After a rigorous review of available metadata schemas, IDBRR determined that selected elements from Project Open Data’s (POD) implementation of W3C’s “Data Catalog Vocabulary” (DCAT) schema was the most appropriate schema for the IDBRR repository, due to POD’s for its authoritative and exhaustive selection of schema elements describing datasets and data “catalogs,” especially those published in an open government context, which many of IDBRR’s published data breach reports are.

There were several characteristics of data breach datasets that IDBRR hopes to include in metadata records in the future, but as of right now cannot locate a stable, trustworthy metadata schema to encode the values. These future elements will describe the following characteristics:

* The size of the data breach, maybe using a controlled vocabulary to describe categories of sizes (e.g., large, small, medium)
* The industry affected by the breach (e.g., government, private sector, agriculture)
* The type of breach that occurred (e.g., cyberattack, phishing, negligence)

All metadata elements used in IDBRR metadata records are included in the below table. All of the requisite information, including each metadata element’s name, namespace, optionality (i.e., “Mandatory,” “Suggested,” or “Optional”), repeatability (i.e., “Repeatable” or “Not Repeatable”), definition, value and/or encoding options (e.g., controlled vocabularly, string, integter), and an encoding example in the JSON format, are included in the below table.

The namespace for all elements is: <https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld>

\*Note that element names that include the @ symbol enclosed in parentheses (i.e., (@)) do so to escape Excel’s formatting stipulations that read the @ symbol as calling a function. The actual metadata element should not include the enclosing parentheses (i.e., not (@)context, but @context instead).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Description Level** | **Element** | **Optionality** | **Repeatability** | **Definition** | **Value Options** | **Example** |
| Catalog | (@)context | Optional | Not Repeatable | "The URL or JSON object for the JSON-LD Context that defines the schema used." (POD v1.1) | String (URI) | {"@context": "https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld"} |
| Catalog | (@)id | Optional | Not Repeatable | "A unique identifier for the Catalog as defined by JSON-LD Node Identifiers. This should be the URL of the data.json file itself" (POD v1.1) | String (IRI) | {"@id": "https://www.agency.gov/data.json"} |
| Catalog | (@)type | Optional | Not Repeatable | "The metadata type as defined by JSON-LD data types. This should be dcat:Catalog for the Catalog" (POD v1.1) | String (IRI) | {"@type": "dcat:Catalog"} |
| Catalog | conformsTo | Mandatory | Not Repeatable | "This is used to identify the schema version using a URI." (POD v1.1) | String (URI) | {"conformsTo": "https://project-open-data.cio.gov/v1.1/schema"} |
| Catalog | describedBy | Optional | Not Repeatable | "This is used to specify a JSON Schema file that defines all fields. By default, it is recommended that the canonical JSON Schema file is referenced (https://project-open-data.cio.gov/v1.1/schema/catalog.json) but if the schema had been extended, publishers may reference a file that defines those extensions." (POD v1.1) | String (URI) | {"describedBy": "https://project-open-data.cio.gov/v1.1/schema/catalog.json"} |
| Catalog | dataset | Mandatory | Repeatable | "This field is a container for an array of Dataset objects. See Dataset Fields below for details" (POD v1.1) | Array of Objects (JSON) | {"dataset": [...]} |
| Dataset | (@)type | Optional | Not Repeatable | "The metadata type as defined by JSON-LD data types. This should be dcat:Dataset for the Dataset" (POD v1.1) | String (IRI) | {"@type": "dcat:Dataset"} |
| Dataset | conformsTo | Optional | Not Repeatable | "This is used to identify a standardized specification the dataset conforms to. If this is a technical specification associated with a particular serialization of a distribution, this should be specified with conformsTo at the distribution level. It’s recommended that this be a URI that serves as a unique identifier for the standard. The URI may or may not also be a URL that provides documentation of the specification." (POD v1.1) | String (URI) | {"conformsTo": "http://www.agency.gov/common-vegetable-analysis-model/"} |
