

The Metadata solution implemented in MG-RAST

Authors:

Travis Harrison, teharrison@anl.gov

Jared Bischof, jbischof@mcs.anl.gov

Andreas Wilke, wilke@mcs.anl.gov

Folker Meyer, folker@anl.gov

v: 1.0

date: 3/11/2013

1. Introduction

Metadata (data describing data) is in constant flux and while many users struggle to add it to their data, as service providers, we need it to enable users to analyze their data. Think of the ability to “color” an ordination plot using user provided data (e.g. BIOME, sampling location, plot name, replicate name, ..).

Rather than using a **single database with a fixed schema** (that would be too restrictive) or a **complete schema-free approach** (this would be too chaotic and not offer validation), we are using a **hybrid approach**, combining the strength of both.

The hybrid approach enables us to capture arbitrary key value pairs **and** have blessed and controlled subsets of the data ⇒ the best of both worlds. Figure 1 shows the general layout

1.1 Scenario points

- the world for data and metadata consists of files that travel in pairs.
- we believe in minimal checklists driven by the practitioners in the field not heavy handed enforced standards designed by committees
- we represent “truth” as a graph where edges describe some transformations and nodes are project, sample, or data.
- A project may contain many samples, for each sample there may be several data sets.
Example: A study of 4 soil samples that each have a 16s amplicon sequence set, a shotgun metagenome, a meta transcriptome and a meta-proteome associated with it.
 - 1 project (including Title, abstract, PI, admin contact, URL, PubMed)
 - 4 samples (GPS coordinates, soil extraction method...)
 - 16 data sets (for sequence data: DNA/mRNA extraction, library creation, sequencing parameters; for non sequence data similar)
- Controlled vocabularies (CV) enable validation (in addition to format or range checks)
 - we use the term CV to mean both controlled vocabularies and some formats (e.g. temperature in Celsius, distances in the metric system, GPS)
- We can implement different levels of metadata stringency for different types / fields.
 - moderately strict: (several required fields with CV terms)
 - enables comparison
 - enables data discovery using CV (e.g. all samples from “marine sediment”)

- loose: (required fields, no CV terms)
 - you can compare to some extent
 - very limited data discovery using text search (some samples will not be found)
- very loose (everything is valid)
 - no comparison without human intervention (e.g. meters vs. feet in depth or Fahrenheit vs. Celsius in temp)
 - no data discovery can't find data ⇒ each user defines separate fields, temp vs temperature, or depth in either meters or feet

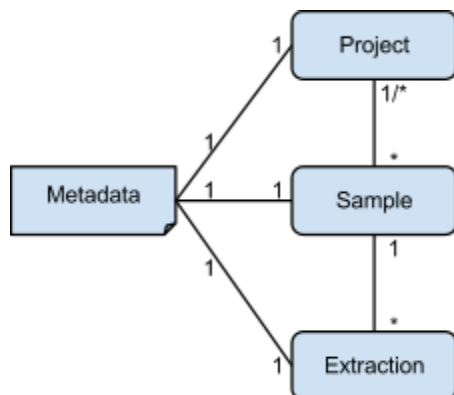


Figure 1: A three tiered hierarchy allows modeling of real life experiments, combining multiple “samples” into a “project” (or study) and allowing for different “extractions” from the sample.

2. Workflow

When data is uploaded (or after the fact) a separate metadata file is created. Using a simple spreadsheet (generated by Metazen in most cases) users can capture minimal metadata required for standard compliance (currently GSC MIxS).

Users can submit their metadata spreadsheet for an entire project with many samples and extractions for validation. For later download spreadsheet and XML/GCDML versions can be created.

