Dataset Aggregator

# Prerequisites

## Background information

There are a series of documents that describe the scope, application and workflow of the application:

* **NIPN Integrated Data Repository**: This document can be found in the *annexes* folder at the top level of the documentation folder. The project started with the idea of creating a generic dataset repository that could be applied to any type of dataset, this document describes the initial idea.
* **SMART Survey Aggregator**: This document can also be found in the annexes folder. This document is a further version of the project in which the application focuses in aggregating [SMART surveys](https://smartmethodology.org/).

Both documents provide an idea of the use and function of the application. Although the second document focuses on a specific type of surveys, the principles, structures and workflow can be applied to any kind of dataset, making both documents a good base for understanding what the final application should do and how it should work.

## Database framework

[ArangoDB](https://www.arangodb.com/) is the database system upon which the application was designed. It is a multi-model database that implements a micro-services framework, [Foxx](https://docs.arangodb.com/3.3/Manual/Foxx/AtAGlance.html), which is used to implement the back-end services. You should become familiar with this database system, since the application relies upon it.

## Structure

The application is implemented with two main components: the *back-end* and the *front-end* as a web-based application.

The back end takes care of storing and serving all the data, it is implemented using the ArangoDB Foxx micro-services framework which is executed in the database itself. The goal of these services is to provide a high-level interface to data for the front-end.

The front-end has not been developed yet, the plan was to create it using [React](https://reactjs.org/), since this framework is well adapted for creating independent components that can be assembled into a working application. The other reason is that both the Foxx framework and React use JavaScript as their main language, which makes for a consistent development environment. The front-end could be developed also using other frameworks and languages, such as PHP, for instance, it should be the decision of the front-end developers to decide what makes the best choice.

# Short summary

This is a very short summary of what the application should do and how it is structured, you can consider it as the summary of the two background documents cited in the beginning and serves the purpose of laying down the principles that apply to the rest of this document.

The application should implement a repository or searchable archive of studies, their documents and the data that they represent.



The goal of this system is to provide access to the studies, their annex documents and to the raw data by allowing users to query the studies metadata and the raw data variables, retrieving two sets of results: the studies and a selection of raw data that satisfies the query conditions.

The study represents a project, survey, or other types of data collection activities that can be considered as an independent entity. The study is tagged by metadata that records when it occurred, which regions it studied, which institutions were involved and all the other information that is necessary to document the conditions, methods and scope of the project.

All annex documents - in the case of surveys these could be the questionnaires, reports and original datasets – should be uploaded and linked to the study, so that it can be considered as an archive of all its information.

Studies will have one or more datasets that represent their data. The dataset exists in two forms: as the original file – Excel, Stata, SPSS, etc. – which will be uploaded as an annex file of the study, and as the data it contains that we will call raw data here. Raw data will be stored in a collection that corresponds to the data domain of the dataset. If we use SMART surveys as the example, all the SMART survey raw data will be stored in a single data collection which represents the domain of SMART surveys. The goal of this structuring is to allow searching all SMART survey data and retrieving aggregated data selections that can be downloaded by users to perform further data analysis or summaries. Data collections can be organised in ontologies to allow more than one data domain, for instance you could have a census collection that holds the data from all census datasets, this structuring allows the system to handle different types of data[[1]](#footnote-1).

The central component of the system is the data dictionary, its role is to provide the definition and documentation for all the fields of all records stored in the database. When importing raw data from a dataset, the first function to perform is to harmonise the variables of the raw data with the data dictionary, so that each column of data in the raw data table corresponds to a descriptor of the data dictionary[[2]](#footnote-2).

Once studies have been registered, their annex documents uploaded, and their data harmonised with the data dictionary, it becomes possible to perform a set of queries whose results will be a selection of studies and a selection of raw data. These two sets of results could be presented in two separate panes. The studies could be chosen and downloaded as a zipped archive in which the study metadata record could be provided as an Excel document along with all the other annex documents. The raw data could also be downloaded as a CSV file or a file of some other format. As the application matures, statistical functions could be applied to the raw data selections in the back-end to provide further functionality and features to the application users.

# Database structure

The database contains a series of predefined collections (tables in the traditional relational nomenclature), each serving a specific purpose:

