GEOME Documentation

This guide gets you started with using GEOME. Part 1 gives you the basics of GEOME, instructing first time users on how to use GEOME. Part 2 gives more depth and background on how GEOME is organized and important contextual information on how GEOME is structured. Part 3 is lower level technical documentation. Part 4 contains frequently asked questions about the software. If, after reading this manual, you still have questions, please feel free to email geome.help@gmail.com.

Part 1: Getting Started With GEOME

Introduction

GEOME is a web-based database which captures metadata on biological samples, used primarily for biodiversity inventories, population studies, and environmental metagenomics. GEOME assigns persistent identifiers for all samples and sampling events and specifies the set of metadata attributes which capture the who, what, where, and when associated with all samples. GEOME provides instant feedback to users on the quality of their data while loading. It enables use of data in laboratory information system (LIMS), and connects to the Biocode LIMS plugin. GEOME also packages submissions for easy delivery to the Sequence Read Archive (SRA) and Genbank's Nucleotide database.

Quick Start

If you want to get started quickly with GEOME, the process is very simple, and consists of two steps:

1. Create an account
2. Create a project

About Teams, Projects and Expeditions

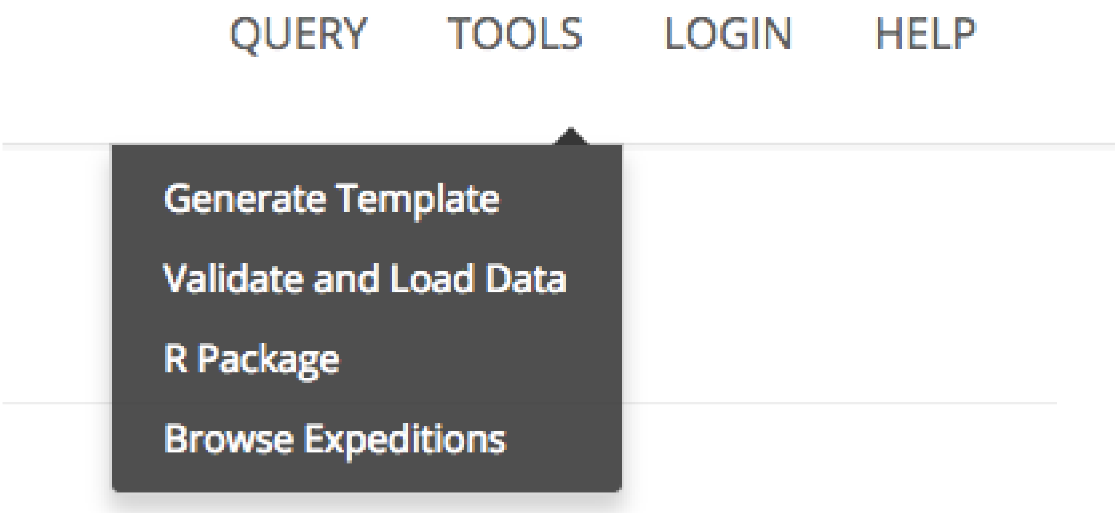
Teams: A team is a specialized research focus with settings relevant to all members. Teams enable users to create projects having a common set of rules, attributes, and controlled vocabulary terms. When you create a project in a team workspace you aagree to use ALL of the attributes, rules, and controlled vocabularies for that team. The team administrator controls all configuration options. To create a project within a team workspace, select "Join team workspace" during the project creation process and then select the appropriate team.

Projects: Please note that all projects have one owner, who may invite additional members. Each of the members in turn can create expeditions within a project. Read further to understand how expeditions work.

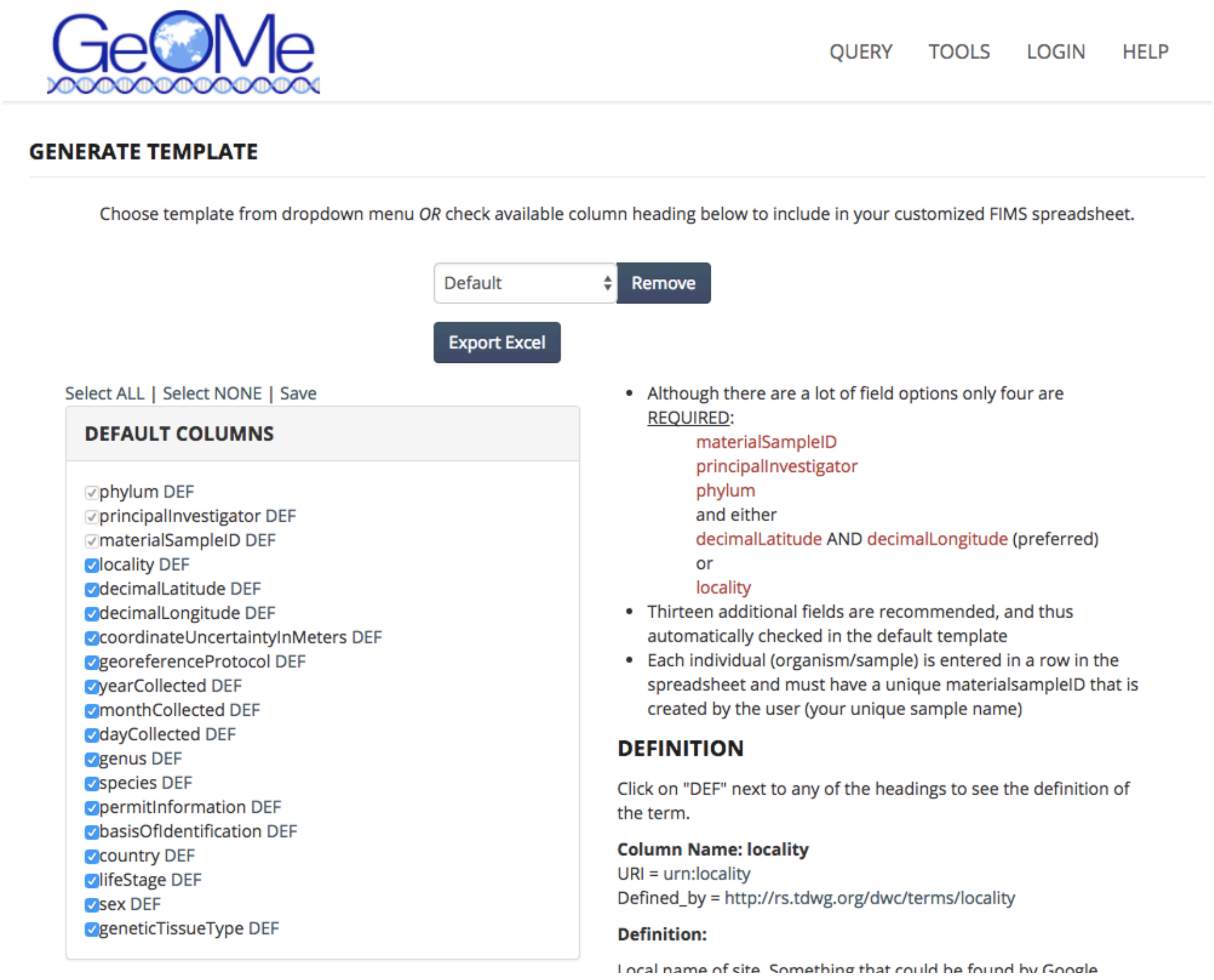
Expeditions: Projects are composed of one or more expeditions. An expedition corresponds to a single spreadsheet, containing all related events, samples, and tissues. All data entered into GEOME must be entered as an expedition. Any member of a project may create an expedition when they first upload a spreadsheet. The expedition owner retains the right to update or alter expedition data as well as setting the expedition to public or private viewing. The project owner also has the capability to alter expedition metadata of any user within the project. Expedition identifiers can be set as unique either within the expedition itself or across the project. Finally, each expedition provides a globally unique and resolvable prefix (expedition root identifier) for each entity. When a local identifier, which is enforced as unique either within an expedition or project, is appended to the expedition root identifier, it services as a resolvable and globally unique representation for each instance of a collecting event, sample, or tissue. The provision of these identifiers happens automatically, and is noted within the system as BCIDs (biocode commons identifiers).

Generate Template

Sample metadata is recorded on an Excel Spreadsheet and you can create and customize your own templates under “Tools -> Generate Template”



On the Generate Template page, you can select columns that you want to include on your spreadsheet. Click on the “DEF” link beside each column name to view the definition of the column name. Columns that are pre-checked and shown in grey, indicate that they are mandatory fields and not able to be unchecked. Columns that are pre-checked and shown in blue indicate they are suggested and can be unchecked. Once you have checked the columns you wish to include in your spreadsheet, press the “Export Excel” button to download an Excel Spreadsheet which you can then use to fill in Sample Metadata.



GENERATE TEMPLATE

Choose template from dropdown menu *OR* check available column heading below to include in your customized FIMS spreadsheet.

