSF)
Batch Number	Max Charge Voltage (Pack)
Batch Name	Gas Gauge Model Number
Batch Manufacture Date	Gas Gauge Model Name
Supplier Part/Model Number	Pack on Frame?
Supplier Part/Model Name	Supplier Name (Pack)
Supplier Name	Batch Manufacture Date (Pack)
Device Form Factor	Module Number
Device Type	Origin
Device Chemistry	Installation Volume
Device Chemistry Common Name	C/10 Capacity
Device Nominal Voltage	C/10 Energy
Device Weight	Keywords Cell
Device Length (without terminals)	Lower Voltage Limit
Device Width (without terminals)	Peak Current
Device Thickness (without terminals)	Continuous Current
Device Diameter (without terminals)	Peak Power
Device Volume	Continuous Power
Initial OCV	Gravimetric Energy Density
Initial ACIR	Volumetric Energy Density
Max cell charge voltage	Gravimetric Power Density
Cell thermal cut off Project Code	Volumetric Power Density
Observed Li Plating	Lower Temperature Limit
N/P Ratio	Main thermal sensor
	Radius of Pole
	Length of Pole in x-direction
	Length of Pole in y-direction
	Length of Pole in z-direction
	Housing Material
	Housing Thickness
	Housing Weight
	Housing Polarity
	Windingcore Mass
	Jellyroll Mass
	Number of jellyrolls
	Number of windings
	Electrode Manufacturing Equipment ID
	Slurry Mixing Equipment ID
	Charge Rate
	Discharge Rate
	CV Steps
	Applied Pressure

- The number of data columns must be consistent throughout a file and match the number of data header columns.

The metadata in Table 5 is recommended for in-field data capture in the metadata header section of the time-series data file, or in a separate file that is linked to the time-series data through filename or metadata header.

Discussion

The battery lifecycle

Battery engineering in industry is cross-disciplinary in nature. To bring a new technology to market in industry, the technology goes through a series of steps spanning multiple engineering teams, time scales, and development processes. Figure 2 is a graphical representation for moving from a cell to an electrified product, representing the multiple cross-disciplinary phases. We have broken up the battery lifecycle into four primary stages representing sequential phases of development: Research and Development, Product Development, Production, and Launch and In-field. In Figure 3, we illustrate the relative volumes of data collected at each stage represented in Figure 2. These phases are, in part, defined by the differences in data collected at each phase: structure, format, conventions, metadata, and auxiliary data, as well as data volumes. Voltaiq Data Format has already been leveraged in industry to standardize battery data collection in each of these primary stages in the battery lifecycle. In each subsequent section, we describe the data collected at each phase and discuss how Voltaiq Data Format can be leveraged.

Research and development

In the Research and Development phase of the battery development lifecycle, data is likely to be collected on the half- or full-cell level as the basic chemistries of the active components in the cells or design criteria on the cell level, such as form factor, are being evaluated. The nature of the data is traditional electrochemical measurements, such as time-resolved current and voltage and all of the derivative measurements made from current and voltage such as capacity, energy, power, etc. Additionally, measurements as a function of throughput are made. In a cell, throughput is generally defined by the throughput during a full charge and discharge of a cell, considered a full 'cycle'. This time-series data is well captured by the conventions and standards outlined in the "Time-Series Data" section.

The types of metadata collected in the research and development phase are normally cell-specific, such as the identity of the active materials, the dimensions of the cells and internal structure, masses of active materials, safety

TABLE 4 Metadata to include in designated metadata file by cell component.

Cathode	Cathode ID
	Cathode material
	Cathode Mass
	Cathode Area
	Cathode Porosity
	Cathode Carbon lot
	Cathode Binder lot
	Cathode type
	Cathode Loading
	Cathode Build date
	Number of Cathode Layers
	Cathode temperature during Compression
	Cathode Water/Solvent ratio
	Cathode slurry mixing time
	Cathode Binder amount
	Cathode edge vs. Center
	Cathode Conductive Additive type
	Cathode Conductive Additive amount
	Cathode Conductive Additive Provider
	Stoichiometry Cathode
	Material current Collector Cathode
	Thickness current Collector Cathode
	Cathode Jellyroll Mass
	Cathode complete coated area
	Length of tab Cathode in x-direction
	Length of tab Cathode in y-direction
	Length of tab Cathode in z-direction
	Number of tabs Cathode
	Position of tabs Cathode
	Length Cathode in x-direction
	Length Cathode in y-direction
	Start of Layer 1 Cathode
	Length of Layer 1 Cathode in x-direction
	Length of Layer 1 Cathode in y-direction
	Start of Layer 2 Cathode
	Length of Layer 2 Cathode in x-direction
	Length of Layer 2 Cathode in y-direction
	Thickness of coating Layer Cathode
	Keywords Cathode
Anode	Anode ID
	Anode material
	Anode Mass
	Anode Area
	Anode Porosity
	Anode Carbon lot
	Anode Binder lot
	Anode Loading
	Anode Build date
	Number of anode Layers

(Continued in next column)

TABLE 4 (Continued) Metadata to include in designated metadata file by cell component.