| Dataset | contactPoint | Mandatory | Not Repeatable | "This is a container for two fields that together make up the contact information for the dataset. contactPoint should always contain both the person’s appropriately formatted full name (fn) and email (hasEmail)." (POD v1.1) | vCard object | "contactPoint": {"@type": "vcard:Contact","@type": "vcard:Contact", "fn": "Jane Doe", "hasEmail": "mailto:jane.doe@agency.gov"} |
| contactPoint Element | contactPoint -> @type | Optional | Not Repeatable | "The metadata type as defined by JSON-LD data types. This should be vcard:Contact for contactPoint" (POD v1.1) | String (IRI) | {"@type": "vcard:Contact"} |
| contactPoint Element | contactPoint -> fn | Mandatory | Not Repeatable | "This should include included with hasEmail as part of a record’s contactPoint (see above example at "contactPoint")" (POD v1.1) | String | {"fn": "Jane Doe"} |
| contactPoint Element | contactPoint -> hasEmail | Mandatory | Not Repeatable | "This should be formatted per vCard specifications (see example) and included with fn as part of a record’s contactPoint (see above example at "contactPoint"))." (POD v1.1) | String | {"hasEmail": "mailto:jane.doe@agency.gov"} |
| Dataset | describedBy | Optional | Not Repeatable | "This is used to specify a data dictionary or schema that defines fields or column headings in the dataset. If this is a machine readable file, it’s recommended to be specified with describedBy at the distribution level along with the associated describedByType. At the dataset level it’s assumed to be a human readable HTML webpage or PDF document. Documentation that is not specifically a data dictionary belongs in “references”" (POD v1.1) | String (URL) | {"describedBy": "http://www.agency.gov/vegetables/definitions.pdf"} |
| Dataset | describedByType | Optional | Not Repeatable | "This is used to identify the media type (IANA Media Type - https://www.iana.org/assignments/media-types/media-types.xhtml - also known as MIME Type - https://en.wikipedia.org/wiki/Media\_type) of the URL used for the dataset’s describedBy field. This should be specified if describedBy is not an HTML webpage." (POD v1.1) | String (IANA Media Type: https://www.iana.org/assignments/media-types/media-types.xhtml) | {"describedByType": "application/pdf"} |
| Dataset | description | Mandatory | Not Repeatable | "This should be human-readable and understandable to an average person." (POD v1.1) | String | {"description":"This dataset contains a list of vegetables, including nutrition information and seasonality. Includes details on tomatoes, which are really fruit but considered a vegetable in this dataset."} |
| Dataset | distribution | Mandatory if the dataset has an "accessURL" or "downloadURL" | Repeatable | "This is a container for one or multiple "distribution" objects which group together the fields: "accessURL", "conformsTo", "downloadURL", "describedBy", "describedByType", "description", "format", "mediaType", and "title". | Array of Objects (JSON) | See combination of below "distribution" element examples, which would be combined into a single JSON object in a real metadata record |
| distribution Element | (@)type | Optional | Not Repeatable | "The metadata type as defined by JSON-LD data types: https://www.w3.org/TR/json-ld/#specifying-the-type. This should be "dcat:Distribution" for each distribution" (POD v1.1) | String (IRI) | {"@type": "dcat:Distribution"} |
| distribution Element | accessURL | Mandatory if the file is accessible indirectly, through means other than direct download. | Not Repeatable | "This should be the URL for an indirect means of accessing the data, such as API documentation, a ‘wizard’ or other graphical interface which is used to generate a download, feed, or a request form for the data. When accessLevel is “restricted public” but the dataset is available online indirectly, this field should be the URL that provides indirect access. This should not be a direct download URL. It is usually assumed that accessURL is an HTML webpage." (POD v1.1) | String (URL) | {"accessURL":"http://www.agency.gov/api/vegetables/"} |