DEFAULT COLUMNS

☑ phylum DEF
☑ principalInvestigator DEF
☑ materialSampleID DEF
☑ locality DEF
☑ decimalLatitude DEF
☑ decimalLongitude DEF
☑ coordinateUncertaintyInMeters DEF
☑ georeferenceProtocol DEF
☑ yearCollected DEF
☑ monthCollected DEF
☑ dayCollected DEF
☑ genus DEF
☑ species DEF
☑ permitInformation DEF
☑ basisOfIdentification DEF
☑ country DEF
☑ lifeStage DEF
☑ sex DEF
☑ geneticTissueType DEF

- Although there are a lot of field options only four are REQUIRED:
  materialSampleID
  principalInvestigator
  phylum
  and either
  decimalLatitude AND decimalLongitude (preferred)
  or
  locality
- Thirteen additional fields are recommended, and thus automatically checked in the default template
- Each individual (organism/sample) is entered in a row in the spreadsheet and must have a unique materialsampleID that is created by the user (your unique sample name)

DEFINITION

Click on "DEF" next to any of the headings to see the definition of the term.

Column Name: locality
URI = urn:locality
Defined_by = http://rs.tdwg.org/dwc/terms/locality

Definition:

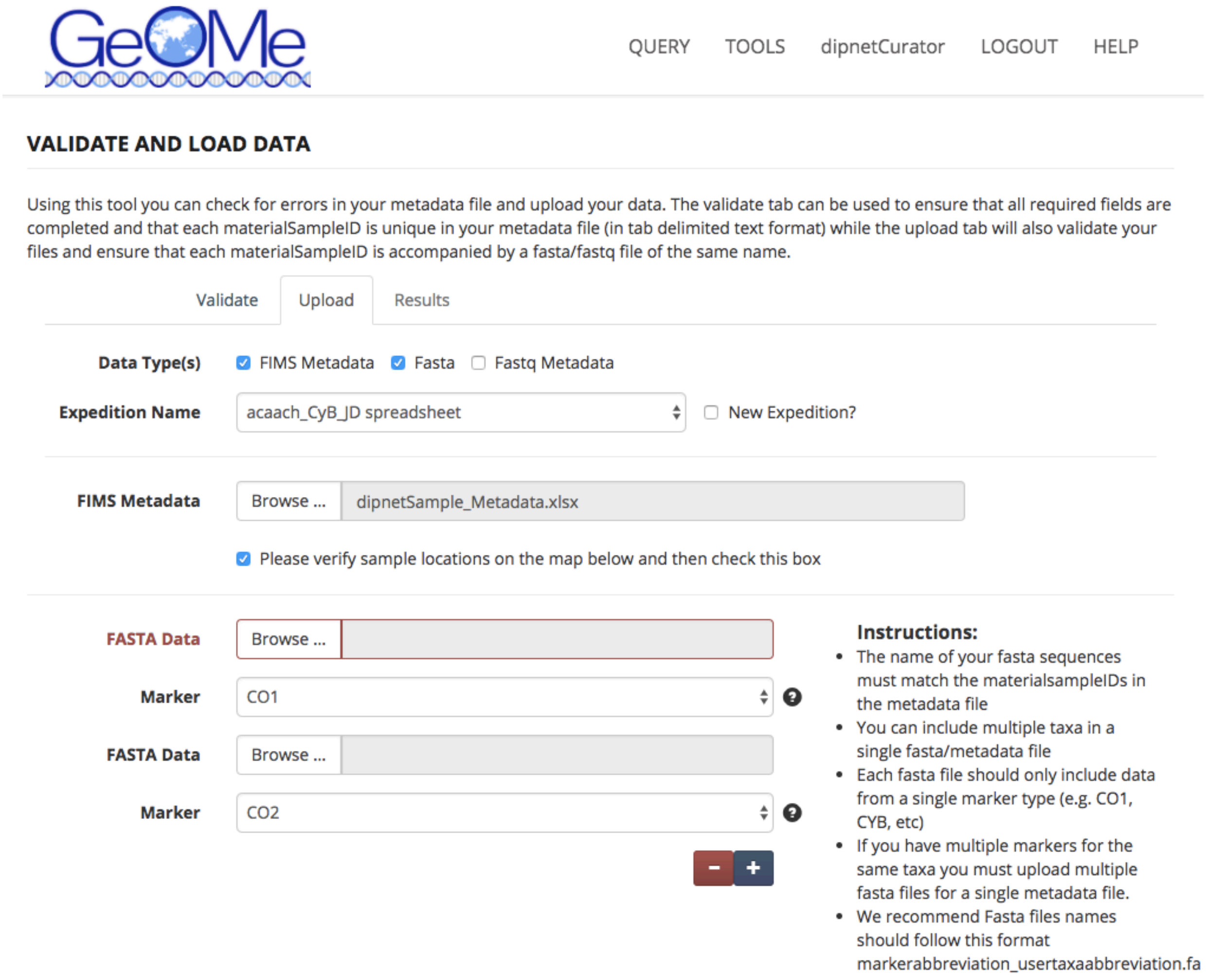
Local name of site. Something that could be found by Google

Validate and Load Data

The Validate and Load Data option can be found under “Tools -> Validate and Load Data”. The first step is validating your sample metadata. Use the Browse button to browse for your file and select the “Validate” button. After data validation, you can Upload your dataset and include just the metadata or include FASTA or FASTQ metadata.

FASTA Upload Example

You must create, or select a pre-existing expedition name for your dataset before continuing. Select your FIMS Metadata file, along with a FASTA filename and a Marker name. After selecting the FIMS Metadata file, you must check a box stating that you have visually verified the sample locations on the map at the bottom of the page. The name of your FASTA sequences must match the sample identifiers in the metadata file. Each FASTA file should only include data from a single marker type. If you have multiple markers for the same taxa you must upload multiple FASTA files for a single metadata file, which can be added by clicking on the "+" button.



VALIDATE AND LOAD DATA

Using this tool you can check for errors in your metadata file and upload your data. The validate tab can be used to ensure that all required fields are completed and that each materialSampleID is unique in your metadata file (in tab delimited text format) while the upload tab will also validate your files and ensure that each materialSampleID is accompanied by a fasta/fastq file of the same name.

Data Type(s)	☑ FIMS Metadata ☑ Fasta ☐ Fastq Metadata
Expedition Name	acaach_CyB_JD spreadsheet ☐ New Expedition?
FIMS Metadata	Browse ... dipnetSample_Metadata.xlsx
	☑ Please verify sample locations on the map below and then check this box
FASTA Data	Browse ...
Marker	CO1
FASTA Data	Browse ...
Marker	CO2

Instructions:

- The name of your fasta sequences must match the materialsampleIDs in the metadata file
- You can include multiple taxa in a single fasta/metadata file
- Each fasta file should only include data from a single marker type (e.g. CO1, CYB, etc)
- If you have multiple markers for the same taxa you must upload multiple fasta files for a single metadata file.
- We recommend Fasta files names should follow this format markerabbreviation_usertaxaabbreviation.fa

FASTQ Uploading

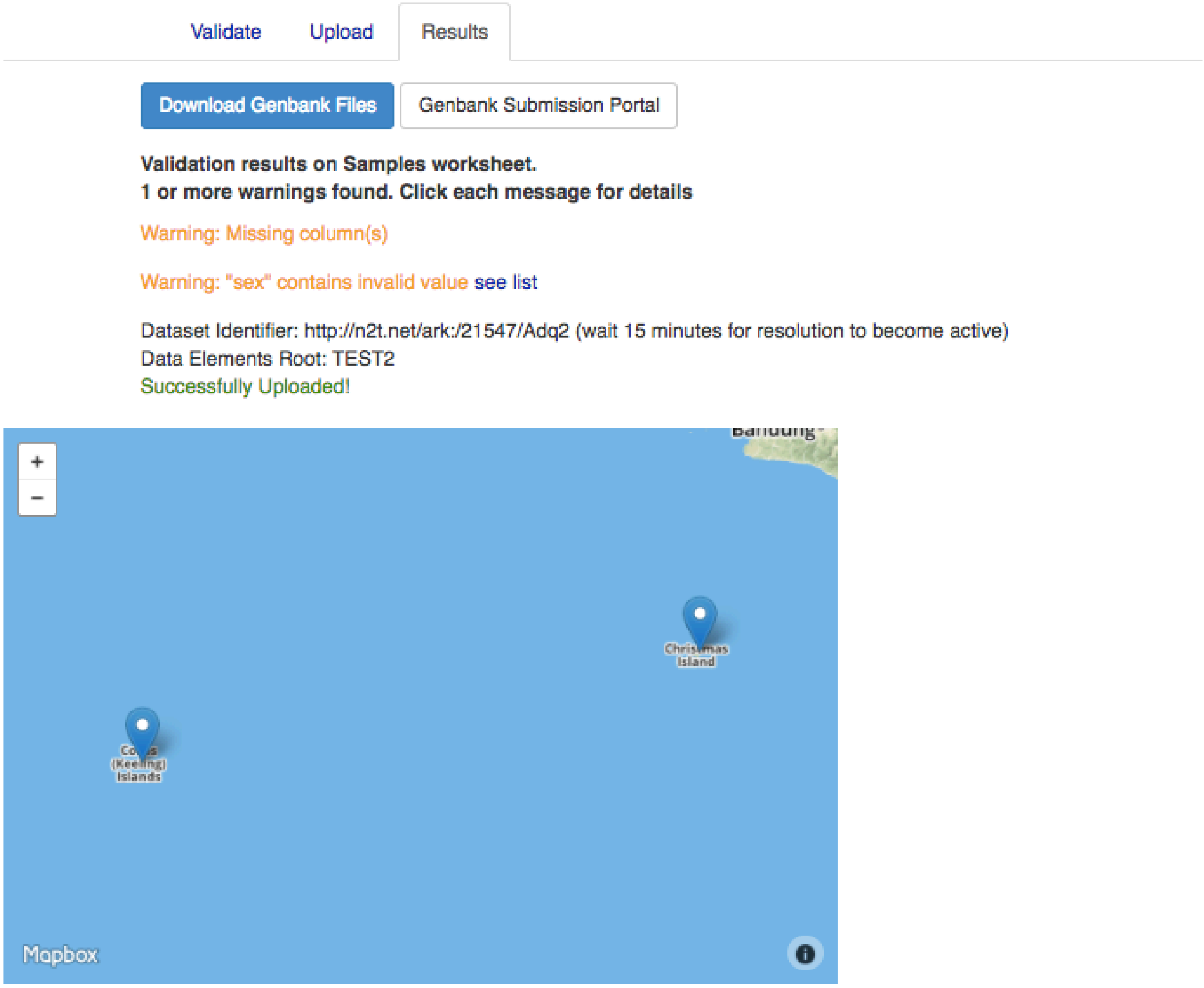
For teams that support the FASTQ data option, GEOME bundles sample and tissue metadata with FASTQ filenames on your local file system to prepare a submission to NCBI's Sequence Read Archive. This submission package will contain pointers to all of your raw sequence data along with validated metadata. Following this method, will ensure you have the highest quality metadata, along with persistent identifiers for all of your samples attached to your submission.