	Anode temperature during Compression
	Anode Water/Solvent ratio
	Anode slurry mixing time
	Anode Conductive Additive amount
	Anode Binder amount
	Anode edge vs. Center
	Material current Collector anode
	Thickness current Collector anode
	Number of anode sheets
	Anode Jellyroll Mass
	Anode complete coated area
	Length of tab anode in x-direction
	Length of tab anode in y-direction
	Length of tab anode in z-direction
	Number of tabs anode
	Position of tabs anode
	Length anode in x-direction
	Length in y-direction anode
	Start of Layer 1 anode
	Length of Layer 1 anode in x-direction
	Length of Layer 1 anode in y-direction
	Start of Layer 2 anode
	Length of Layer 2 anode in x-direction
	Length of Layer 2 anode in y-direction
	Thickness of coating Layer anode
	Keywords anode
Electrolyte	Electrolyte ID
	Electrolyte Material
	Electrolyte Additive Material
	Electrolyte Additive [wt%]
	Actual Electrolyte Fill Mass
	Electrolyte Fill Date
	Electrolyte Filling Procedure
Separator	Separator ID
	Separator Base Film Material
	Separator Coating Material
	Separator Coating Particle Size
	Separator Binder Type
	Separator Solvent Type
	Separator Batch Name
	Separator Supplier Name
	Separator Grade
	Separator Base Film Thickness
	Coating Separator anodeside
	Coating Separator cathodeside
	Keywords Separator
	Length Separator in x-direction
	Length Separator in y-direction

TABLE 5 Exemplary Metadata header entries for in-field data collection.

| Metadata header	Required	Definition
Logger ID	Recommended for in-field data	An identifying string unique to the in-field data logger
Device ID	Recommended for in-field data	An identifying string unique to the device/cell/pack/system being operated across the organization
{Application} ID	Recommended for in-field data	An identifying string unique to the application that the device is in. For instance, a vehicle ID or electronic ID.
BMS Hardware Version	Recommended for in-field data	Version of hardware collecting the in-field data
BMS Firmware Version	Recommended for in-field data	Version of software collecting the in-field data
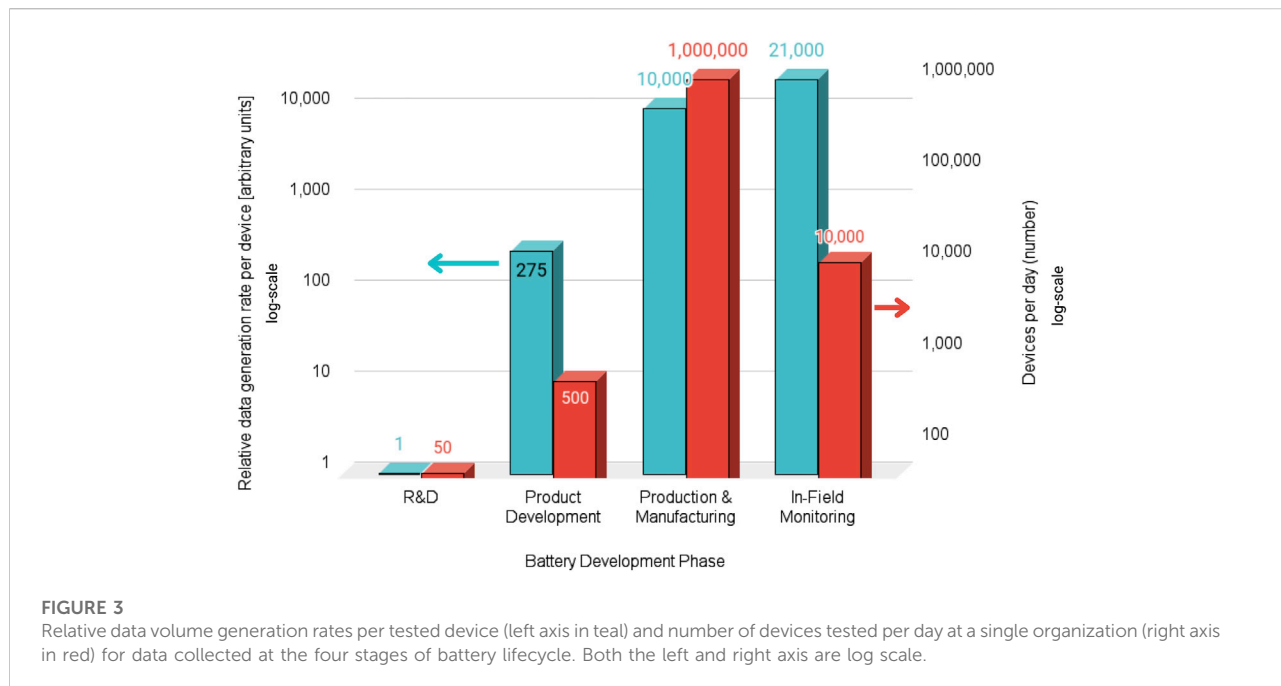
FIGURE 3
Relative data volume generation rates per tested device (left axis in teal) and number of devices tested per day at a single organization (right axis in red) for data collected at the four stages of battery lifecycle. Both the left and right axis are log scale.