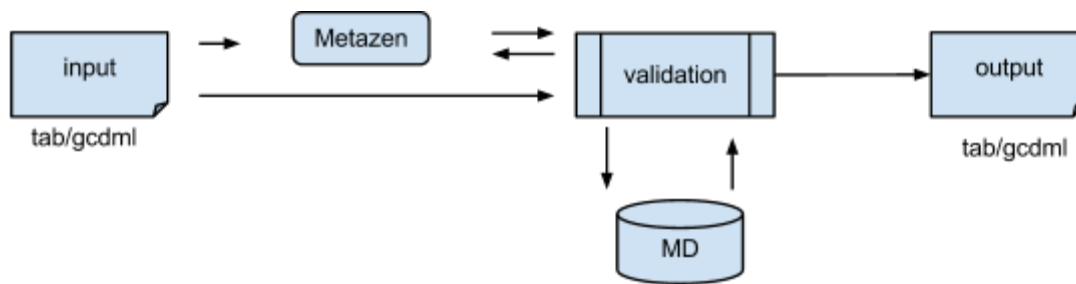


Figure 2: All data files are accompanied by metadata. The tuples of payload file(s) and a metadata file are uploaded together and the metadata file is potentially validated. Separate metadata files describe the data, different “schemata” described Project, sample and data, however the lower tiers can inherit data from the higher tiers i.e. Project and sample.

3. Implementation

We use two schemas, a metadata template mechanism, and a validator for controlled vocabulary.

We rely on external controlled vocabularies e.g. BioOntology.org via web-services or downloaded flat files.

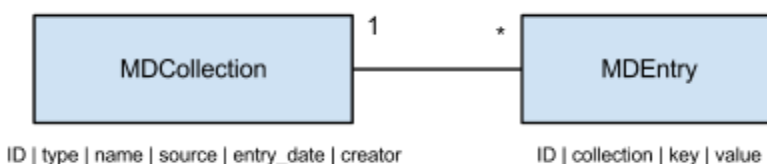
3.1. Schema 1: Basic data

1. Generalized metadata storage

This is our “capture all” storage mechanisms, enabling us to capture arbitrary key value pairs. This is important in a field that is currently discovering the need for metadata capturing and the tools for that.

We use a very simple schema: (id | collection | key | value) to store arbitrary metadata on any data type. By (optionally) organizing metadata into collections we can for each individual collection:

- require certain fields (Keys)
- validate the use of controlled vocabulary terms (when required for certain values) for given fields.



type: Project, Sample, Library, Microarray, Proteome, ...

Figure 3: The backend storage mechanism provides two different tables. MDEntry to store all data and MDCollection to group entries into sets.

3.2 Schema 2: A meta-metadata database

This database describes the required fields and the controlled vocabularies for each “namespace|domain”. Example domains are GSC MlxS, GSC MIMS.

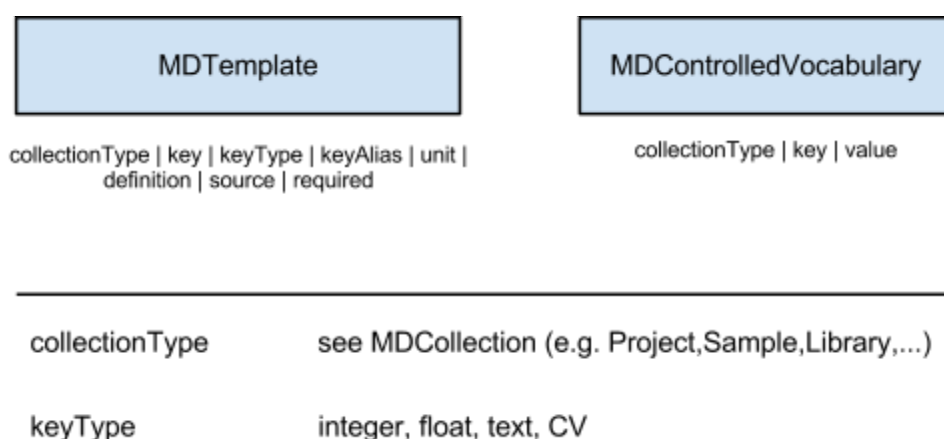


Figure 4: The meta-metadata database provides essential two tables describing structure and syntax of certain meta data. For a given metadata collection type the MDTemplate table provides information about metadata categories, value format for the category and source name (e.g. MIAME). MDControlledVocabulary contains valid values for defined keys with type CV in MDTemplate.