When you check the FASTQ Metadata box on file upload, you will be prompted to upload a file that contains all of the filenames of your FASTQ raw reads that live in your local file system. The following points should be followed when constructing your FASTQ metadata file:

- Each line contains a unique filename
- Each FASTQ filename should contain reads from a single individual
- Filenames listed in the FASTQ filenames file must start with the tissueID column value in your Sample metadata file. For example, given a tissueID value of 'sample1': **Single End** filename values will be sample1.fastq.gz or sample1.fq.gz, and **Paired End** filenames must have either a (-1 or -2) or (-F or -R) immediately preceding the *.fastq.gz or *.fq.gz names, and valid values will be sample1-1.fq.gz, sample1-2.fq.gz OR sample1-F.fastq.gz, sample1-R.fastq.gz
- For the DIPNet team, the materialSampleID and tissueID are synonyms. Other projects and teams allow for multiple tissueIDs for each materialSampleID. Thus, we always match on the tissueID.
- Once uploading is complete the FIMS system will produce two files (SRA metadata and BioSample attributes files) that will ease the upload process to NCBI's Short Read Archive (SRA). When these files are downloaded a set of simple instructions are included that will speed your SRA submission.

- The actual FASTQ sequence files will not be uploaded here and stored on the FIMS system. Instead the metadata file will be uploaded and stored here.
- If you get an error when uploading your FASTQ filenames, you may need to convert your FASTQ filename to ASCII see https://github.com/biocodellc/geome-db/issues/41 for a resolution

Once you have validated and uploaded FASTQ file, a screen is presented that shows you two buttons and your validation results. One button enables you to download pre-generated Genbank submission files. The second button is available which opens a browser window taking you to Genbank's SRA Portal.



Validation results on Samples worksheet.
1 or more warnings found. Click each message for details

Warning: Missing column(s)

Warning: "sex" contains invalid value see list

Dataset Identifier: http://n2t.net/ark:/21547/Adq2 (wait 15 minutes for resolution to become active)
Data Elements Root: TEST2
Successfully Uploaded!

NOTES:

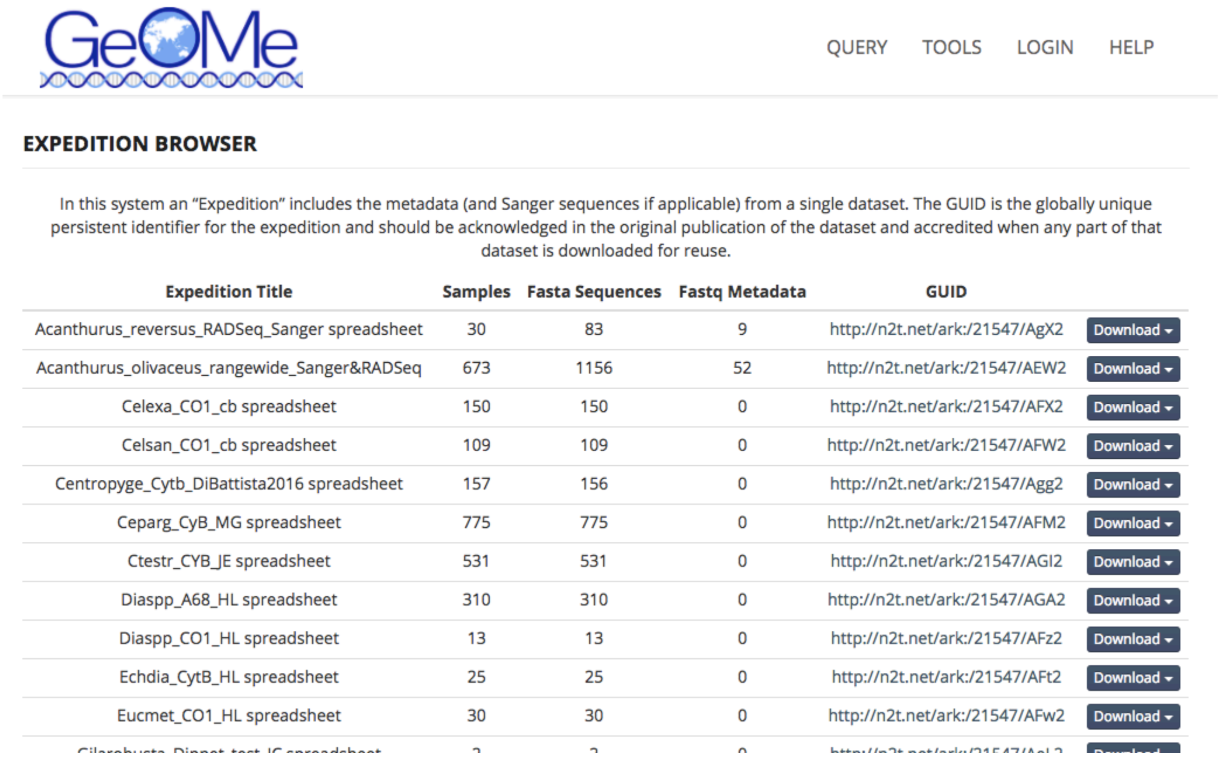
- NCBI does not allow colons in identifier file names when uploading: they will be replaced with an underscore.
- If you have existing sample data in NCBI with the same identifier you will need to work with NCBI directly to update your data.

GEOME R Package

A link is available under the tools menu which takes you to the GEOME R package github page, located at https://github.com/biocodellc/geomedb. More instructions are available at that link.

Browse Expeditions

The "Browse Expeditions" option shows all available uploaded expeditions that are part of GEOME. This page shows you the number of samples, FASTA sequences, and FASTQ metadata provided for each sample. Here you have the option of downloading CSV, FASTA, or FASTQ formatted metadata.



EXPEDITION BROWSER

In this system an "Expedition" includes the metadata (and Sanger sequences if applicable) from a single dataset. The GUID is the globally unique persistent identifier for the expedition and should be acknowledged in the original publication of the dataset and accredited when any part of that dataset is downloaded for reuse.

Expedition Title	Samples	Fasta Sequences	Fastq Metadata	GUID	
Acanthurus_reversus_RADSeq_Sanger spreadsheet	30	83	9	http://n2t.net/ark:/21547/AgX2	Download ▾
Acanthurus_olivaceus_rangewide_Sanger&RADSeq	673	1156	52	http://n2t.net/ark:/21547/AEW2	Download ▾
Celexa_CO1_cb spreadsheet	150	150	0	http://n2t.net/ark:/21547/AFX2	Download ▾
Celsan_CO1_cb spreadsheet	109	109	0	http://n2t.net/ark:/21547/AFW2	Download ▾
Centropyge_Cytb_DiBattista2016 spreadsheet	157	156	0	http://n2t.net/ark:/21547/Agg2	Download ▾
Ceparg_CyB_MG spreadsheet	775	775	0	http://n2t.net/ark:/21547/AFM2	Download ▾
Ctestr_CYB_JE spreadsheet	531	531	0	http://n2t.net/ark:/21547/AGI2	Download ▾
Diaspp_A68_HL spreadsheet	310	310	0	http://n2t.net/ark:/21547/AGA2	Download ▾
Diaspp_CO1_HL spreadsheet	13	13	0	http://n2t.net/ark:/21547/AFz2	Download ▾
Echdia_CytB_HL spreadsheet	25	25	0	http://n2t.net/ark:/21547/AFt2	Download ▾
Eucmet_CO1_HL spreadsheet	30	30	0	http://n2t.net/ark:/21547/AFw2	Download ▾
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Query

The GEOME query interface enables users to filter on geographic information, any word string as part of the metadata (e.g. "Moorea"), Darwin core terms, expedition names, or any other column that is part of the GEOME specification. The Query interface returns results either in map form or table form, selectable by clicking on the "Map" or "Table" buttons on the upper right corner of the interface. The "Download" link enables metadata download of the queried results.