considerations, and initial performance measurements such as resistance or OCV. All of this metadata can be handled in the VDF Metadata File. The linking of the aforementioned time-series data to the metadata file is done through the metadata header in the time-series data file. This linking allows for these disparate measurements (time-series and metadata) to be linked together electronically and handled at scale.

Auxiliary data collection at this phase consists of measurements to better understand safety, chemistry degradation methods, and changes of the actual cell with aging. Electrochemical impedance spectroscopy (EIS), temperatures of the cell or environment during testing, and changes in cell dimensions with cycling are examples of the types of auxiliary measurements and data collected in this phase of battery development. These auxiliary measurements are captured in the VDF "Time-Series Data" file.

The volumes of data collected within research and development are small-to-moderate compared to traditional 'big data', but can still be cumbersome to process manually

depending on the scale of the program. The data generation rate is less than a kb/min per cell. Additionally, the number of cells tested per day at a typical R&D organization is on the order of 50. Due to the exploratory nature of research and development, the cleanliness of the data is generally quite low, meaning that conventions, definitions, and standards to experimental protocols and data collection are lacking. This is an excellent opportunity for the data collected in this phase of development to be written in or mapped to the Voltaiq Data Format.

Product development

As we move to the Product development phase, the nature of the product greatly impacts the data collected. For the purposes of this publication, we focus on applications that depend on a battery pack to operate, for example an electric vehicle, power tools, grid scale storage, or consumer
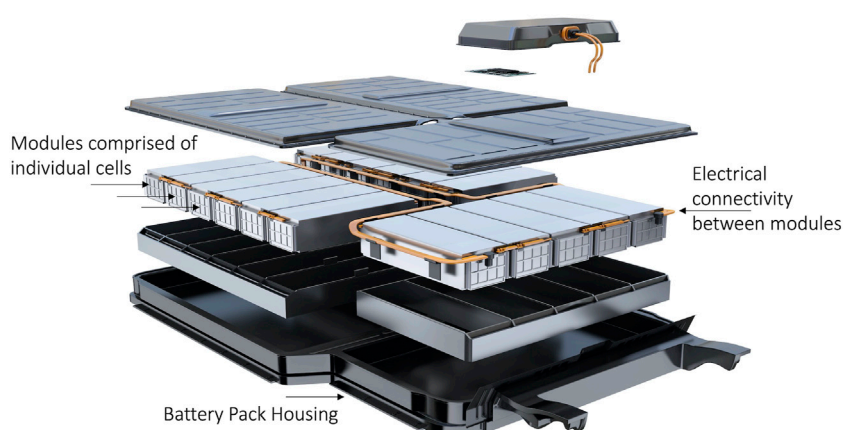
FIGURE 4
Diagram of a battery pack showing the pack and module components in a battery pack. Electronic connectivity between modules is shown in orange wires. Chesky/shutterstock.com.

electronics. In these applications, during the product development phase, the building out of the larger-format storage system begins. This task no longer involves the collection of data only from a single cell but focuses additionally on collecting data from the larger system. The components and environment are designed and engineered such as the cell or module connectivity, thermal interface materials, electrical connectors, contact system, communication system, battery management system, and housing as shown in Figure 4.

The cells, modules, environment of the system, and the system itself all have sensors and time-resolved data streams associated with them. Traditional electrochemical measurements such as time-resolved current and voltage are still recorded, however there are now multiple current and voltage measurements per system, as each cell and module within the system can be instrumented, as well as the pack itself, with electrochemical data recorded for each component. The time-series data for each of these instrumented components should be recorded according to the same conventions for the data recorded during the research and development phase. The conventions for the signing of current, power, capacity, the accumulation of capacity, should all remain the same as outlined in the VDF conventions. This allows for the insights gathered at the research and development phase to be readily and meaningfully compared to the insights in the product development phase. Additionally, since there are now multiple module- and cell-level traces, it is important that each measurement in the datafile is associated with the physical structure of the energy storage system as outlined in the Time-Series Data section. The convention is to label the hierarchy for the trace from high to low, for example "Mod01_Cell01_Voltage".