* descriptors: This collection contains the variables definitions, each record represents a variable, its type, label, description and all the other information that is necessary to document and validate data associated with that variable.
* terms: A term is an item that has a code and a series of descriptive properties in several languages whose code is used as a reference. Terms are organised as ontologies (tree or graph structures) and implement all the relational structures of the data dictionary, such as controlled vocabularies (or enumerations), forms, ontologies, etc.
* schemas: This collection contains all the links between the data dictionary elements, parent-child relationships in controlled vocabularies and forms, as well as relationships between terms and their instances are recorded in this collection.
* users: This collection contains all the user records.
* hierarchy: This collection contains all the edge documents that relate users with other components of the data dictionary, for instance, such as the relationship between a user and its manager.
* studies: This collection contains the studies metadata records.
* annexes: This collection contains the study annex document records[[3]](#footnote-3).
* toponyms: This collection contains a set of toponyms which are related to terms defining geographic units. Toponyms are not needed by the application but represent a multilingual repository of geographic locations and administrative units, along with their nomenclature.
* shapes: This collection contains a set of GeoJSON shapes related to toponyms.
* edges[[4]](#footnote-4): This collection contains the relationships between toponyms, shapes and terms.
* errors: This collection contains error type definitions. The elements of this collection can be compared to the JavaScript Error object name.
* messages: This collection contains errors and other messages, the errors can be compared to the JavaScript Error message property, while the other entries are used to store the descriptions of the services in several languages.
* settings: This collection contains the settings for the application, it currently contains one entry that indicates the status of the application.
* sessions: This collection contains session data, it can be used by the front-end to store session user specific states.
* logs: This collection contains the log of the services, it records what services were called by which user and the user status at the call time.
* groups: This collection was supposed to contain user groups. User groups are not implemented, although they are earmarked in the code.
* smart: This collection should contain the SMART survey raw data as SMART datasets are harmonised and added to the database.

# Data Dictionary

The main component of the application is the data dictionary, all fields of all collections *must* have an entry in the data dictionary, this means that all data elements are defined, documented and available.

## Common default fields

The key fields of all collections follow the ArangoDB standards:

* **\_id**: This field contains the unique identifier of the record within the *database*. It is the concatenation of the collection name, followed by a ‘/’ token and the value of the \_key field. This field will generally never be set programmatically since the database will automatically set it, but it is required when referring to a record belonging to an unknown collection. Once set, the value cannot be changed.
* **\_key**: This field represents the unique identifier of the record within its *collection*. This will be the value used to identify a specific record in a collection. In some cases this field will have to be set explicitly, in other cases it is computed from other fields in the record. Once set, the value cannot be changed.
* **\_rev:** This field represents the record revision. It is never explicitly set, but it is used when updating or replacing a record: if the record selector contains a revision and that revision is different than the current record revision, the operation will fail. This is useful to ensure nobody modifies a document that is currently under modification.

All collections, document and edge, feature these fields and they have the same function in both types of collection.

Edge collections have two additional fields:

* **\_from**: This field represents the relationship source node \_id. Once set, the value cannot be changed.
* **\_to:** This field represents the relationship destination node \_id. Once set, the value cannot be changed.

## Global and local identifiers

Some collections, in particular terms and descriptors, use an additional identification scheme, this scheme uses the following fields:

nid: This field represents the namespace of the identifier. It is the \_id of another record. The referenced record must either implement this identification scheme, or have the global identifier field (gid). This field is generally not required, since there would not be any way to define a namespace.

lid: This field represents the local identifier. The value must be unique within its namespace. This field is required and connot be empty[[5]](#footnote-5).

gid: This field represents the global identifier. The global identifier should be unique among all namespaces. It is a computed field that is set by concatenating the gid of the record referenced by the nid (namespace), followed by a colon (:) and ending with the current record’s nid (local identifier). If the record does not have the namespace, the global identifier will be its local identifier.

For instance, suppose we want to create a namespace in the terms collection:

{

\_id: “terms/ISO”,

\_key: “ISO”,

lid: “ISO”,

gid: “ISO”

}

This namespace could be used to create another terms record such as:

{

\_id: “terms/ISO:ITA”,

\_key: “ISO:ITA”,

nid: “terms/ISO”,

lid: “ITA”,

gid: ISO:ITA”

}

This scheme is used in particular for controlled vocabularies where the code may be ambiguous depending on its domain. For instance, we have the it code for *Italy*, the country and we also have the it code for *Italian*, the language: to be able to uniquely identify both codes we make use of the namespace.

This means that all controlled vocabulary elements in the database will have the full code, namespace and local identifier; this also means that given a local identifier we can easily select all entries that might be a match, something useful when importing data from different sources.

## References and Relations

Fields values may be references to another record of another collection. It is the equivalent of a relation in a relational structure. The value of the field is a reference to the record that describes that field. In this application we use this especially for controlled vocabularies: for instance, you may have a “country” field that contains the key of the record that defines that country. This kind of relationship, that we call here *reference*, is used whenever you are referencing a single object and the relation links two vertices.

References are implemented in three ways:

* \_id reference. The value of the field is the \_id of the referenced record. In this case, since the \_id contains the collection to which the referenced record belongs, it is the only bit of information needed.
* \_key reference. The value of the field is the \_key of the referenced record. Since the collection is unknown, this kind of reference is indicated by a specific data type. For instance the :type:value:term data type indicates that the value of the field must be the \_key of a record in the terms collection. This is the data type of all controlled vocabulary variables and currently the only data type of this kind.
* gid reference. The value of the field is the gid of the referenced record. As for the previous type of reference, it is the data type that indicates the referenced collection. This data type is used to reference descriptors, the :type:value:field is currently the only data type of this kind[[6]](#footnote-6).