The "Query Entity" field lets you select which entity you want to focus your query on. If you choose one of the photo entities, you may choose to browse photos (button available on the right pane).

Accession Numbers and Sample Identifiers

When you submit your work for publication you may be asked for Genbank accession numbers, dataset identifiers, or even sample identifiers. GEOME creates identifiers for physical samples and datasets, as well as automatically syncing sequence read archive SRA numbers. The following information describes how to handle these identifiers.

Physical Sample Identifiers

As you may have seen, you can obtain a globally unique form of the materialSampleID in the "bcid" column at then end of the row of metadata when you download a CSV file and it looks like: https://n2t.net/ark:/21547/CXs2MBIO1040

Dataset/Expedition Identifiers

You can find dataset identifiers by going to "Tools -> Browse Expeditions" and you'll see a column called "GUID" that if you click on will bring you to information about your expedition.
E.g. https://n2t.net/ark:/21547/Apc2

Sequence Identifiers

For nextgen sequences that have followed the GEOME path described in this document you can enter the resolvable GUID for the materialSample and find links to the BioProject and BioSample identifier, e.g. check out the following record:

http://n2t.net/ark:/21547/le2Acaoli_CAS44

GEOME currently doesn't link Genbank Accession identifiers for FASTA data submissions, so these will need to be researched independently.

Uploading your data to the NCBI Short Read Archive (SRA)

After submitting your metadata to GEOME two files will be produced: the bioSample-attributes.tsv and the sra- metadata.tsv files and you will be directed to SRA to upload your data. There are several steps but the creation of those two files will streamline the process significantly!

If you don't already have a NCBI account you will need to create one. If you do have an account then sign in using the tab at the top right corner of the page.

After you sign in start a new submission

Step 1: Submitter
Enter your personal information

Step 2: General Info
You will be asked two important questions here:

1. Did you already register a BioProject for this data set?

2. Did you already register BioSamples for

this data set?

In the majority of cases the answer to both

questions will be NO

The following instructions are based on the user answering "NO" to both of the above questions.

Step 3: Project Info

Fill in project information. For example:

Project Title: Acanthurus_reversus_RADSeq_data

Project Description: RADSeq data for the reef fish Acanthurus reversus Relevance: Evolution

Is your project part of a larger initiative that is already registered with NCBI?

Most likely No

External links: Add if relevant

Select your grants: If relevant

Step 4: Biosample type

Here you choose your sample type. Most DIPnet members will check either "Invertebrates" OR "Model organism or animal sample" for vertebrates.

Step 5: Biosample attributes

Upload the bioSample-attributes file (.tsv) produced by GeOMe. You may see additional warnings or error messages produced by the SRA validator. You must fix error messages. In some cases, you may safely ignore warnings. For example, we have seen cases for users working in marine system where locality is often based on nearby terrestrial locations, and the SRA responds with a warning that the locality is invalid since it is located in the warning. This particular message may be ignored for marine users where this is intentional.

Step 6: SRA metadata

Check the Upload a file option and ppload the sra-metadata file (.tsv) produced by GeOMe

Step 7: Files

Follow the directions on SRA and upload your files. You will be asked to download the latest version of Aspera Connect. This will speed upload tremendously. Once Aspera is installed go directly to the Choose Files option, choose your zipped folder, and Aspera will automatically open.

Step 8: Overview Submit!

Part 2: More About How GEOME Works

Minimum information requirements

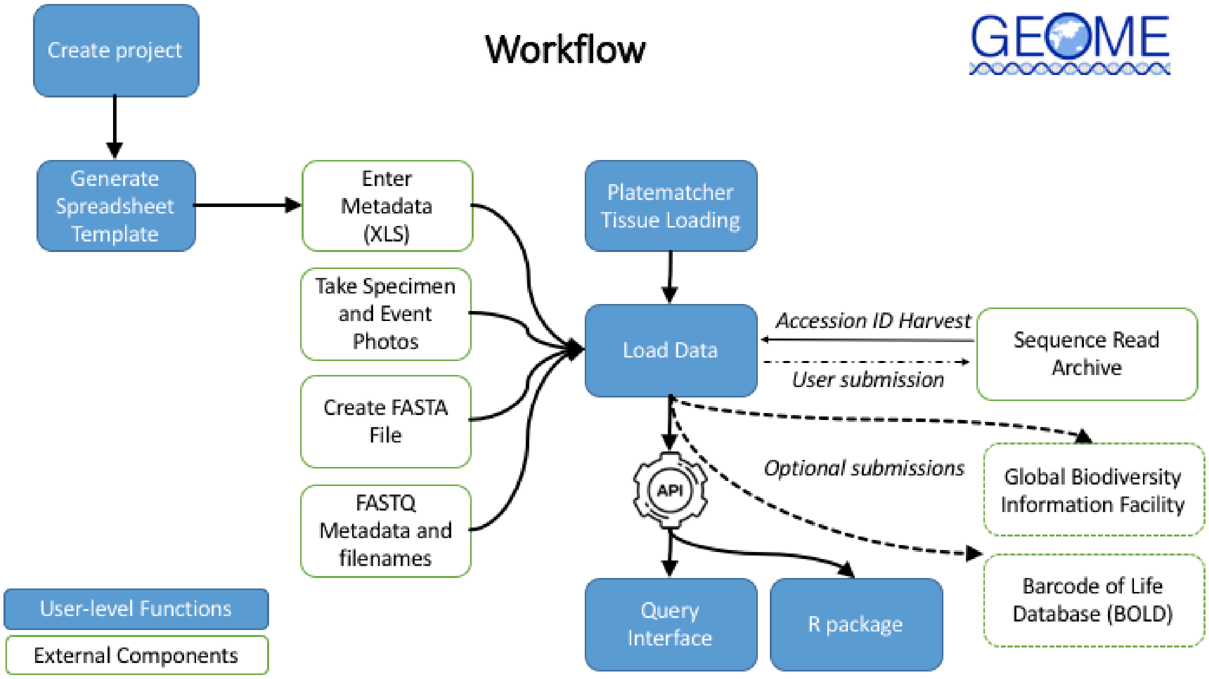
GEOME requires the following fields to be entered for ALL projects:

- materialSampleID
- yearCollected
- country
- locality

Each Team can set its own requirements for fields. The following table illustrates the minimum information requirements for teams and projects:

	Fields	Biocode Team	Other Teams
Required Fields	materialSampleID	required	required
	locality	required	required
	yearCollected	required	required
	country	required	required
Additional (Example) Attributes	decimalLatitude	required	configurable
	decimalLongitude	required	configurable
	phylum	required	configurable
	institutionCode	required	configurable
	enteredBy	required	configurable
	kingdom	optional	configurable
	scientificName	optional	configurable
	principalInvestigator	optional	configurable
Format In Which You Load Data		**Individual Sheets for Event, Samples, Tissues, and Photos**	**Configurable**

Workflow: How information is managed in the GEOME environment



The GEOME R Package

The GEOME R package is used to retrieve GEOME data for analysis. Please visit our github page to run the current code. The GEOME R package is no longer available under CRAN and only available using the github installer.

The Biocode LIMS Plugin

All data uploaded to GEOME can be manipulated in a laboratory information management system (LIMS) using a specially built LIMS plugin that operates within the Geneious environment. The purpose of the LIMS tool is to help manage lab and sequence analysis workflows.Use the LIMS plugin wiki to learn about how to use the LIMS.

- **Field Database Connection:** GEOME FIMS
- **Host:**https://api.geome-db.org/
- **Username/password:** User your GEOME username and password to access your data in the LIMS system.

Video Library

https://youtu.be/WyJKmFsUVKc (Instructions for the FuTRES team… Using GEOME to load data)

https://youtu.be/GX-2Zk9MVws (This video discusses uploading data, logging in, creating a new project, generating spreadsheet templates, and loading data.)

https://youtu.be/hYlyqtW-bn8 (This video talks about download functions including querying and working with GEOME in the GEOME R package.)

https://youtu.be/cuAN9LbDO-U (This video talks about GEOME's technical architecture, and how metadata is handled.)

History

GEOME has its roots in the Moorea Biocode project database, developed from 2006 to 2011 to support data collection for the Moore Foundation funded Moorea Biocode Project: an all taxa biotic inventory of a single tropical island involving 6 teams, 50 researchers, and thousands of collecting events. The Moorea Biocode field information management system, also known as "FIMS1", was developed to ingest spreadsheet data from researchers working on the project and employed data validation on data ingest. The tools were written in Perl and Java. FIMS1 has been running from 2006 through 2018 and developed tools such as the Plate Matcher (to easily map tissues to 96 well plates), the bioValidator (for loading and validating spreadsheets), and a web interface for managing data.