Measurements as a function of throughput are also made for large format systems, however, the concept of a traditional charge and discharge cycle may not be applicable depending on the application for the storage system. For instance, in the automotive industry, drive cycles are used for throughput measurements, while in grid storage, cycles representative of the grid operation may be an appropriate throughput measurement. That the frequency of data collection in this phase is normally commensurate with the collection frequencies at the research and development phases. The volumes of data are on the order of 100x larger than data volumes collected in the research and development phase per tested device though, as each system has many more components being measured compared to the research and development phase. For example, in a battery pack for an automotive application in this phase, there can be on the order of 4–24 battery modules per pack, each module with on the order of 10–100 cells, as shown in Figure 4. For one single recorded trace such as voltage, that is one trace per cell for 10 to 100 cells/module. For 10 modules, 100x to 1,000x more data per individual measurement trace, as illustrated in Figure 3. This results in on the order of 100 s kb/min per device data recording rate. It is noteworthy that every single cell per module in a pack may or may not be instrumented for measurement.

The Battery Management System (BMS) is also developed and finalized during this phase of development. The BMS is a software system which operates the energy storage device and environment in an intelligent manner to ensure the energy storage system is safe and operates within safe operation conditions. Additionally, the BMS is used to communicate the state of the battery to the rest of the system, and to alert the end user or the system on how to operate the storage system most effectively. The BMS both takes in and produces data—for the electrochemical data produced by the BMS, it can be treated

effectively the same as the time-series data measurements in the testing phases of product development.

The types of metadata collected in the product development phase, distinct from the research and development phase, include the electrical configuration of the pack such as number of cells per module, number of modules per pack, series and parallel configuration of cells or modules within the pack, environmental variables such as the cooling system, type of coolant, pack voltage, or specific energy. However, all this metadata is still well captured in the VDF metadata file format, where each row in the file represents an energy storage device, each column defines the metadata entry, and the subsequent column defines the units of measurement.

Production

Once a system has finished development, and the commercial organization is confident they can meet the requirements and warranty conditions for the electrified product, production can begin. Battery production is generally undertaken by a company that specializes solely in producing batteries, as opposed to the organization responsible for the electrified product or the original equipment manufacturer (OEM). The scale of production and subsequently the scale of the data are orders of magnitude larger than any of the previous stages in the battery lifecycle. Data collection rates per device per minute can be on the order of 10,000x more than on the research and development phase as shown in Figure 3 left axis. Once production ramps up, a typical large battery manufacturer can produce over 5 GWh per year, or several million cells per day, as shown on the right axis of Figure 3.

In the production phase, cells, modules, and packs are produced. We will only be commenting on the cell production phase. The data that is collected in the cell production phase is notably different in nature than that in the previous phases, reflective of the fact that the outcome goals are different. In earlier phases, the primary outcome is a deep understanding of the cell, battery, and system, across all conditions relevant to operation of the electrified product. The goal of gaining deep understanding of the system is to ensure that the cell and technology selection, pack design, system and product designs are selected and engineered to ensure a high quality, highly performant, dependable, and safe product capable of withstanding variable operation and environmental conditions and capable of meeting warranty demands. In battery production, in addition to the above goals, it is also critical to mass produce the cell designed and developed in the previous phases to the quality standards contracted between the manufacturer and their OEM customers. Metrics important to maintaining quality and process stability during mass production of the cell are therefore collected.

There are three broad processes in cell production: Electrode Manufacturing, Cell Assembly, and Cell Formation, as shown in

Figure 5 (Liu et al., 2021). Time-series electrochemical data is collected throughout the production process, typically as part of in-line or sampled quality testing. However, much of the data is collected at the 'end of line' (EOL) during or immediately after the Cell Formation process, once the cell has been assembled. At the end of the production line, standard electrochemical tests are performed to evaluate the cell and ensure the cell meets the required performance standards. The types of tests at EOL generally consist of several carefully controlled, full charge-discharge cycles, resistance or impedance testing, open circuit voltage measurement, and storage testing (Wolter et al., 2012; Weng et al., 2021). For the typical large battery manufacturer that is producing several million cells per day, the amount of time-series data that is generated and must be stored can easily require hundreds of fields and millions of rows. The time-series data is still collected in the same format and with the same conventions as all of the previous phases in the battery lifecycle. This again allows for meaningful comparisons of data across the lifecycle and can enable rapid troubleshooting.

The metadata collected during the production process differs from the metadata collected in the earlier phases of development, reflecting the differing outcome goals of this process compared to the previous lifecycle stages (Figure 1). In production, metadata associated with each manufacturing process step in Electrode Manufacturing, Assembly, and Formation (Figure 5) is collected to be able to control, maintain, and troubleshoot problems in the cell production process. Examples of metadata collected at each phase of the production process are shown in Figure 6. While this is not an exhaustive list of metadata, these are the primary metadata categories related to mass production of a cell. This metadata is often used to capture process parameters, settings, and performance indicators. Due to the large amount of equipment, instrumentation, and personnel involved, it is necessary to generate large volumes of data to power the key technologies of so-called Industry 4.0: data analytics, artificial intelligence, and robotic equipment. Typical data generation rates range from 0.1 to 1 mb/min per device, depending on the choice and quantity of instrumentation.