When we want to consider a relationship that connects more than two vertices, such as the paths that connect a set of cities, we use a different way of relating elements: we use *edge collections*. Each record in this type of collection contains the source and destination vertex references and other information that quantifies and qualifies the relationship. Since ArangoDB implements natively directed graphs, this mechanism allows us to consider a hierarchical structure as a single entity. Controlled vocabularies, forms, data types and many other concepts are represented in this way.

A relationship is uniquely identified by its *significant fields*, these will include the \_from and \_to fields, plus any other field that uniquely identifies the relationship. In this application we apply a specific identification scheme to prevent having to create a unique index for these significant fields. We concatenate the significant fields, separated by a TAB character, and hash the result, this value is set in the edge \_key field, which is by definition unique.

The application implements three types of edge:

* ***Edge***: This type of edge has as significant fields the \_from, \_to and predicate fields. The predicate represents the type of relationship or its verb, it is a controlled vocabulary. All edges in the application *must* have at least these three fields as significant. The \_key is the MD5 hash of these fields.
* ***Branched edge***. Suppose you have a hierarchical controlled vocabulary of geographic locations. You would have as root the term that defines the controlled vocabulary. As the first level children of that term you would have the continents, which would have the regions as children, which would have countries as their children, which would have sub-national administrative units as children and so on. If you want to implement a subset of that tree with the first type of edge, you would be forced to duplicate each of these edges to include only those elements in which you are interested in. This could become tedious, especially if the edge contains a lot of other data that characterises the relationship, so this type of edge can be used to re-use existing edges while keeping track of which branches they belong to. This edge adds the branches field which is an array containing the list of branches that use this relationship. This field is not significant, meaning that it is not used to compute the edge key, but it is used when traversing the graph to only follow paths belonging to a specific branch: you start with the origin vertex and follow all edges that have the predicate you are interested in and that have in their branches field the branch of interest. Controlled vocabularies use this type of edge, the branch represents which specific controlled vocabulary tree you want to traverse in the graph.
* ***Attribute edge***. This type of edge adds the attributes field to the significant fields. This field is an array of controlled vocabulary elements which is sorted just before computing the edge key, so these attributes become integral part of edge identification.

Significant fields cannot be modified once they are created. For instance, the branches field in branched edges is dynamic and elements are added or deleted in time, the attributes field in attribute edges is instead immutable.

## Data types



Data types are defined as a controlled vocabulary of term elements. This controlled vocabulary defines the format, scope and domain of values belonging to a specific descriptor. The data type identifier is the \_key of the term, which is copied from the term global identifier, see above.

By standard, no values can be null, a variable may either have a non-empty value, or not be there.

Data types are organised as a hierarchy and implemented as branched edges where the branch is the specific data type and the predicate is terms/:predicate:type-of. The first level elements, those in orange in the above diagram, represent the data domains. Each domain has a specific set of modifiers, that are essentially constraints applied to the domain. For instance, the text domain includes a modifier which constrains the minimum and maximum number of characters that the text may have; numbers have a similar modifier that restricts the value to a range.

Some of these modifiers are implemented in the data type, other are implemented in the descriptor. For instance, the :type:value:term data type represents a reference to the \_key field of a term, the data type sets terms as the default collection; descriptors of that type may include a list of controlled vocabularies to which values of that type must belong.

To create a specialised data type you derive from a type and modify the inherited modifiers. For instance, the string data type sets a constraint limiting the maximum number of characters to 254. Or you implement a descriptor and override the modifiers in the descriptor.

Besides reading the description below, I advise you to look at the definition of the data type terms in the terms table, the description fields contain further information to get a better idea of their function.