From 2012 to 2015, the National Science Foundation funded the BiSciCol (Biological Sciences Collections) project. While BiSciCol focused on identifiers, ontologies, and semantic technologies for linking biodiversity data, the primary use case was to work on a solution for linking events, samples, and tissues across many different systems using linked data technology with the Moorea Biocode Project data integration with member institution databases. BiScicol brought us the following products: a clearer understanding of the role of persistent identifiers, the BiSciCol triplifier, the development of the Biological Collections Ontology (BCO), and the development of FIMS2, also known as the BiSciCol FIMS. The BiSciCol FIMS adopted ARK identifiers (through California Digital Library's EZID system) for all samples, events and tissues. In addition, all data inserted into BiSciCol FIMS was backed by a Fuseki triplestore with metadata configurations specified using an XML configuration file, stored and managed separately by each project.

From 2014-2017, BiSciCol FIMS began hosting the Diversity of the IndoPacific Network (DIPNet) to develop a coherent metadata repository for assembling metadata from across the IndoPacific region. DIPNet added the ability to upload FASTA and FASTQ metadata to the FIMS2 site and also developed the name "GEOME" (eventually becaming the brand name for FIMS3). During this time-frame, another FIMS system was developed for the Smithosnian National Museum of Natural History ("NMNH FIMS") with the goal of creating a centralized field data ingestion system for all NMNH field data. The NMNH FIMS has since been discontinued with SI users using the BiSciCol FIMS (Barcode of Wildlife Project, ARMS portal, and Global Genome Initiative) and FIMS1 installations (Invertebrate Zoology and LAB applications).

Beginning in 2016, John Deck and RJ Ewing started work on a redesigned FIMS system, utilizing a Postgres backend to control project management features and JSON metadata objects to store data and configurations. This FIMS system (FIMS3) assumed the name GEOME. It incorporates features from FIMS1 (including photo uploading, specimen, tissue, and event pages, plate-matching tools) along with features of the BiSciCol FIMS (persistent id generation, flexible project configuration with attributes stored in a non-relational system guided by a configuration file). In addition, FIMS3 brings in the concept of a network which governs multiple configuration templates and the ability to easily create projects through a user interface. The goal of GEOME is to integrate all previous FIMS1 and FIMS2, with FIMS1 and FIMS2 projects removed from their current hosting environment by Spring of 2019.

GEOME was hosted at the University of Florida before 2018, and at the University of Arizona (CyVerse) from 2019 through 2025, and from 2025 to current hosted at Indiana University (Jetstream2).

Data Usage Policy

GEOME has the following objectives:

- To advance genetic diversity research worldwide
- To aggregate sample and genetic data in raw formats in a searchable database such that original datasets can be utilized for further investigation.
- To promote and advocate open and collaborative science as a best practice for conducting biodiversity research
- To promote and provide capacity-building within developing countries for monitoring, study and protection of their biodiversity resources.

Recalling that access to and utilization of genetic resources and data taken should be consistent with the provisions of the Convention on Biological Diversity (CBD) taking into account their specifications by the Bonn Guidelines on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits arising from their Utilization, and, where appropriate, the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits arising from their Utilization (NP),

Recalling that according to these provisions non-monetary and/or monetary benefits from the utilization of the genetic resources shall be shared with the Country of Origin if the same so requires and as it is set out in mutually agreed terms,

Acknowledging that research and development on genetic resources can be for the public domain (non-commercial) or for commercial purposes and,

Recalling that according to these provisions, non-commercial research purposes may contribute to the conservation and sustainable use of biodiversity

By uploading data to or downloading data from the GEOME database, GEOME contributors and Recipients of GEOME data agree as follows:

ARTICLE 1 – UPLOADING OF LEGAL AND ACCURATE GENETIC DATA AND ASSOCIATED METADATA

By uploading Genetic Data and associated metadata to GeOME database, Contributors certify the following:

1.1 The Genetic Resources from which Genetic Data and Associated Metadata were derived were collected under appropriate and legal access permits or their equivalent by each Country of Origin, and

1.2 All co-authors on the data's primary publication have consented to share the data in the GEOME database.

ARTICLE 2 – USAGE OF DOWNLOADED GENETIC DATA AND ASSOCIATED METADATA

2.1 The Recipient shall be entitled to the Use of the Genetic Data and Associated Metadata for the Public Domain.

2.2 Should the Recipient intend to utilize the Genetic Data and Associated Metadata for Commercial Purposes they will seek consent of the Country of Origin, subject to the mutually agreed terms between the GEOME contributor and Country of Origin.

2.3 When used for Commercial Purposes or in the Public Domain, data should be recognized as follows: "Data were made available through GEOME (https://geome-db.org/)".

2.4 Every attempt should be made to cite the original studies having contributed data to the new analyses. Attribution would include any publications first describing these data as well as reference to electronic depositions (such as DOI associated with dataDryad). This information may often be found in the "associatedReferences" field of the database.

2.5 The Recipient may transfer downloaded Genetic Data and Associated Metadata to Subsequent Recipients provided that Subsequent Recipients agree in writing to use the data under the terms of this agreement.

2.6 GEOME does not undertake to monitor the rights of any Country of Origin for their Genetic Resources downstream from our database portal.

Steering Committee

The GEOME Steering Committee is responsible for guiding the development of GEOME software, its implementation, and setting strategic direction. The GEOME steering committee consists of Neil Davies (UC Berkeley), John Deck (UC Berkeley and Biocode, LLC), Rob Toonen (University of Hawaii), Cynthia Riginos (University of Queensland), Libby Liggins (Massey University), Chris Meyer (Smithsonian

National Museum of Natural History), Rachel Toczydlowski (USDA Forest Service), and Michelle Gaither (University of Central Florida).

Part 3: Technical Documentation

Creating Local Identifiers

A crucial part of the GEOME is converting local identifiers that you construct and use in your own research, and turning these into globally unique, resolvable identifiers. Globally unique identifiers are created by appending your local identifier onto a unique root that is generated for every resource within every expedition. Examples of locally unique identifiers are “Grinnell1213”, “MooreaEvent2”, or “MBIO56_1”.

Each identifier that is minted will be resolvable via HTTP using California Digital Library’s Name-to-thing resolver. Since the name-to-thing resolver is sensitive to certain characters, we have limited the characters that are suitable for use as local identifiers. Allowable characters are validated on data load so if you choose an invalid character you will get an error message. The following are the allowed local identifier characters:

- A-Z
- a-z
- 0-9
- + (plus)
- = (equals)
- : (colon)
- . (period)
- _ (underscore)
- ((open parantheses)
-) (close parantheses)
- ~ (tilde)
- * (asterisk)

The following are valid identifiers: “MVZ:Herp:1234”, “Grinnell (1234)”

The following would be invalid identifiers: “MVZ-Herp-1234”, “Grinnell/Alexander 1234”

Once data is made loaded and made public, you can search for your newly minted globally unique and resolvable identifiers in the Query page, and they will be listed under the “BCID” column. If the identifier is shown as “ark:/21547/R2MBIO564” you can substitute “http://n2t.net/ark:” for the “ark:” to make a a resolvable identifier as `https://n2t.net/ark:/21547/CXs2MBIO564`_, where MBIO564 is the locally uinque identifier.

GEOME Queries

GEOME provides a custom sql-like query syntax to help you find the data you need. The following documentation supplements the Swagger Applicaiton Programming Interface. By default, the query terms are executed against all columns in the project. To execute a query against a specific column, you can construct the query in the form columnName:query. The full text search query japan would return all results where a column contains the word japan. Where as the full text search query column1:japan would return all results where column1 contains the word japan.

All queries can be constructed using the sql operators *AND*, *OR*, and *NOT* as well as groupings within (); The query _expeditions_:myExpedition and not japan would return all results in the *expedition* myExpedition which do not contain the word japan.

Below you will find more information about the supported queries.

Supported Queries

The following queries are supported:

- full text search
- comparison
- project
- `expedition`_
- `exists`_
- like
- phrase
- range
- select

full text search

This the default query, and will perform a search on the tokenized version of the uploaded data.

col1:value - will perform a fts on col1 for value col1:val* - will perform a fts on col1 for words starting with val value - will preform a fts on all columns for value

comparison

This query is used to compare 2 values. The following operators are supported:

NOTE: for correct comparison results when using <, <=, >, >=, the Attribute dataType should be one of (Integer, Float, Date, Datetime, Time). This can be set via the project configuration. Talk to your project administrator about this.

> = - equals <> - not equals > - greater then >= - greater then or equal to < - less then <= - less then or equal to

project query

This query is will filter the results based on the project(s) that they belong to.

The query `_projects_:1` would return everything uploaded under project `1` The query `_projects_:[1, 2]` would return everything uploaded under project `1` or `2`

expedition query

This query is will filter the results based on the expedition(s) that they belong to. Note: as expeditions are only unique within a project, you most likely want to specify a project query as well.

The query `_expeditions_:myExpedition` would return everything uploaded under `myExpedition` The query `_expeditions_:[myExpedition1, myExpedition2]` would return everything uploaded under `myExpedition1` or `myExpedition2`

exists query

This query returns results where a column has a value.

The query `_exists_:column1` would return all results where `column1` has a value. The query `_exists_:[column1, column2]` would return all results where `column1` or `column2` has a value.

like query

This query performs a sql `ILIKE` (case-insensitive `LIKE`) query.