The conventions, data formats, and standardization practices around metadata collection in cell production are not consistent across the industry. There are some large manufacturers with their own internal standardized practices, but there are many new battery manufacturing plants in the planning or early construction and commissioning stages, while standard operating procedures are still lacking across the industry. This lack of standardization makes it slow, difficult, and in some cases, incompatible to collaborate between industry, academia, suppliers, and equipment vendors. The metadata format outlined in *Voltaiq Metadata File Section*: Metadata File provides a standard for how to capture this metadata. Each row in the file represents an energy storage device, each column defines the metadata entry, and the subsequent column defines the units of measurement. Additionally, the set of metadata
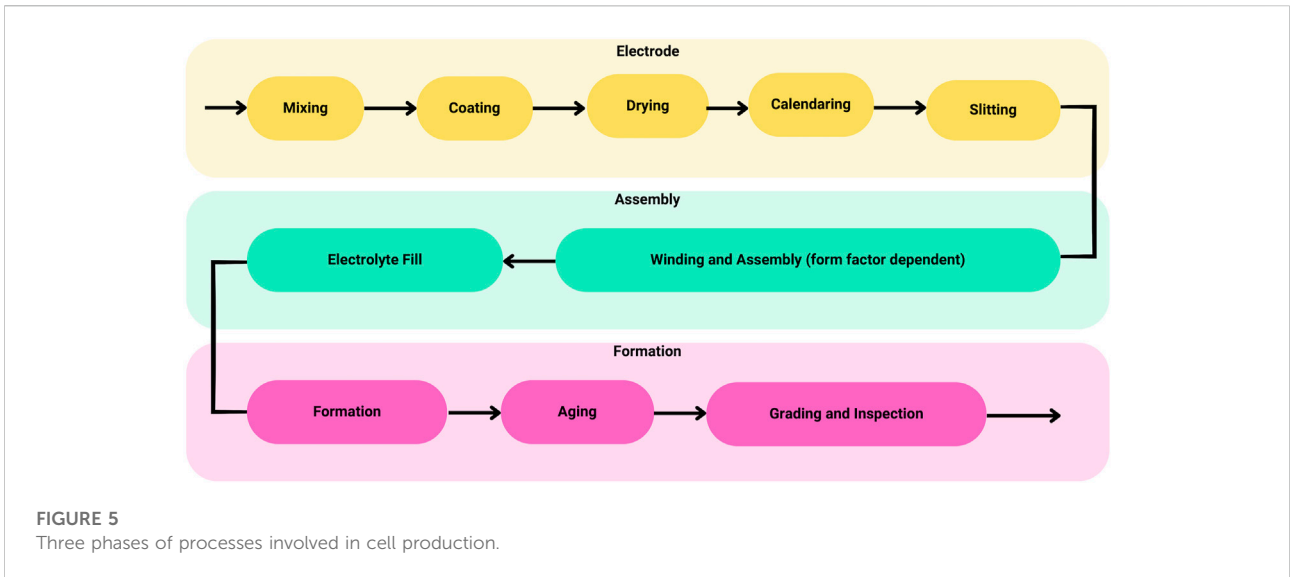
FIGURE 5
Three phases of processes involved in cell production.
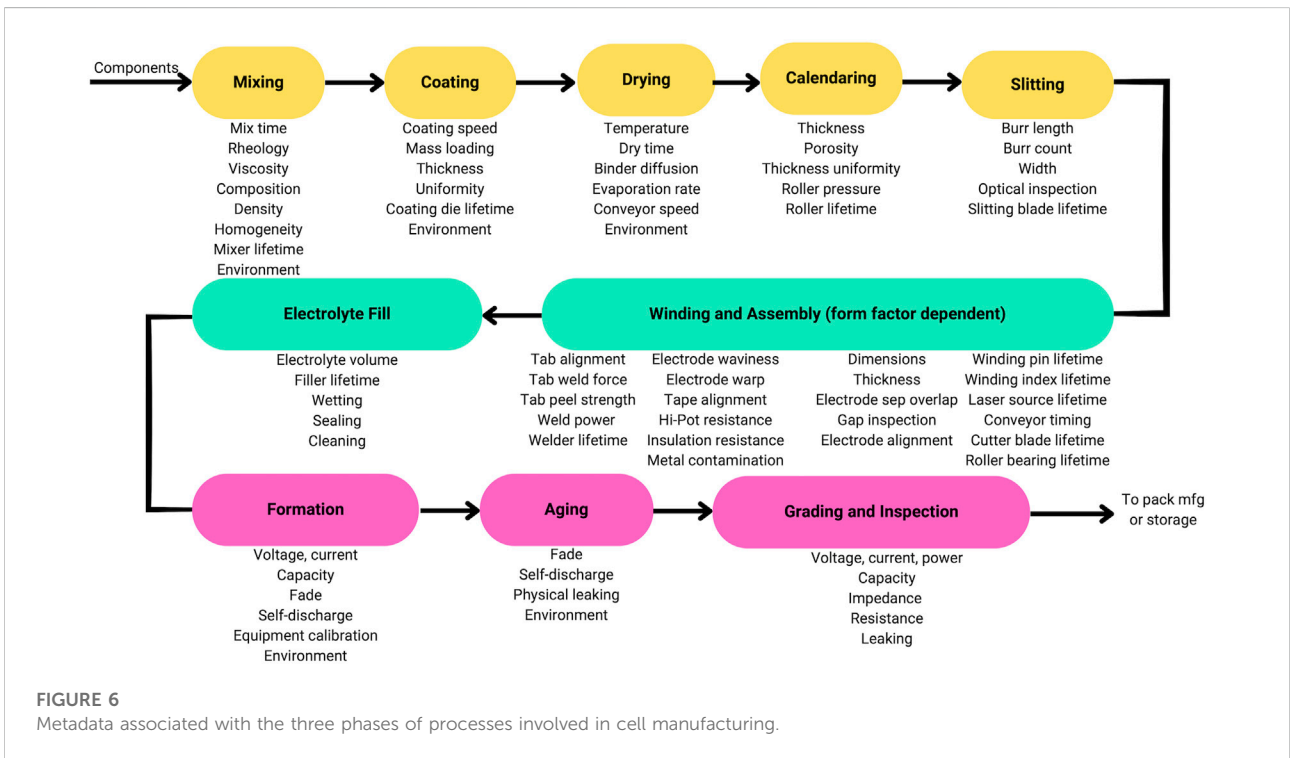
FIGURE 6
Metadata associated with the three phases of processes involved in cell manufacturing.