| distribution Element | conformsTo | Optional | Not Repeatable | "This is used to identify a standardized specification the distribution conforms to. It’s recommended that this be a URI that serves as a unique identifier for the standard. The URI may or may not also be a URL that provides documentation of the specification." (POD v1.1) | String (URI) | {"conformsTo": "http://www.agency.gov/vegetables-data-standard/"} |
| distribution Element | downloadURL | Mandatory if the file is available for public download. Accepted Values | Not Repeatable | "This must be the **direct** download URL. Other means of accessing the dataset should be expressed using "accessURL". This should always be accompanied by "mediaType". | String (URL) | {"downloadURL":"http://www.agency.gov/vegetables/listofvegetables.csv"} |
| distribution Element | describedBy | Optional | Not Repeatable | "This is used to specify a data dictionary or schema that defines fields or column headings in the distribution. If this is a machine readable file the media type should be specified with describedByType - otherwise it’s assumed to be a human readable HTML webpage." (POD v1.1) | String (URL) | {"describedBy": "http://www.agency.gov/vegetables/schema.json"} |
| distribution Element | describedByType | Optional | Not Repeatable | "This is used to identify the media type (IANA Media Type also known as MIME Type) of the URL used for the distribution’s "describedBy" field. This is especially important if "describedBy" is a machine readable file." (POD v1.1) | String (IANA Media Type: https://www.iana.org/assignments/media-types/media-types.xhtml) | {"describedByType": "application/schema+json"} |
| distribution Element | description | Optional | Not Repeatable | "This should be a human-readable description of the distribution." (POD v1.1) | String | {"description":"Vegetable data as a zipped CSV file with attached data dictionary"} |
| distribution Element | format | Optional | Not Repeatable | "This should be a human-readable description of the file format of the dataset, that provides useful information that might not be apparent from "mediaType". Note that API should always be used to distinguish web APIs." | String | {"format":"CSV"} |
| distribution Element | mediaType | Mandatory if the file is available for public download | Not Repeatable | "This must describe the exact files available at "downloadUR"L using a media type (IANA Media Type also known as MIME Type). For common Microsoft Office files, see Office Open XML MIME types" (POD v1.1) | String (IANA Media Type: https://www.iana.org/assignments/media-types/media-types.xhtml) | {"mediaType":"text/csv"} |
| distribution Element | title | Optional | Not Repeatable | "This should be a useful title for the distribution. Acronyms should be avoided." (POD v1.1) | String | {"title":"listofvegetables.csv"} |
| Dataset | identifier | Mandatory | Not Repeatable | The identifier assigned by IDBRR to the dataset, likely a DataCite DOI | String | {"identifier":"http://dx.doi.org/10.7927/H4PZ56R2"} |
| Dataset | isPartOf | Optional | Not Repeatable | "This field allows the grouping of multiple datasets into a “collection”. This field should be employed by the individual datasets that together make up a collection. The value for this field should match the "identifier" of the parent dataset." (POD v1.1) | String | {"isPartOf":"http://dx.doi.org/10.7927/H4PZ56R2"} |
| Dataset | issued | Mandatory | Not Repeatable | The date when the dataset was published by IDBRR. "Dates should be ISO 8601 - https://en.wikipedia.org/wiki/ISO\_8601 - of least resolution. In other words, as much of YYYY-MM-DDThh:mm:ss.sTZD as is relevant to this dataset." (POD v1.1) | ISO 8601 Date | {"issued":"2001-01-15"} |