`col1:"%value"` - `col1 ILIKE '%value'`

phrase query

This query performs a sql `ILIKE` (case-insensitive `LIKE`) query.

col1:"some value" - col1 ILIKE '%some value%'

range query

This is a shorthanded way to perform a comparison query.

NOTE: for correct comparison results, the Attribute dataType should be one of (Integer, Float, Date, Datetime, Time) This can be set via the project configuration. Talk to your project adminstrator about this.

col1:[1 TO 10] - >= 1 AND <= 10 col1:[1 TO 10} - >= 1 AND < 10 col1:{1 TO 10} - > 1 AND < 10 col1:{* TO 100] - <= 100

select query

Used to select related parent/child data along with the queried entity. The provided value should be the conceptAlias of the Entity to select. The provided conceptAlias' do need to be related to the query entity, but do not need to be directly related. For example, if you are querying a parent entity, you can also select the grandChildren and the grandParents. Any combination of related entities can be selected.

NOTE: _select queries should not be preceded/followed by and or or keywords and can not be preceded by the not keyword.

select:parentEntity - selects both child and parent entity results for the query _select_:[parentEntity, grandParentEntity] - selects both child and parent entity results for the query

Tokenization

Text fields go through a tokenization process before they are indexed. This process attempts to breakdown text into words and numbers as well as converting words to their normalized form.

Tokenization Ex:

```
"many donkeys" -> ["many", "donkey"]
```

For more information, you can view the psql tokenization.

Installation

This content is for people wishing to install GEOME on their own server.

Details

GEOME consists of a core set of Java classes and REST services. Developers have a choice of interacting with the REST services running _BCID, which has built in EZID minting capabilities, or running their own instance of _BCID and installing their own EZID instance requiring a purchase of an EZID account.

To run an instance of FIMS you will need the following components:

- A unix-based server * A java servlet container e.g. Tomcat, Glassfish, Jetty * Connection to a BCID service

Installation and Build – Migrating an existing installation

- Source code is available on this site via github
- Building is done via an Gradle build file (provided as part of the distribution)
- a properties file needs to be configured by copying biocode-fims.template to biocode-fims.props (in the root directory of the distribution)

Install the following software

- postgres
- jetty9
- java8
- bcid and geome-db repositories (from github)

Properties file

- update properties files in src/main/environment/production

gradle war deploy build/libs/geome-db.war (do the above for both bcid and geome-db) generate openapi document using gradle resolve

Configuration Files

GEOME has a network level configuratino file which defines network level rules and all available data properties and entities. Each project has its own configuration file as well which supplements the GEOME network configuration file. All configuration files are written in JSON This is where the projects specific configuration is specified. This includes resources, attributes, validation rules, and relations.

Attributes

DataType

Each attribute may specify a dataType. A dataType can be specified to provide additional validation, and in the case of date, datetime, and time, can be used for data formatting. This is especially helpful for standardizing the data to aid in querying and analysis.

The following dataType are supported:

- String (default if not specified)
- Integer
- Float
- Date
    - must specify dataformat as well
- Time
    - must specify dataformat as well
- Datetime
    - must specify dataformat as well

Record

Fims validation and upload is based around the concept of a Record. A Record is a single instance of an Entity.

A Record is typically a k:v map of properties. The key should be the columnUri. It is the responsibility of the DataReader implementation to map any columnName -> columnUri when creating an instance of a Record.

Each type of Record will have a RecordValidator implementation that is responsible for handling the validation of that Record type. The default Record type is a GenericRecord. A GenericRecord is the most common in the fims system will be. The validation for a GenericRecord is strictly controlled by the project configuration w/o any additional validation logic.

currently we have support for the following types:

- GenericRecord
- FastaRecord
- FastqRecord
- PhotoRecord

RecordSet

A collection of Record instances.

Dataset

A collection of RecordSet instances. If a Dataset has any RecordSet`s for a child `Entity, then the Dataset will contain the both the parent and child RecordSet`s. The `DatasetBuilder should be used to help construct a valid Dataset instance.

Data Readers

DataReader implementations contain the logic for reading and converting a specific file type into a RecordSet (TODO: more info about RecordSets). When a file is uploaded for validation, it is passed to the DataReaderFactory which will return the appropriate DataReader implementation for the provided file. A DataReader should return true when handlesExtension is called if that reader can handle the provided ext.

A current limitation of DataReaders is that if multiple DataReader implementations handle the same file ext, only 1 can be enabled at a given time. This restriction may be lifted in the future.

TODO more info about current DataReader implementations

Entity

Custom entities can be created and must subclass the Entity class. All subclasses must exist in the biocode.fims.digester package to be properly registered as a valid subtype for polymorphic serialization/deserialization via Jackson. An Entity subclass provides the ability to fix certain parts of a given entity, as well as provide additional validation logic (to be executed on ProjectConfig updates) to ensure the entity is well formed and not missing any pertinant information.

REST Services

FIMS REST Services are available at: http://www.biscicol.org/apidocs/

Versioning

FIMS REST Services are now versioned. v1 is the default version. You may specify the version by including the header:

```
Api-Version: {version}
```

or via the url:

```
http://biscicol.org/biocode-fims/rest/{version}/...
```

We currently support the following versions:

- v1
- v1.1

more info about the specific version resources to come...

User Accounts

User accounts are not required to lookup/resolve BCIDs. However, they are required to work with projects, expeditions, or create new BCIDs. Here we describe how to obtain a user account for Biocode-

Account Creation

User accounts can be created by either by the Biocode-Fims instance owner or by project administrators. Project administrators can add any existing user in the Biocode-Fims system as an authorized expedition creator. Talk to your project administrator to be added to a particular project.

[https://github.com/biocodellc/biocode-fims-commons/wiki/OAuth2 Information about Open Authorization]

Project Administrators

Project administrators are set by the Biocode-Fims instance owner upon request. There is only one designated project administrator per project. The project administrator can add, create, and remove users, set the location of the validation XML file, and define the project abstract.

oauth2

All developers need to register their app. Please contact the system admin to register. You will be issued a client_id and client_secret. The client_secret should be kept private.

Authorization

Client app will make a GET request to /id/authenticationService/oauth/authorize. This request will contain the following query parameters:

- client_id (Required) - The client_id your app was issued during when registered.
- redirect_uri (Required) - The absolute URI you would like the response directed to.
- state (Optional) - Will be returned, unmodified, in the response.

The response will contain the following query parameters:

- code - The random 20 character string used to exchange for an access_token. This code expires in 10 mins and can only be used 1 time.
- state - Only if this parameter was included in the request.

Access Token

Client app will make a POST request to /id/authenticationService/oauth/access_token. This request will contain the following parameters in the request body:

- client_id (Required) - The client_id your app was issued during when registered.
- client_secret (Required) - The client_secret your app was issued during when registered.
- code (Required) - The authorization code received in the authorization request.
- redirect_uri (Required) - The absolute URI you would like the response directed to. Must be identical to the redirect_uri provided in the authorization request.
- state (Optional) - Will be returned, unmodified, in the response.
- grant_type (Optional) - If grant_type is “password”, and a username and password is provided, the username and password will be used for authentication. If authentication is successful, an access_token and refresh_token will be returned
- password (Optional) - Required if grant_type is “password”.
- username (Optional) - Required if grant_type is “password”.

The JSON response will contain the following parameters:

- access_token - The random 20 character string used to access a user's profile.
- refresh_token - The random 20 character string used to obtain a new access_token. This expires after 24 hrs.
- token_type - currently we only issue bearer tokens.
- expires_in - the number of seconds the token is good for.
- state - Only if this parameter was included in the request.

Refresh Token

Client app will make a POST request to /id/authenticationService/oauth/refresh. This request will contain the following parameters in the request body:

- client_id (Required) - The client_id your app was issued during when registered.
- client_secret (Required) - The client_secret your app was issued during when registered.
- refresh_token (Required) - The refresh_token you were issued with you access token.

The server will validate the refresh token and if the refresh token is less then 24 hrs old, a new access token will be issued. The current refresh token will be expired and a new one will be issued.

The JSON response will contain the following parameters:

- access_token - The random 20 character string used to access a user's profile.
- refresh_token - The random 20 character string used to obtain a new access_token. This expires after 24 hrs.
- token_type - currently we only issue bearer tokens.
- expires_in - the number of seconds the token is good for.

API Access

In order to obtain a user's profile information, make a GET request to /id/userService/profile with the access_token as a query parameter.