that is shared between battery production companies, OEMs, materials suppliers, and other commercial stakeholders in the battery lifecycle is not standardized, nor is the format in which the data is shared. The set of metadata in Tables 3, 4 serve as a set of standard metadata that can be requested and shared across commercial entities. The format for the metadata sharing is standardized and outlined in *Voltaiq Metadata File Section*.

The cells that are deemed to be good quality are then assembled into a pack, either on-site at the existing battery manufacturing plant or at a secondary location. While there is additional metadata and processes associated with pack assembly, we are not addressing this explicitly in this publication. However, the same general metadata collection process and standards apply, the identity of the metadata entries merely change.

In-field use and monitoring

Once a product has been deployed to the field and is in the end user's hands, for many applications, the monitoring and measurement of the battery continues. This is accomplished through on-board systems which record data and send it either wirelessly or over a wired network to large data storage centers for further analysis. For the measurements taken from the battery in the deployed application, sampling frequencies vary from kHz to hourly. However, the battery data is included amongst many other data traces describing other active processes in the electrified product, and the data is recorded at varying frequencies, or out-of-band. This results in exorbitantly large data volumes, 1–5 mb/min of data generated per device in the automotive industry (Wang et al., 2017). The types of data included in in-field monitoring can be as descriptive as cell-level or module-level data but can also only include pack-level readings, depending on the configuration of the monitoring system. There can be measurement data as well as calculated data by the Battery Management System, such as the state of charge (SOC), state of health (SOH), and current and voltages. In-field data capture is an emerging space, we propose here guidelines on how to modify Voltaiq Data Formation for in-field data in *In-field Data Collection section* of the results. Additionally, we open-source our guidelines and invite the community to evolve the guidelines into standards that can be adopted and accepted as best-practices across the community.

Conclusion

Standardization is powerful in solidifying conventions, methods, and practices across a field of researchers and can accelerate processes and insights from data. Within the battery field, there are variations in data collection methods, conventions, and a lack of standardization. The lack of conformity is well justified by the widely varying chemistries, measurement equipment, use cases, applications, practices, as well as the history of the commercialization of the battery. Sets of experiments at varying phases in the battery lifecycle have differing outcome goals and therefore protocols vary, resulting in differences in collection standards. However, open-access comparison of data, reproduction of scientific findings, conversations about best practices, and innovation at an unprecedented scale are all needed in present-day commercialization of the lithium-ion battery. Adherence to a convention on data format can aid in traceability of materials, supply chain, cells, and products. Information transfer across the battery lifecycle chain can be established through standardization of data formats, standard definitions, and standard sets of metadata shared across the lifecycle. Some of the most pressing questions in battery development and engineering can be further elucidated by data science and machine learning techniques, which are deeply dependent on standardization in

data format and conventions in the underlying data. In this vein, we have published and provided open access to our standard data format, the Voltaiq Data Format, which we have found to meaningfully accommodate data collected across each phase of battery development we have outlined here. This standardization is the first step towards enabling open-access comparisons of data collected across the battery lifecycle and can enable the battery field to innovate quicker and leverage the powerful techniques of data science and machine learning more readily.