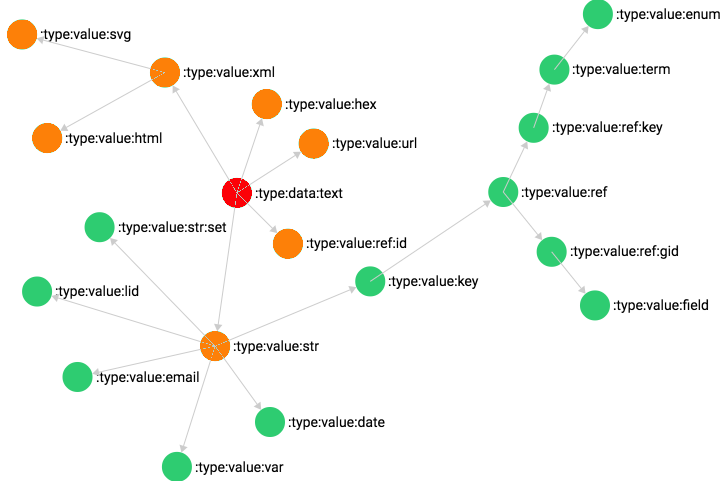
### Base data types



* **[:type:data:any**]: This data type represents a value that can take *any* type. Values of this type will not be validated, since they can take any value. This data type has no derived types.
* **[:type:data:bool**]: This data type represents a true/false *boolean* value, it can be represented in a form as a radio button or checkbox. This data type has no derived types.
* **[:type:data:text**]: This data type represents *text* data, it is by definition utf8 characters and can take any size. Because of the lack of size limits this variable can be represented with a scrollable text area in a form. This base type can have the following modifiers: minimum and maximum number of characters, regular expression and range.
* **[:type:data:numeric**]: This data type represents a number that can either be an integer or a float[[7]](#footnote-7). This base type can have the following modifiers: range and the number of decimal positions[[8]](#footnote-8).
* **[:type:data:list**]: This data type represents a list or array of items of undefined type. This base type has one modifier which determines the minimum and maximum number of elements.
* **[:type:data:struct**]: This data type represents an object or structure whose elements must follow the data dictionary rules. This means that when the validation script encounters such a value, it will recursively validate its elements. This base type does not have modifiers.
* **[:type:data:object**]: This data type represents an object or structure whose elements are parsed as a key and value pair. This base type has two modifiers, type-key:type and type-value:type, that can be used to indicate respectively the data type of the pair key[[9]](#footnote-9) and value.

### Text data types

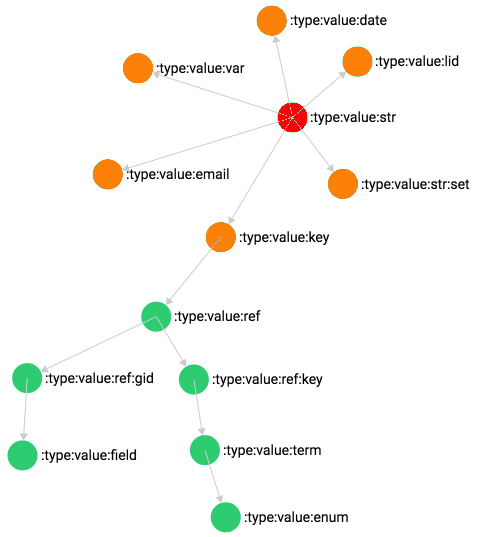
Text data types represent generic character data, ArangoDB treats all text as a sequences of UTF8 characters.



* [**:type:value:hex**]: This data type represents a hexadecimal value.
* [**:type:value:url**]: This data type represents an URL or internet link.
* [**:type:value:xml**]: This data type represents an XML text.
* [**:type:value:svg**]: This data type derives from the XML type and represents an SVG image.
* [**:type:value:html**]: This data type derives from the XML type and represents an HTML text.
* [**:type:value:str**]: This data type derives from the text type and represents a string that is limited to 254 characters. This type is used for text variables that can be indexed and used as object identifiers.
* [**:type:value:ref:id**]: This data type derives from the text type and represents the \_id field value of another record. This data type is used to directly refer to another object.

### String data types

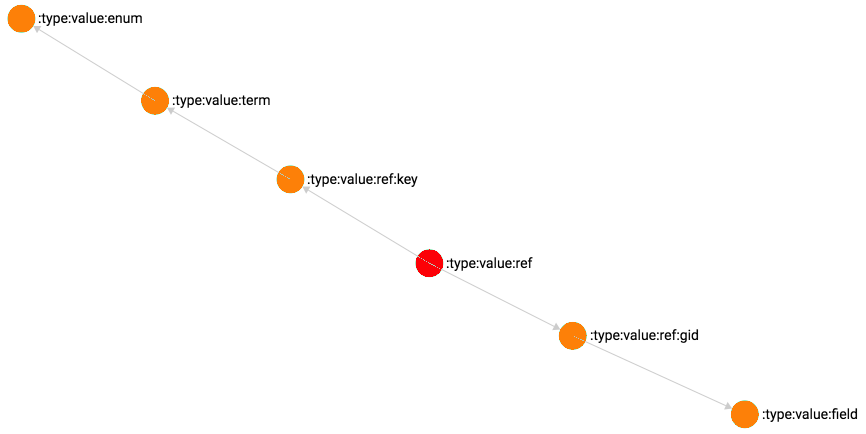
This set of data types derive from the string type, these will be generally used to represent short strings of data such as titles, labels and names.



* **[:type:value:key**]: This data type is used for strings that are used as record keys, the string cannot contain spaces, it allows a subset of punctuation and it must be non-empty and up to a length of 254 characters.
* **[:type:value:var**]: This data type is used for strings representing variable names, the string must be alphanumeric, non-empty and up to 64 characters long; it allows the use of underscores.
* **[:type:value:lid**]: This data type is used for strings representing local identifiers (lid), the string cannot contain spaces, it allows a subset of punctuation and it must be non-empty and up to a length of 64 characters.
* **[:type:value:email**]: This data type is used to represent e-mail addresses.
* **[:type:value:date**]: This data type is used to represent a string date with optional month and day in the YYYYMMDD format.
* **[:type:value:str:set**]: This data type is used to represent a list of unique strings.