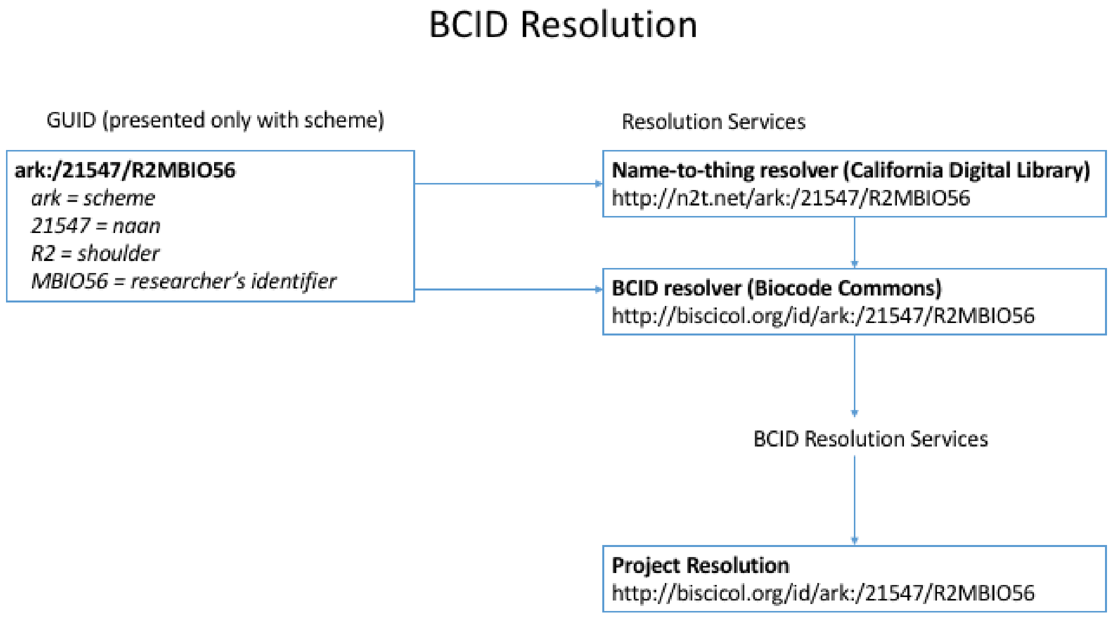
If the token is still valid, you will receive a JSON response with the following user information:

- firstName
- lastName
- email
- institution
- userId
- username
- projectAdmin
- hasSetPassword

We also support access to any rest services on behalf of the user. Just append "?access_token=your_access_token" to the url in order to access the service.

Resolution System

The following illustration shows how BCIDs work with local identifiers, the world wide web, and EZID's name-to-thing resolution service. A field researcher uses their own numbering system (e.g. 'MBIO56'), and uploads their data to FIMS, which assigns it to a resource category (e.g. 'R2'). The FIMS system itself is registered under the ark: scheme, and has a name assigning authority number (NAAN) of 21547. Resolution requests coming through name-to-thing are re-directed to the BCID resolution service.



BCID Resolution

GUID (presented only with scheme)

ark:/21547/R2MBIO56
ark = scheme
21547 = naan
R2 = shoulder
MBIO56 = researcher's identifier

Resolution Services

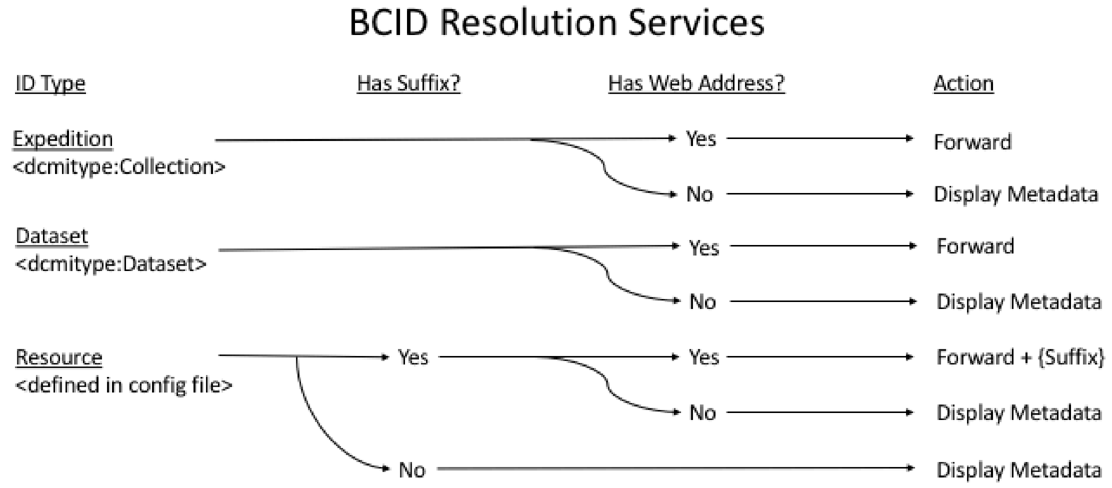
Name-to-thing resolver (California Digital Library)
http://n2t.net/ark:/21547/R2MBIO56

BCID resolver (Biocode Commons)
http://biscicol.org/id/ark:/21547/R2MBIO56

BCID Resolution Services

Project Resolution
http://biscicol.org/id/ark:/21547/R2MBIO56

The following chart shows how BCID resolution works for expeditions, datasets, and resources in the FIMS system with actions falling under forwarding, or metadata display. Forwarding behaviour is determined by either the specification of a target webaddress in the database, or absent that, a specification in the project's configuration file.



BCID Resolution Services

Forward Logic:

```
If (bcid.webaddress != null) return bcid.webaddress;   // From database
else {
    If (ID Type = Expedition) return <metadataParam.expeditionForwardingAddress>{ark};          } Uses apache
    else if (ID Type != Dataset) return metadataParam.conceptForwardingAddress/{ark}/{suffix};  } strSubstituor
    else return "Display Metadata Address"
}
```

Types Of Identifiers

FIMS uses a centralized minting service to assign identifiers for three types of identifiers: expeditions, datasets, and resources. The three types of identifiers are described below.

Each FIMS system installation must use its own name assigning authority number and register with California Digital Library's EZID service to mint Archival Resource Keys (ARKs).

Expedition identifiers

- resourceType: http://purl.org/dc/dcmitype/Collection
- Mutable, representing the most current version of a particular spreadsheet
  - Metadata:
    - expeditionCode
    - expeditionTitle
    - userId (who created this expedition)
    - ts (when loaded)
    - projectId (project this belongs to)
    - public (public or not)

Dataset identifiers

- resourceType: http://purl.org/dc/dcmitype/Dataset
- Immutable
- Belongs to a specific expedition
  - Metadata:
    - webAddress (where this dataset can be found, in its native format, depending on installation)
    - userId (who uploaded this dataset)
    - doi (an optional doi, in addition to the created ARK)

Resource identifiers

- resourceType: defined in configuration file
- Belongs to an expedition. Multiple resources may be specified for each expedition.

- Implements suffix-passthrough feature to identify individual resources within each dataset. For example, a single "Material Sample" identifier is created for each expedition. If the expedition has 1000 rows representing physical samples, 1000 identifiers can be resolved by appending a locally unique suffix on to the Resource Identifier root.
- A resource identifier plus the locally unique primary key loaded for the most recent dataset in an expedition forms the globally unique identifier for a particular resource

Part 4: Frequently Asked Questions

Getting Started Questions

What is GEOME?

The Genomic Observatories Metadatabase (GEOME, www.geome-db.org) is a web-based database which captures metadata for biological samples.

How can I make accessioning my data easier in future uploads to GEOME?

We recommend you use the metadata template as your standard sample record keeping and management spreadsheet when doing field work and lab work. The template can be modified as much as you like (column order, extra fields added, etc.) as long as fields mandatory for upload to GEOME remain unaltered, and are filled. To upload the data in your modified template to GEOME, you would need to save it as either an Excel spreadsheet following the format that is delivered under "Generate Template" or from downloading expedition metadata. The format may also be CSV with a single row of column headers. Be sure to check the box appropriate for your format on data upload.

When I publish data uploaded to GEOME for the first time, what do I report about dataset accessibility in my publication?

The GUID is the globally unique persistent identifier for the expedition (i.e. dataset) metadata and should be acknowledged in the original publication of the dataset and accredited when any part of that dataset is downloaded. For example, you might write: 'All FASTQ sequence files are available from the GenBank at the National Center for Biotechnology Information short-read archive database (accession number: [*your numbers*]). Associated metadata are also available at GEOME (GUID https://n2t.net/ark:/[*five numbers that represent the project*]/[*alphanumeric code that precedes the materialSampleIDs*])' or something similar for sanger sequence (FASTA) or microsatellite datasets. To find the GUID related to your own datasets, go to Workbench, and then 'My Expeditions'. Your GUID for a dataset is beside 'Identifier'.

How do I cite, or acknowledge use of GEOME?

The original GEOME publication can be cited as follows: "The Genomic Observatories Metadatabase (GEOME): A new repository for field and sampling event metadata associated with genetic samples", John Deck , Michelle R. Gaither, Rodney Ewing, Christopher E. Bird, Neil Davies, Christopher Meyer, Cynthia Riginos, Robert J. Toonen, Eric D. Crandall Published: August 3, 2017.

How do I reference the dataset in my project, team or in GEOME?

Use the GUID - this is a globally unique persistent identifier for the expedition (i.e. dataset) metadata which should be acknowledged in the original publication of the dataset and accredited when any part of that dataset is downloaded. To find the GUID related to your own datasets, go to Workbench, and then 'My Expeditions'. Your GUID for a dataset is beside 'Identifier'. If you would like the GUID for several datasets, go to 'Project Overview' (from Workbench, for the relevant project). Here you will see a table of all the datasets under 'Expedition Title' and each will have a GUID associated.

How do I access my project or sample metadata once uploaded to GEOME?

