PHEMI Internal Style Guide

Contents

A to Z.....3

A to Z

access policy

Lowercase (i.e., not "Access Policies").

An access policy is a set of logical rules that determines how users can consume data stored in the PHEMI system. Access policies can be optionally applied to data sources and datasets.

authorizations

Lowercase (i.e., not "Authorizations").

Authorizations are properties assigned to system users. Authorizations are combined with *data visibilities* to determine what permission a user has to interact with different data. For example, a user with Clinician authorization might be allowed to access all forms of health data, including confidential or identifiable information, while a user with Researcher authorization might be allowed to consume only with de-identified or nonidentified information.

big data

Do not capitalize. Do not hyphenate when used as an adjectival phrase (i.e., "big data warehouse" not "big-data warehouse").

cell

A cell is the smallest unit of data storage in the PHEMI system. A cell can range from a single byte up to gigabytes. *Raw data* is stored in a single cell. Elements of *derived data* (transformed from the raw data) are also stored individually in cells. Any cell can be protected through the application of data visibilities.

code library

A code library is a package of executable code that is included in a *DPF archive*. Whether the code is source or compiled depends on the coding language. Code libraries must be portable and self-contained; that is, all dependencies required for the *DPF* to function must be bundled inside the library, in the appropriate way, for whatever language is being used.

dataset

One word (i.e., not "data set").

A dataset is a view, or map, of an underlying set of data. Data items in a dataset can be selected from across multiple data sources and DPFs. The dataset is a view, or map, to the underlying data. The actual content of the dataset (that is, the dataset's data) is generated when the dataset is executed or when it is queried against.

data category

Not "data source category" or "datasource category."

Data categories are a way to classify data into broader groupings. Examples of data categories are "Research Reports," "X-Rays," and "Prescriptions."

data source

Two words (i.e., not "datasource").

A data source represents an external source of data that will submit data to be *ingested* into the PHEMI system. A data source definition specifies data ownership, privacy, retention rules, and other properties of the source data. These definitions are used to create *metadata* for every piece of data that is ingested into the system, enabling the tracking and management of all data in accordance with the properties.

data visibilities

See visibilities.

derived data

Derived data is data that is extracted from the *raw data* ingested into the PHEMI system. Derived data is produced from raw data by means of a *DPF*. The set of derived data items can be searched, further processed, or exported from the system.

digital asset

A digital asset is any piece of data stored in the system. This may be *raw data*, or it may be *derived data* that has *metadata* defining how the data is to be managed.

Data Processing Function, DPF

Capitalized.

A DPF is an executable piece of code that supplies the instructions for parsing *raw data* (for example, a log message or medical report) into *derived data* (such as a temperature reading or blood glucose measurement). The output of a DPF is structured elements, which includes a type property (for example, INT or STRING) and can include other attributes (for example, SECRET or IDENTIFIABLE). A DPF is "registered" with a specific *data source*, by uploading the *DPF archive* as part of data source configuration.

DPF archive

Note that "archive" is lower case.

A DPF archive is the set of code comprising a *Data Processing Function*. A DPF archive as a ZIP file archive, which contains a *manifest file* and a *code library*. To associate a DPF with a *data source*, the DPF archive is ``registered`` with the data source by uploading the archive during data source configuration.

ingestion

Ingestion is the process by which data is brought into the PHEMI system. The sending system (the *data source*) submits the data to the PHEMI system, which listens for the data using a web service. The specific characteristics of data ingestion can be specified per data source as part of the data source configuration.

JSON

JSON stands for JavaScript Object Notation. JSON is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate. JSON is used in the body of several *REST* requests in the PHEMI RESTful API.

key-value pairs

A key-value pair is a set of two linked data items: a key which uniquely identifies some item of data, and the data value itself.

logical rows

When a collection of *key-value pairs* from a single document are grouped together, the result is a logical row.

M2M

M2M is a way of referring to machine-to-machine interfaces, used in machine-to-machine communication.

manifest file

A manifest file is a JSON file that specifies the output of a *DPF*. With the *code library*, the manifest file makes up the *DPF archive* that is uploaded to register the DPF with a *data source*. The manifest file should include the properties of the DPF along with the details of each *derived data item* to be generated.

metadata

One word (i.e., not "meta data")

Metadata is information about a piece of data. In the PHEMI system, metadata is information about how a given piece of data is to be managed. When a piece of *raw data* is *ingested* into the PHEMI system, raw data is ingested into the PHEMI system, information from the connection together with information configured for the *data source* is used to create a variety of metadata properties that are stored with the raw data. For example, the

system uses the timestamp from when the object was ingested together with the retention policy in the data source configuration to generate a time-to-live metadata property for the item. When a piece of raw or *derived data* is associated with metadata, it is considered a *digital asset*.

MongoDB

(From "huMONGOus"). MongoDB is an open-source document database.

PII

Personally Identifiable Information, or PII, is a legal concept used in US privacy law and information security to mean information that can be used on its own or with other information to identify, contact, or locate a single person or to identify an individual in context. When thinking about PII, it is important to distinguish legal requirements to remove attributes uniquely identify an individual from a general technical ability to identify individuals. Because of the versatility and power of modern re-identification algorithms, together with the amount of information freely available from all sources, the absence of PII data does not guarantee that de-identified data cannot be used, perhaps in combination with other data, to identify individuals.

PHEMI Central (or PHEMI Agile)

No article. If modifying a noun phrase that would normally take an article (e.g., "a big data warehouse") use an article ("the PHEMI Central big data warehouse").

privacy-level visibilities

Privacy-level visibilities are *data visibilities* that characterize the privacy level of a data item. The PHEMI system includes predefined privacy-level visibilities designed to apply to data domains where privacy is important.

- **IDENTIFIED**. The data contains Personally Identifying Information that potentially identifies an individual. Examples of information of this type include name, Social Insurance Number, and date of birth.
- **DE-IDENTIFIED**. The data contains IDENTIFIED information that has been masked or encrypted.
- **NON-IDENTIFIED**. The data is not identifying in and of itself. Examples of this type of information include weight or favorite food.

privacy, security, and governance

Always use the same order. Do not capitalize (unless in a heading or title).

Privacy by Design

Title case (i.e., capitalize "Privacy" and "Design"; not "by"). If shortened, the acronym is "PbD."

raw data

In the PHEMI system, raw data items are files, objects, records, images, and so on that are submitted for *ingestion* into the system. Raw data is stored exactly as received, along with the *metadata* generated for it on ingestion.

REST, RESTful API

"REST" is all caps. The "ful" in "RESTful" is lower case.

Representational Statement Transfer (REST) is an architectural style that uses HTTP requests and associated methods (POST, PUT, GET, and DELETE) to create, update, read, and delete data. A RESTful API is an application programming interface (API) based on REST.

visibilities

All *raw data* and *derived data* stored in the PHEMI system can be tagged with attributes that provide information about the data's sensitivity and the visibility it should have to different system users. These attributes are called data visibilities.

ZIP file

"ZIP" is all caps.