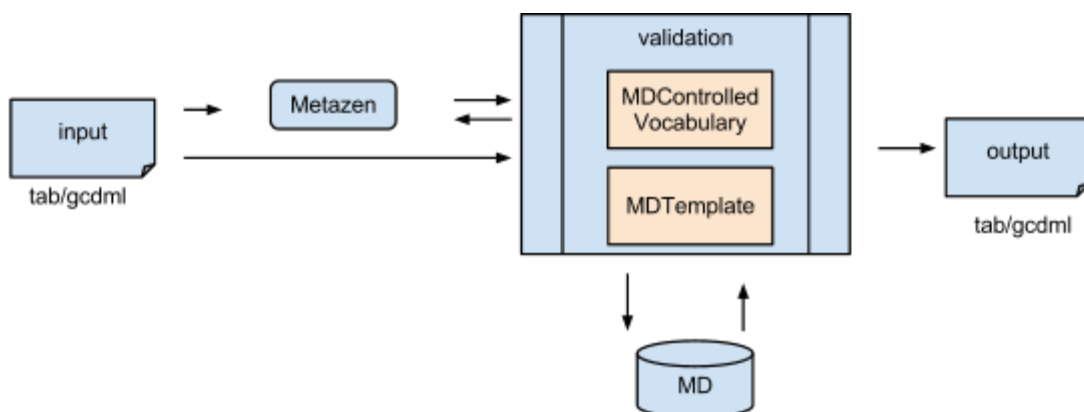


Figure 5: Using MDTemplate and MDCV we enable checking for required keys and controlled

vocabulary terms.

3.3 Example for a metagenome in cartoon

sd

3.4 Example for a metagenome and a 16s data set from the same sample

sd

4. Metazen a metadata validation tool

4.1 Recording metadata

Metazen is an interactive tool that aids users in recording their metadata in a complete and valid format. A defined set of mandatory fields captures vital information while the option to add fields provides flexibility. Project and user level information is stored so users can reload that information without having to enter it repeatedly. The following figure shows what the user sees when they log into Metazen.

The screenshot shows the Metazen web interface in a browser window. The browser's address bar displays `http://metagenomics.anl.gov/metazen.cgi`. The page header features the **MG-RAST** logo and the text "metagenomics analysis server". Navigation links include **LOGIN**, **REGISTER**, **PASSWORD**, and **FORGOTT?**, with a **login** button. The main content area is titled **MetaZen** and contains the following sections:

- ABOUT THIS TOOL:**

This tool is designed to help you fill out your metadata spreadsheet. The metadata you provide, helps us to analyze your data more accurately and helps make MG-RAST a more useful resource.
- Terminology:**
 - sample - a single entity that has been obtained for analysis
 - library - a prepped collection of DNA fragments generated from a sample (also, in this case, corresponds to a sequencing run)
 - environmental information - the characteristics which describe the environment in which your samples were obtained
 - sample set - a group of samples sharing the same library and environmental characteristics

This tool will help you get started on completing your metadata spreadsheet by filling in any information that is common across all of your samples and/or libraries. This tool currently only allows users to enter one environmental package for your samples and all samples must have been sequenced by the same number of sequencing technologies with the same number of replicates. This information is entered in tab #2 below. If your project deviates from this convention, you must either produce multiple separate metadata spreadsheets or generate your spreadsheet and then edit the appropriate fields manually.
- PREFILL FORM:**

To prefill the project tab with your information please select from the contact status options below and click 'prefill form'. Optionally, you can also prefill other project fields and the number of samples in the form by selecting a previous project below before clicking the 'prefill form' button.

Select your contact status:

Select your previous project from which to prefill the form:
- ENTER METADATA**
 - 1. enter project information
 - 2. enter sample set information

The next figure indicates fields that have been pre-filled by the stored user information, required fields versus optional fields, and fields that are validated (Note: the PI Organization is a required

field that was left blank).