### Reference data types

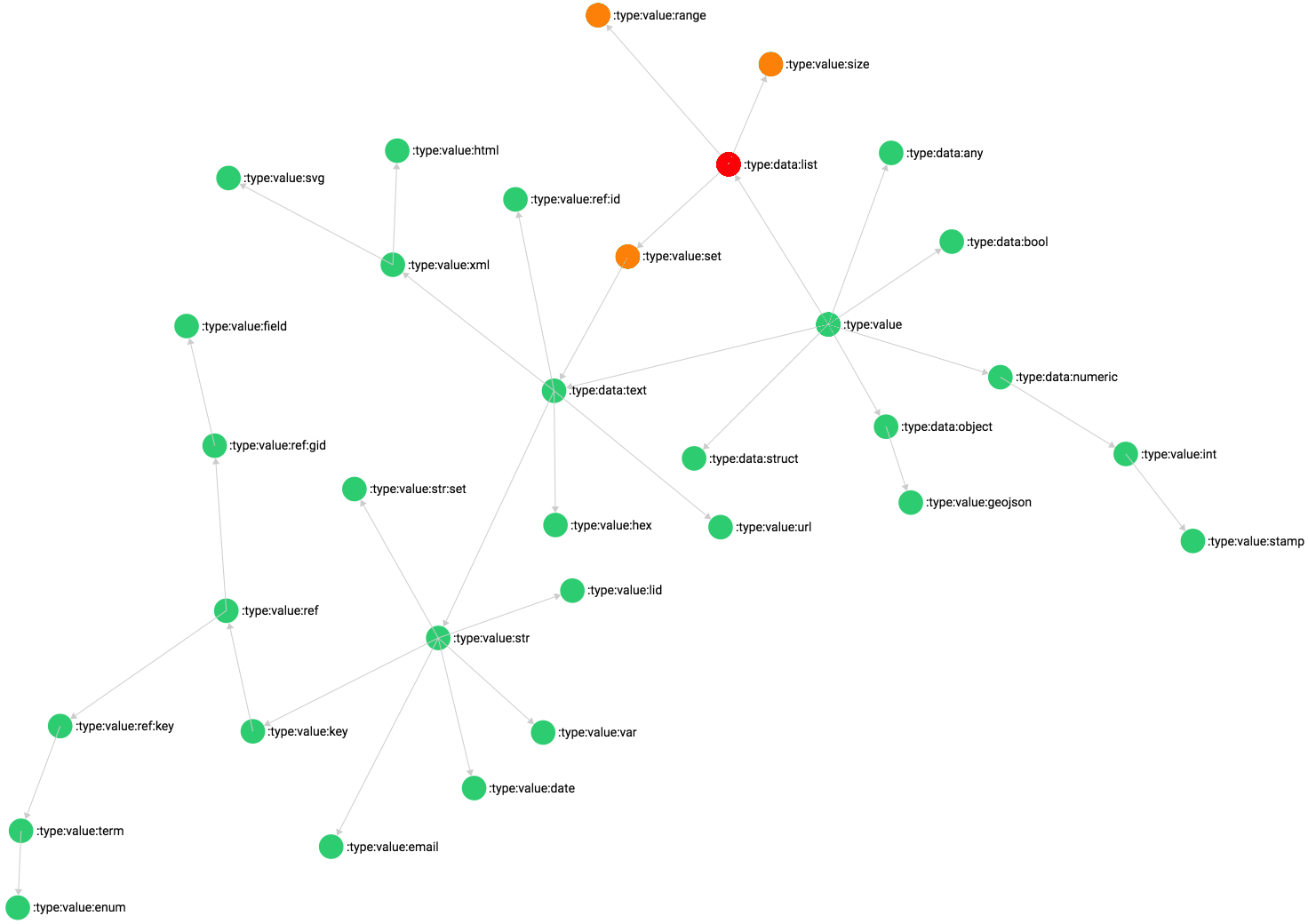
Reference data types are used for fields that reference another object through its \_key or gid field[[10]](#footnote-10).



* **[:type:value:ref:gid**]: This data type represents a reference to the gid field of another object.
* **[:type:value:field**]: This data type derives from the :type:value:ref:gid type and sets descriptors as its default collection. This data type is used to reference data fields or descriptors.
* **[:type:value:ref:key**]: This data type represents a reference to the \_key field of another object.
* **[:type:value:term**]: This data type derives from the :type:value:ref:key type and sets terms as its default collection. This data type is used to reference terms.
* **[:type:value:enum**]: This data type derives from the :type:value:term type, as its parent, it expects the value to be a reference to a term, except that the referenced term must define a controlled vocabulary[[11]](#footnote-11). This data type is used to reference enumerations that apply to a controlled vocabulary field.