| Dataset | keyword | Mandatory | Repeatable | Terms used to "tag" the dataset with relevant concepts that users will search for. "Surround each keyword with quotes. Separate keywords with commas. Avoid duplicate keywords in the same record." (POD v1.1) | Array of strings | {"keyword":["vegetables","veggies","greens","leafy","spinach","kale","nutrition"]} |
| Dataset | language | Mandatory | Repeatable | "This should adhere to the RFC 5646 standard: https://datatracker.ietf.org/doc/html/rfc5646. This language subtag lookup - https://rishida.net/tools/conversion - provides a good tool for checking and verifying language codes. A language tag is comprised of either one or two parts, the language subtag (such as en for English, sp for Spanish, wo for Wolof) and the regional subtag (such as US for United States, GB for Great Britain, MX for Mexico), separated by a hyphen. Regional subtags should only be provided when needed to distinguish a language tag from another one (such as American vs. British English)." (POD v1.1) | Array of strings | {"language":["es-MX","wo","nv","en-US"]} |
| Dataset | license | Mandatory | Not Repeatable | The license under which the dataset is published. See IDBRR's "Data Licensing" section for allowable licenses. | String (URL) | {"license":"http://creativecommons.org/publicdomain/zero/1.0/"} |
| Dataset | modified | Mandatory | Not Repeatable | "Dates should be ISO 8601 - https://en.wikipedia.org/wiki/ISO\_8601 - of highest resolution. In other words, as much of YYYY-MM-DDThh:mm:ss.sTZD as is relevant to this dataset. If there is a need to reflect that the dataset is continually updated, ISO 8601 formatting can account for this with repeating intervals: https://en.wikipedia.org/wiki/ISO\_8601#Time\_intervals. For instance, R/P1D for daily, R/P2W for every two weeks, and R/PT5M for every five minutes." (POD v1.1) | ISO 8601 Date | {"modified":"2012-01-15"} or {"modified":"R/P1D"} |
| Dataset | publisher | Mandatory | Not Repeatable | This should be "The Interstate Data Breach Report Repository" unless otherwise discussed with an IDBRR curator | String | {"publisher": "IDBRR"} |
| Dataset | spatial | Mandatory if the dataset is spatial | Not Repeatable | "This field should contain one of the following types of content: (1) a bounding coordinate box for the dataset represented in latitude / longitude pairs where the coordinates are specified in decimal degrees and in the order of: minimum longitude, minimum latitude, maximum longitude, maximum latitude; (2) a latitude / longitude pair (in decimal degrees) representing a point where the dataset is relevant; (3) a geographic feature expressed in Geography Markup Language using the Simple Features Profile - https://www.ogc.org/standard/gml/; or (4) a geographic feature from the GeoNames database: https://www.geonames.org/." (POD v1.1) | See definition | {"spatial":"Lincoln, Nebraska"} |
| Dataset | temporal | Mandatory if applicable | Not Repeatable | "This field should contain an interval of time defined by the start and end dates for which the dataset is applicable. Dates should be formatted as pairs of {start datetime/end datetime} in the ISO 8601 - https://en.wikipedia.org/wiki/ISO\_8601 - format. ISO 8601 specifies that datetimes can be formatted in a number of ways, including a simple four-digit year (eg. 2013) to a much more specific YYYY-MM-DDTHH:MM:SSZ, where the T specifies a seperator between the date and time and time is expressed in 24 hour notation in the UTC (Zulu) time zone. (e.g., 2011-02-14T12:00:00Z/2013-07-04T19:34:00Z). Use a solidus (“/”) to separate start and end times. If there is a need to define the start or end of applicability using a duration rather than a date, ISO 8601 formatting can account for this with duration based intervals: https://en.wikipedia.org/wiki/ISO\_8601#Time\_intervals. For instance, applicability starting in January 2010 and continuing for one month could be represented as 2010-01/P1M or 2010-01/2010-02. However, when possible, full dates are preferred for both start and end times." (POD v1.1) | ISO 8601 Date | {"temporal":"2000-01-15T00:45:00Z/2010-01-15T00:06:00Z"} or {"temporal":"2000-01-15T00:45:00Z/P1W"} |
| Dataset | title | Mandatory | Not Repeatable | "Acronyms should be avoided" (POD v1.1) | String | {"title":"Types of Vegetables"} |