Data availability statement

Freely available battery data in the Voltaiq Data Format has been open-sourced and can be found on the Voltaiq Community platform at www.voltaiqcommunity.com.The Voltaiq Data Format and examples can be explored in more detail and contributed to at https://github.com/vq-clininger/VoltaiqDataFormat.

Author contributions

CL: Content conceptualization, formal analysis. Writing—original draft and review and editing. Original figure generation. TT: Content conceptualization. Writing—review and editing. Original figure generation. TJ: Content conceptualization. Writing—review and editing. EL: Writing—review and editing. TS: Content conceptualization. Writing—review and editing.

Conflict of interest

CL, TT, TJ, EL, and TS were employed by the company Voltaiq.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2022.1059154/full#supplementary-material

References

Bartlett, Jeff S., and Preston, Ben (2022). Automakers are adding electric vehicles to their lineups. Here's what's coming. https://www.consumerreports.org/hybrids-evs/why-electric-cars-may-soon-flood-the-us-market-a9006292675/.

Baure, G., and Dubarry, M. (2019). Synthetic vs. Real driving cycles: A comparison of electric vehicle battery degradation. *Batteries* 5, 42. doi:10.3390/batteries5020042

Berdichevsky, Gene, and Yushin, Gleb (2020). *The future of energy storage towards A perfect battery with global scale.* https://www.silanano.com/uploads/Sila-_-The-Future-of-Energy-Storage-White-Paper.pdf.

Blair, N. (2021). *Global overview of energy storage performance test protocols.* Available at: https://www.nrel.gov/docs/fy21osti/77621.pdf. doi:10.2172/1696786

Brodd, Ralph J. (1999). Recent developments in batteries for portable consumer electronics applications. *Electrochem. Soc. Interface* 8, 20–23. doi:10.1149/2.f05993if

California California moves to accelerate to 100% new zero-emission vehicle sales by 2035. *Calif. Air Resour. Board*, 22–30. 2022.

Christopherson, J. P. (2015). *Battery test manual for electric vehicles.* Idaho: Idaho Falls. Available at: https://inldigitallibrary.inl.gov/sites/sti/sti/6492291.pdf.

Christopherson, J. P. (2015). Battery test manual for plug-in hybrid electric vehicles. *Contract* 1, 1720–1723.

Chu, X., Ilyas, I. F., Krishnan, S., and Wang J. Data cleaning. Proceedings of the 2016 International Conference on Management of Data 2201–2206 ACM. San Francisco California USA. June 2016.(2016). Association for Computing Machinery. doi:10.1145/2882903.2912574

Conover, D. R. (2016). *Protocol for uniformly measuring and expressing the performance of energy storage systems.* Available at: http://www.osti.gov/servlets/purl/1249270/. doi:10.2172/1249270

Electric Vehicle Toolkit (2022). *Policy and regulation.* https://greeningthegrid.org/electric-vehicle-toolkit/topics-resources/policy-and-regulation/.

European *About the European alternative fuels observatory.* https://alternative-fuels-observatory.ec.europa.eu/general-information/about-european-alternative-fuels-observatory.2015.

Garnett, P. J., and Treagust, D. F. (1992). Conceptual difficulties experienced by senior high school students of electrochemistry: Electrochemical (galvanic) and electrolytic cells. *J. Res. Sci. Teach.* 29, 1079–1099. doi:10.1002/tea.3660291006

Global EV (2019). Global EV Outlook: *Scaling-up the transition to electric mobility.* https://www.iea.org/reports/global-ev-outlook-2019.

Guidelli, R., Compton, R. G., Feliu, J. M., Gileadi, E., Lipkowski, J., Schmickler, W., et al. (2014). Definition of the transfer coefficient in electrochemistry (IUPAC Recommendations 2014). *Pure Appl. Chem.* 86, 259–262. doi:10.1515/pac-2014-5025

Ilyas, I. F., and Rekatsinas, T. (2022). Machine learning and data cleaning: Which serves the other? *J. Data Inf. Qual.* 14, 1–11. doi:10.1145/3506712

Isse, A. A., and Gennaro, A. (2010). Absolute potential of the standard hydrogen electrode and the problem of interconversion of potentials in different solvents. *J. Phys. Chem. B* 114, 7894–7899. doi:10.1021/jp100402x