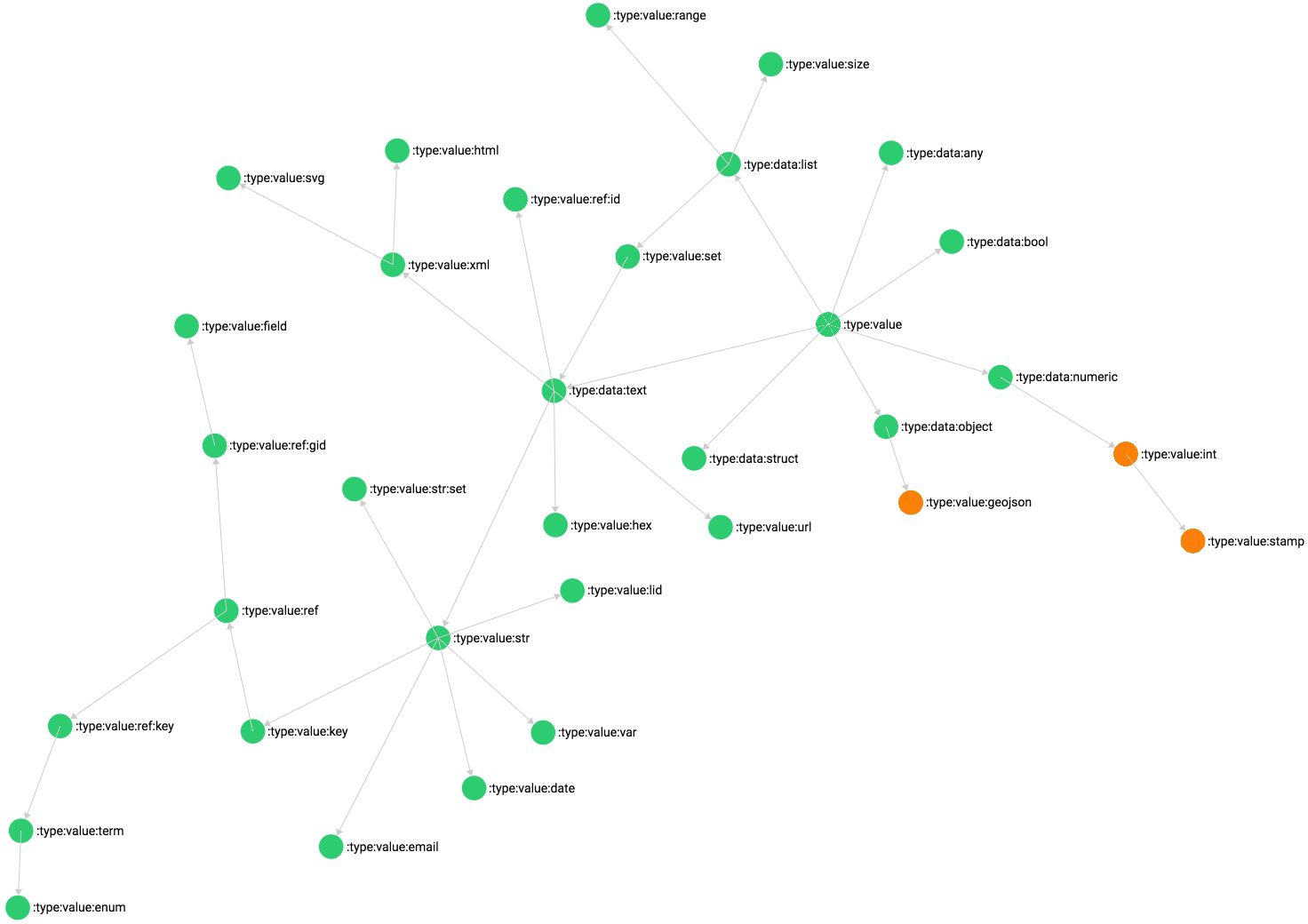
### List data types

List data types represent list of elements of a specific type. Although descriptors have a property that can be used to indicate that the variable contains a list of elements of the descriptor data type, list data types are inherently lists.



* **[:type:value:set**]: This data type represents a list of unique elements.
* **[:type:value:range**]: This data type represents a range, it is implemented as an array of 4 elements which are the minimum, maximum, include minimum and include maximum. The first two elements are numbers, the last two elements are Booleans which indicate whether or not the limits are included in the range. The first Boolean refers to the minimum bound, the second to the maximum bound.
* **[:type:value:size**]: This data type represents a size range, it is structured as the previous data type, except that it indicates a size, rather than a value range. It is used to restring string length and list elements count.

### Other data types



* **[:type:value:int**]: This data type represents integer values.
* **[:type:value:stamp**]: This data type derives from the inteher data type and is used to represent time stamps.
* **[:type:value:geojson**]: This data type represents a GeoJSON structure, it can be used to store and validate geometric shapes.

## Terms

Terms are used to implement elements of controlled vocabularies, forms, types, predicates and all other items that represent root elements of ontology structures. All terms reside in the terms collection.

Terms implement the local/global identifier scheme and the value of the global identifier (gid) is copied by default to the object’s \_key. Once a term is inserted, the nid, lid and gid field values become locked.

* **\_key**: This field contains the same value of the gid field, it represents the global identifier of the term.
* **nid**: This field contains the \_id reference to the term that represents the term’s *namespace*. All terms should have a namespace, except terms that define namespaces.
* **lid:** This field contains the term local identifier unique within its namespace (nid). This field is required.
* **gid:** This field contains the term global identifier, it is a required computed field described in the “Global and local identifiers” section at the beginning of this document.