The screenshot shows a web browser window titled 'MetaZen' with the URL 'http://metagenomics.anl.gov/metazen.cgi'. The page is titled 'ENTER METADATA' and has a sub-header '1. enter project information'. Below this, there is a paragraph of instructions: 'Please enter your contact information below. It is important that the PI and technical contact information (if different than the PI) is entered so that if a technical contact is no longer available, the PI can still gain access to their data. Note that selecting 'other' under 'Project Funding' enables an additional text field where you can type in your funding source. Required project fields are marked with a red asterisk (*) and must be entered before continuing on to the rest of the form. The selected (checked) optional fields will be included in your spreadsheet whether you complete those fields now or after download.'

The form is divided into two main sections: 'Principal Investigator (PI) Information:' and 'Technical Contact Information:'. Each section has several fields, some of which are required (marked with a red asterisk) and some are optional (checked with a blue checkmark).

Principal Investigator (PI) Information:

- ☒ PI e-mail^[?]: folker@anl.gov
- ☒ PI First Name^[?]: Folker
- ☒ PI Last Name^[?]: Meyer
- ☒ PI Organization^[?]: [blank]
- ☒ PI Org Address^[?]: 9700 South Cass Avenue, Argonn
- ☒ PI Org Country^[?]: USA
- ☒ PI Organization URL^[?]: http://www.anl.gov

Technical Contact Information:

- ☒ Contact E-mail^[?]: wilke@mcos.anl.gov
- ☒ Contact First Name^[?]: Andreas
- ☒ Contact Last Name^[?]: Wilke
- ☒ Organization^[?]: ANL
- ☒ Organization Address^[?]: [blank]
- ☒ Organization Country^[?]: [blank]
- ☒ Organization URL^[?]: http://www.anl.gov

dbXref ID's:

Below you can enter project ID's from different analysis tools so that your dataset can be linked across these resources.

- ☒ Greengenes Project ID^[?]: [blank]
- ☒ MG-RAST Project ID^[?]: [blank]
- ☒ NCBI Project ID^[?]: [blank]
- ☒ QIIME Project ID^[?]: [blank]
- ☒ VAMPS Project ID^[?]: [blank]

Project Information:

[blank]

Complex, controlled vocabularies are more easily navigated and chosen through Bioportal widgets seen in the following figure.

MetaZen

http://metagenomics.anl.gov/metazen.cgi

# of samples	environmental package	# of shotgun metagenome libraries per sample	# of meta transcriptome libraries per sample	# of amplicon metagenome (16S) libraries per sample
9	soil	1	1	0

submit

3. enter environment information

ENTER ENVIRONMENT INFORMATION

Use three different terms from controlled vocabularies for biome, environmental feature, and environmental material to classify your samples. Note that while the terms might not be perfect matches for your specific project they are primarily meant to allow use of your data by others. You can enter your detailed project description in the project tab at the top of this form.

Biome

Environment Ontology

Find: Large river delta biome Go

- biome
 - aquatic biome
 - freshwater biome
 - Large lake biome
 - Large river biome
 - Large river delta biome**
 - marine biome

enter selected term as biome

Large river delta biome

Environmental Feature

Environment Ontology

Find: Go

- water body
 - well**
 - wetland
- physiographic feature
- habitat
- marine feature
- mesopelagic abundant biota

enter selected term as env feature

well

Environmental Material

Environment Ontology

Find: Go

- sediment
 - anaerobic sediment**
 - biogenous sediment
 - boulder sediment
 - clay sediment
 - cobble sediment
 - sediment

enter selected term as env material

anaerobic sediment

4. enter sample information

5. enter shotgun metagenome information

6. enter meta transcriptome information

click to generate spreadsheet

Users can search for the latitudinal and longitudinal coordinates of where their samples were obtained and enter valid metadata into strictly formatted fields (e.g. date, time, timezone, decimal versus integer values, and other controlled vocabularies). Some of these options are depicted in the next figure.