Jafari, M., Gauchia, A., Zhang, K., and Gauchia, L. (2015). Simulation and analysis of the effect of real-world driving styles in an EV battery performance and aging. *IEEE Trans. Transp. Electrific.* 1, 391–401. doi:10.1109/tte.2015.2483591

Johnson, B. (1999). Environmental products that drive organizational change: General motor's electric vehicle (EV1). *Corp. Environ. Strategy* 6, 140–150. doi:10.1016/s1066-7938(00)80024-x

Kumar, V., and Khosla, C. (2018) Data cleaning-A thorough analysis and survey on unstructured data. 2018 8th International Conference on Cloud Computing,

Data Science & Engineering (Confluence). Noida, India. Jan 2018. IEEE, 305–309. doi:10.1109/CONFLUENCE.2018.8442950

Lawder, M. T., Northrop, P. W. C., and Subramanian, V. R. (2014). Model-based SEI layer growth and capacity fade analysis for EV and PHEV batteries and drive cycles. *J. Electrochem. Soc.* 161, A2099–A2108. doi:10.1149/2.1161412jes

Liu, Y., Zhang, R., Wang, J., and Wang, Y. (2021). Current and future lithium-ion battery manufacturing. *iScience* 24, 102332. doi:10.1016/j.isci.2021.102332

Matsui, T., Kitagawa, Y., Okumura, M., Shigeta, Y., and Sakaki, S. (2013). Consistent scheme for computing standard hydrogen electrode and redox potentials. *J. Comput. Chem.* 34, 21–26. doi:10.1002/jcc.23100

Moran, P. J., and Gileadi, E. (1989). Alleviating the common confusion caused by polarity in electrochemistry. *J. Chem. Educ.* 66, 912. doi:10.1021/ed066p912

Moy, K., Lee, S. B., Harris, S., and Onori, S. (2021). Design and validation of synthetic duty cycles for grid energy storage dispatch using lithium-ion batteries. *Adv. Appl. Energy* 4, 100065. doi:10.1016/j.adapen.2021.100065

Petrova-Antonova, D., and Tancheva, R. (2020) Data cleaning: A case study with OpenRefine and trifacta wrangler. International Conference on the Quality of Information and Communications Technology. Auguest 2020. Springer, 32–40. doi:10.1007/978-3-030-58793-2_3

Reddy, M. V., Mauger, A., Julien, C. M., Paolella, A., and Zaghib, K. (2020). Brief history of early lithium-battery development. *Materials* 13, 1884. doi:10.3390/ma13081884

Rosewater, D., and Ferreira, S. (2016). Development of a frequency regulation duty-cycle for standardized energy storage performance testing. *J. Energy Storage* 7, 286–294. doi:10.1016/j.est.2016.04.004

Santini, D. J. *Electric vehicle waves of history: Lessons learned about market deployment of electric vehicles* 2011.

Tourani, A., White, P., and Ivey, P. (2014). Analysis of electric and thermal behaviour of lithium-ion cells in realistic driving cycles. *J. Power Sources* 268, 301–314. doi:10.1016/j.jpowsour.2014.06.010

Tyutyunnik, V. M. (2021). Another breakthrough in power supply technology – lithium-ion batteries: 2019 nobel prize winners in chemistry John goodenough, Stanley Whittingham and Akira Yoshino. *Image J. Adv. Mat. Technol.* 6, 163–166. doi:10.17277/jamt.2021.03.pp.163-166

United States Council for Automotive Research (1996). *USABC manual.* Available at: https://uscar.org/usabc/. doi:10.2172/214312

Wang, B., Panigrahi, S., Narsude, M., and Mohanty, A. Driver identification using vehicle telematics data. in WCX™ 17: SAE World Congress Experience (2017). March 2017. doi:10.4271/2017-01-1372

Wang, X., and Wang, C. (2020). Time series data cleaning: A survey. *IEEE Access* 8, 1866–1881. doi:10.1109/access.2019.2962152

Weng, A., Mohtat, P., Attia, P. M., Sulzer, V., Lee, S., Less, G., et al. (2021). Predicting the impact of formation protocols on battery lifetime immediately after manufacturing. *Joule* 5, 2971–2992. doi:10.1016/j.joule.2021.09.015

Wolter, M., Fauser, G., Bretthauer, C., and Roscher, M. A. (2012). End-of-line testing and formation process in Li-ion battery assembly lines. Int. Multi-Conference Syst. Signals Devices. Chemnitz, Germany. 20-23 March 2012. IEEE, 1–3. doi:10.1109/SSD.2012.6198092

Yang, Z., and Rutherford, D (2019). *Japan 2030 fuel economy standard.* Washington, DC: The International Council on Clean Transportation. Available at: https://theicct.org/publication/japan-2030-fuel-economy-standards/ (Accessed September 29, 2022).