Terms, in their minimalist form, will contain the identifier fields and a set of text description fields that represent the name, description and other information. These fields are:

* **label**: This is the name, title or label of the term, whenever a term is referenced this field can be used as a label for the term. *This field is required*.
* **definition**: This field represents the definition of the term, while the label should not be long, this can be a long text and should provide the necessary information elements. This field could be used for an info button.
* **description**: This field can be used to expand the definition with a more extensive description.
* **note**: This field can be used to add notes and comments.
* **example**: This field can be used to provide examples, if that is relevant.

These description fields are multilingual and they are implemented by using the :type:data:object base data type and setting the specific key and value modifier data types in the descriptor. The property names of the object are language codes and the values are text, for label and definition, or html for the other fields. This is the default structure for all multilingual content.

Terms also contain a set of properties that represent alternative identification or keywords:

* **var**: This field represents a variable name, it can be used as such when referring to a term. In practice, this field is set in all default terms, there is a service in the application that generates a JavaScript file in which the \_key and var fields can be consulted as a dictionary: in the source code I always refer to the var value when referencing terms, this allows me to change the global identifier without needing to revise the source code.
* **sym**: This field represents the term symbol, it is a string that represents the acronym of the term, such as cm. for centimeters.
* **syn**: This field is an array of synonyms for the term, by default it contains the local identifier and the variable name, additional used synonyms can be added to aid in selecting the right term.
* **keyword**: This is a set of keywords that can be applied to the term.
* **type-cast**: This field is a controlled vocabulary that is used only in data types for validation purposes. The value indicates what type casting is to be applied to values of the specific data type before performing the validation. For instance :rule:castNumber indicates that the value must be cast to a number.
* **type-custom**: This field is a controlled vocabulary that is used only in data types for validation purposes. The value indicates what custom validation procedure is to be applied to values of the specific data type in addition to the default procedures[[12]](#footnote-12). For instance :rule:customHex indicates that the value should correspond to a hexadecimal number.
* **collection**: This field is a controlled vocabulary that references terms defining collections. It is used in instance controlled vocabularies to indicate in which collection the instance should be stored.
* **instance**: This field is a controlled vocabulary that references the instances controlled vocabulary. I actually forgot what its use is and there is only one term that has this field, so let’s ignore it for the time being...
* **traversal**: This field is a controlled vocabulary that represents graph traversal directions, it is used exclusively in predicate terms to indicate the direction of the predicate. For instance the :predicate:category-of predicate has the :type:traversal:is value: this means that the relationship source *is* a category of the relationship destination. This value is useful to determine in what direction a graph should be traversed according to the predicate.
* **length**: This field is a value of :type:value:size type that is used exclusively in data type terms as the modifier indicating a character count range restriction.
* **regex**: This field is a string used exclusively in data type terms to provide the regular expression for validating text data.
* **deploy**: This field is required both in terms and descriptors, it is a controlled vocabulary that indicates what deployment status of the current object:
  + **[:state:application:embedded**]: It is an object implemented by the database; only descriptors can fall under this category.
  + **[:state:application:default**]: It is default object.
  + **[:state:application:standard**]: It is an object that implements a standard.
  + **[:state:application:user**]: It is a user-defined object.
* **level**: This field is a number that is used as a priority or hierarchical indicator, lower numbers represent higher priorities or top hierarchical levels.
* **class**: This field is a controlled vocabulary that indicates the class of the term, it is used in the instances controlled vocabulary to indicate the class of the instance.
* **store**: This field is a controlled vocabulary that indicates the collection type, it is used in the collection controlled vocabulary to indicate whether the collection is document or edge.
* **instances**: This field is a controlled vocabulary set of values that indicates which instances the current term has taken. For instance, a term may represent a controlled vocabulary, in that case this field would contain :instance:enumeration; the same term may, however, also be used as a controlled vocabulary selection element, in that case the field would also contain :instance:selection. This field is used to select terms according to how they were used.
* **info**: This field is an array of URLs providing information resources.
* **schema**: This field is an URL pointing to a resource that represents the term schema. For instance the JSON schema of ISO standards is published in the web, the term that defines the corresponding ISO controlled vocabulary could also include this information.

## Descriptors

The descriptors collection contains all variable definitions, the \_key field value represents the record field name and the other fields contain the data type, definitions and all other information necessary to document and validate values belonging to that variable.

Although the descriptors collection implements the global/local identifier scheme, in this case the global identifier is not copied into the \_key field, this is because the value of this field becomes the name of the record field which may pose problems both because of the characters used and because the value may become long, making the structure of the record inefficient[[13]](#footnote-13).

Since descriptors share most of their properties with terms, we shall only include here those fields that terms do not have, or that have a different function; note that the fields described below are descriptors, which means that the data dictionary uses itself for its definitions.