4. enter sample information

ENTER SAMPLE INFORMATION

Please enter only information that is consistent across all samples. This data will be pre-filled in the spreadsheet.

Required sample fields are marked with a blue asterisk (*), because unlike required project fields, they can be entered after downloading your spreadsheet.

The selected (checked) optional fields will be included in your spreadsheet whether you complete those fields now or after download.

Date/Time Information:

* Collection Date^[?]: 2012-08-08

* Collection Time^[?]: 14:58:30

* Collection Timezone^[?]: (UTC+6:00) Bangladesh

Location Information:

By entering an address or location below we can search google maps to try and identify the latitude and longitude of your location. If you find that this does not work for your location, you can try using Google Maps [here](#) and the instructions [here](#) to identify the latitude and longitude of your desired location.

* Location/Address^[?]: Khulna Division Bangladesh

Search for Latitude/Longitude

Searched the location: Khulna Division Bangladesh

Top Google Result: Khulna Division, Bangladesh

Latitude and Longitude for this address were entered below. If Google returned address is not correct, please try searching again or refer to the Google Maps links above.

* Latitude^[?]: 22.8087816

* Longitude^[?]: 89.2467191

* Country/Water body^[?]: Bangladesh

Optional Fields: ☒ Select all ☐ Deselect all

☒ Altitude^[?]: meters

☒ Biotic Relationship^[?]:

☒ Continent^[?]:

☒ Depth^[?]: meters

☒ Elevation^[?]: meters

☒ pH^[?]: (decimal value)

☒ Rel to Oxygen^[?]:

☒ Sample ID^[?]:

☒ Sample Collect Device^[?]:

☒ Sample Size^[?]:

☒ Temperature^[?]: celsius (decimal)

[Click to view/hide more fields](#)

Other sample information:

4.2 Editing and validating metadata

Once users are finished entering their metadata into Metazen, a spreadsheet can be generated for download and further editing. Often times users will want to edit their spreadsheet manually if they have many samples and/or sequencing runs for which they would like to enter specific metadata on a per sample or per run basis. Having both the web and spreadsheet to interact with, provides users with guidance to follow standards and flexibility to extend and modify their metadata. Once they are finished editing their metadata, we have a validation tool to help users identify metadata errors, and to ensure that the submitted spreadsheet contains the vital (required) metadata fields and completely valid data.

5.0 Extensions to the current state

5.1 Model for adding new data types

e.g. GWAS, metaproteomes or microarrays

We believe that adding other data types should be easy, the fields in the spreadsheets might look fundamentally different, but the general approach will translate just fine.

5.2 Model for addition of new domain specific metadata

The GSC metadata mechanism is plug-in enabled. Domain scientists can create new required subsets of terms (and the corresponding CV etc). to capture more data on their field of study. Several of these “environmental packages” have been created already and voted into standards by the GSC board.

5.3 Extending existing CVs (locally)

The slowness of the curators of the various CVs leads to user unhappiness in many areas, why are NIH researchers required to list Terrestrial Biome for their Human microbiome samples? Adding a local “ontologies on the fly” extension capability could fix that. This capability basically amounts to adding the “other” field that users can type in IFF their sample/data does not fit into the existing categories.

A number of conditions need to be met in the implementation:

- We’d need to ensure that users only do this after trying to find a good CV term for their data. The default user action here is to select the path of least resistance and select other and type in “blah”. Users are forced to classify their samples and their instinct (as evidenced by hundreds of users) is to provide the most specific classification more akin to a sample name than a sample classification. E.g. “camel” gut vs. “other mammalian” gut.
- Other users should be presented with the new “candidate CV” terms before they can attempt to create their own. This way we can avoid creating too many candidate terms
- We need a procedure for interacting on candidate terms with the owners of the CVs. The alternative here is to become CV curators ourselves which we should try to avoid at all cost.
- Validation and export need to recognize the candidate CV terms.