* **\_key**: This field contains the unique identifier of the descriptor. The value will become the field name associated with that variable definition.
* **nid**: This field contains the \_id reference of the term or descriptor that represents the variable’s *namespace*, this field is required for descriptors.
* **lid:** This field contains the local identifier unique within its namespace (nid). This field is required.
* **rank:** This field is a controlled vocabulary associated with users, it represents the user rank. In descriptors it is used to restrict usage to users having a rank with a level lower or equal to the current session user’s.
* **kind:** This field is a controlled vocabulary that defines the data kind, it indicates whether the value is qualitative, categorical, ordinal, discrete or quantitative.
* **type:** This field is a controlled vocabulary that defines the data type, these have been described earlier.
* **format:** This field is a controlled vocabulary that defines the data format, it can be used to indicate that the descriptor contains a scalar, a list or a set of values of the descriptor’s data type.
* **terms:** This field is a controlled vocabulary that applies exclusively to descriptors of type :type:value:term or :type:value:enum, it is a controlled vocabulary set of terms that define controlled vocabularies. In :type:value:term descriptors it limits the scope of the term reference to one of the controlled vocabularies belonging to the list. The field is used to restrict the choice of term references to a defined set of values.
* **fields:** This field is identical to the terms field above, except that the elements or choices of the controlled vocabulary will not be terms, but gid descriptor references. This field is used to provide a choice of fields.
* **type-key:** This field is an object of type :type:data:struct that is used exclusively by :type:data:object data type descriptors. It serves the purpose of defining the scope and type of the object keys. It *must* contain the type field and any other field that is used to define the data type. For instance multi-language strings have type = :type:value:term and terms holding the list of controlled vocabularies that categorise languages. The data type applies to all property names of the object.
* **type-value:** This field is an object of type :type:data:struct that is used exclusively by :type:data:object data type descriptors. It serves the purpose of defining the scope and type of the object property values. It *must* contain the type field and any other field that is used to define the data type. For instance multi-language strings have type = :type:value:str or another text type variant. The data type applies to all values of the object.
* **size**: This field is a value of :type:value:size type that is used exclusively for lists and sets as the modifier indicating an element count range restriction.
* **size**: This field is a value of :type:value:range type that is used as the modifier indicating the valid range for values of that descriptor.
* **regex**: This field is a string that represents the validation regular expression, it is used by descriptors of type text.
* **unit:** This field is a controlled vocabulary that defines the descriptor values unit.
* **units:** This field is a controlled vocabulary set of unit elements that lists all the other units in which values could be encountered. This field can be useful when harmonising datasets: it can be used to provide users with a choice of unit conversions

1. The idea is that each data domain represents a compatible set of data. While the use of a NoSQL database would allow you to store all types of data into a single collection, when the number of records grows, it is advisable to aggregate similar data structures together to allow specific indexing strategies. [↑](#footnote-ref-1)
2. The process of harmonizing data with the data dictionary has been described extensively in the *SMART Survey Aggregator* background document. [↑](#footnote-ref-2)
3. The management of study annex documents should be performed by the front-end component, ArangoDB does not support a viable binary format that would allow to store physical binary files as records – as MongoDB [GridFS](https://docs.mongodb.com/manual/core/gridfs/) allows. [↑](#footnote-ref-3)
4. ArangoDB has two types of collection: document and edge. This collection is one of the edge collections used in the database, the name may be misleading, this collection only contains edge documents related to toponyms. [↑](#footnote-ref-4)
5. There is only one record in the database that violates this rule: it is the default namespace, this namespace is reserved for default terms and descriptors, the exception is used so that it is not possible to use that namespace in other types of terms and descriptors. [↑](#footnote-ref-5)
6. This reference type exists only because with descriptors the \_key represents a database field name which could be different from the descriptor global identifier. Normally descriptors would have the same structure and identification scheme as terms. [↑](#footnote-ref-6)
7. ArangoDB does not make a distinction between integer and real numbers: everything is a floating point number, as in JavaScript. If a value must be an integer, there is a specific data type for it. [↑](#footnote-ref-7)
8. Only relevant for non-integers. [↑](#footnote-ref-8)
9. The key data type can be any data type derived from the base text data type, it is useful when defining an object whose field names should be taken from a controlled vocabulary. [↑](#footnote-ref-9)
10. Note that there is another reference data type that defines fields that point to the \_id of another object: this type does not derive from string since it adds the collection name in the value which may exceed the string maximum character length. [↑](#footnote-ref-10)
11. This means that the referenced term must exist in branched edge branches field. [↑](#footnote-ref-11)
12. There are many of these procedures that are declared, but do nothing yet, only the essential ones have been implemented: the controlled vocabulary is there as a list of possible options. [↑](#footnote-ref-12)
13. In a previous prototype of this system the \_key of the descriptor was the hexadecimal value of a counter, while the gid represented the “human-readable” representation of the identifier. This allowed to have a large number of descriptors and keep the field name short at the same time. In this version we did not implement this feature, which means that the choice of the \_key is the responsibility of the user defining the data dictionary. [↑](#footnote-ref-13)