Your project metadata will be permanently accessible through the GEOME web interface query function, either using a geographic bounding box or search terms relevant to the metadata you provided in the metadata template. It will exist as a unique *expedition* with a globally unique persistent identifier known as a *GUID* (e.g. GUID https://n2t.net/ark:/[*five numbers that represent the project*]/[*alphanumeric code that precedes the materialSampleIDs*]) and each sample (*materialSampleID*) will also have a permanent unique identifying code called a *BCID*. To find the GUID related to your own datasets, go to Workbench, and then 'My Expeditions'. Your GUID for a dataset is beside 'Identifier'. If you would like the GUID for several datasets, go to 'Project Overview' (from Workbench, for the relevant project). Here you will see a table of all the datasets under 'Expedition Title' and each will have a GUID associated.

Why can't I download all the genetic data for the query I made in GEOME?

GEOME is for the metadata associated with that sequence data, rather than the genetic data itself. You can use GEOME to discover what genetic data exists based on your query and this will provide you with metadata for those genetic sequences, including their accession numbers in NCBI/GenBank. You can then query the NCBI/GenBank repository to gather the genetic data. In some cases the FASTA files may be provided to you based on your GEOME query, this is because the person uploaded their FASTA files to GEOME alongside their metadata (this is optional when uploading metadata).

What are the advantages of depositing metadata in GEOME?

GEOME assists in the deposition of raw genetic data to INSDC's sequence read archive (SRA) while maintaining persistent links to standards-compliant ecological metadata held in the GEOME database. This approach facilitates findable, accessible, interoperable and reusable data archival practices (adhering to the FAIR guiding principles). Moreover, GEOME enables data management solutions for large collaborative groups, and expedites batch retrieval of genetic data from the SRA. These advantages and several others are fully described in Deck et al. (2017) and Riginos et al. (2020).

What are derived data?

Derived data are any data that have been created through an analytical process or pipeline from the original raw genomic reads. We are specifically interested in SNP calls (VCF, Genepop, Structure format), but can also take microsatellite or other genetic datasets created for the same samples. What we are specifically looking for URIs for datasets posted online, e.g. in Dryad. Please see instructions below if you are providing more than one set of derived data.

How do I find out more about these metadata initiatives and contribute?

The rationale for, and capability of, the Genomics Observatories Metadatabase (GEOME) is described in Deck et al. (2017) and Riginos et al. (2020). We encourage you to also visit the webpage (www.geome-db.org) and explore GEOME. You are welcome to request to join an existing *project* or *team*, or to create your own. We also invite feedback on current functionality and ways in which GEOME may better accommodate your research needs. If you see the value of GEOME, you can promote its wider use by encouraging colleagues, students, and authors of manuscripts you review or handle as an editor.

Technical Questions

How can i update multiple expeditions at once?

Yes, It is possible to update multiple expeditions at the same time by following the procedure for updating all project data. First, click on “Project Overview” and then select the button labelled “PROJECT CSV ARCHIVE”. This will download a zip archive to your local computer. You will want to make a copy of the archive and saving the original as a backup in case you need to revert changes later. When you extract the second copy, you will find you one or more CSV files (one for each worksheet that you use to load your data). Make updates and changes to these CSV files, making sure to retain the columns “bcid”, “expeditionCode”, and “projectId”: these columns will be used by GEOME to map each sample to the appropriate expedition. Finally, you will want to select “Load Data” on the menu and check the box for each CSV entity that you have updated. Select the files to load and make sure to select “Multiple Expeditions” in the expedition selector.   Press the “LOAD” button and you will be done.

Using a pre-existing materialSampleID formed as a URI

GEOME uses the materialSampleID property to uniquely identify all samples. In GEOME, materialSampleID is a field identifier, assigned by the researcher, and is a basis for creating persistent URIs by appending the provided materialSampleID onto a URI root.   The materialSampleID can on only contain the following characters: [a-zA-Z0-9+=:._()~*]+ This means that characters existing in URIs, such as “/” and “-” are NOT allowed in the materialSampleID field. If you already have a materialSampleID that is formed as a URI, this creates a problem in that you will not be able to load this identifier as a materialSampleID.

The solution is to insert your existing materialSampleID formatted as a URI into the voucherURI field (e.g. http://n2t.net/ark:/65665/34f224976-105e-4392-9d3d-4a4f1f22f048) and then use the suffix of that identifier, or mint a new one, for the materialSampleID field (e.g. 34f224976105e43929d3d4a4f1f22f048).

How do I delete an existing metadata record in GEOME?

Even if you only want to delete a record for a single sample, it is easiest to download the entire expedition in question, edit the contents in Excel, and then select the "replace expedition data" box.

How to update data and what the "Replace Expedition Data" box means?

First, you will want to download GEOME's current copy of your expedition metadata. Goto your Project Overview and then select the expedition you wish to update and download the Excel Workbook.

Make all the required changes in your filled metadata template. Go to 'Load data' in your workbench and make sure you are submitting the metadata to the same project as before. Browse to your revised metadata template – if you are deleting data then you will need to check the box that says 'replace expedition data'. *Please note that you need to retain all columns that you downloaded from GEOME when you re-upload it if you choose this option*. If you are not deleting data then you can leave this check-box unchecked. Select the same 'Expedition' that you used last time. Then go ahead and upload the revised metadata. The process is the same, even if you only intend to upload a FASTA file, or make a FASTQ file ready for upload to SRA.

What does 'concatenated and separated' and 'delimited list' mean?

When there are multiple entries in the same cell, we need a way to programmatically access all the entries and recognise that they are separate entries for the same field. They should be entered sequentially and separated by the pipe '|' symbol. For example in the field 'collectorList' I might want to enter 'Charles Darwin and Alfred Wallace', so I would write 'Charles Darwin | Alfred Wallace'.

What do I do if I have multiple values for a particular metadata field (e.g. *associatedReferences, permitInformation, environmental_medium*, *derivedDataXXX*)

When there are multiple entries in the same cell, we need a way to programmatically access all the entries and recognise that they are separate entries for the same field. For everything except *derivedDataXXX*, multiple values may be added to each field and delimited by a pipe (|).

If you have multiple derived datasets, follow these steps:

1. First fill in all other metadata
2. Highlight and copy all metadata rows, pasting them below the first group of entries.
3. Change ONLY the derived dataset fields to point to the second derived dataset.

4. Continue this process of copying rows and changing derived dataset fields for every derived dataset that you link to.

My samples are from a market OR I'd like to protect the exact geographic location of where the samples were taken - how can I do this and still contribute spatial metadata?

If the sample was purchased freshly killed (as in a fish market) and you can enter the geographic coordinates of the market as sampling location, and set *coordinateUncertaintyInMeters* to 100,000 (100 km). Similarly if the sample was taken from a location that should not be disclosed (for the protection of the population or custodians over that location), we recommend you still provide spatial metadata, but only provide coordinates that are proximal to the sampling location (e.g. set the *coordinateUncertaintyInMeters* to 100,000, and round the actual geographic coordinates).

Working With Sequence Data

Is there a quick way to replace project sequence data that's similar to the metadata replace function?

Yes, you can re-load FASTA sequences using the "Load Data" functions. Make sure that your FASTA file references your tissue identifiers you have previously loaded and update your sequence data. When you load the FASTA file, select the same marker you selected previously and it will re-load FASTA data for this marker.

How can I upload metadata and link to genetic data that are already in the SRA?

It is best practice to first load data into GEOME and then push this data to SRA. This ensures that GEOME and SRA both have the proper links from the start and enables GEOME synchronize metadata with SRA. However, If you already have data uploaded to SRA and want to link your GEOME metadata to SRA, use the following procedure.

1) Adopt a template that uses the Tissue entity.
2) Insert the bioSampleAccession field into the tissue metadata sheet. Note that for some samples, there may be multiple bioSampleAccessions. In this case, multiple tissues will be created when you upload your data.
3) Add biosampleAccession numbers (e.g. SAMDXXXXXXXX) that correspond to each of your tissueIDs (which each correspond to a materialSampleID).
4) Upload to GEOME.

How do I find SRA-specific information such as library strategy, sequencing platform or read type?

Using the https://github.com/ropensci/rentrez package to link to the SRA metadata along with the GEOME metadata.

How do I refer to and cite sample metadata and related genetic data that I retrieved through GEOME?

Individual materialSamples, tissues, events can be cited using their ARK Identifier with the name-to-thing prefix. For example, https://n2t.net/ark:/21547/CVJ2BMOO_00004 This format is comprised of a resolution service that has a persistence mission (n2t.net) along with an open identifier scheme (ark), a GEOME expedition identifier (CVJ2), and a locally unique suffix (BMOO_00004). Genetic data that has been accessioned into a sequence repository should use the relevant identifier, such as https://www.ncbi.nlm.nih.gov/biosample/SAMN10240383

What if my genetic data is already on NCBI/GenBank?

If your sequence data is already in NCBI/GenBank in the Nucleotide database you can include the accession number/s in the 'associatedSequences' column of the metadata template for your Sample. For next-generation data already submitted to the Sequence Read Archive (SRA), insert the biosampleAccession number into the biosampleAccession field for your Tissue as 'SAMNxxxxxxxxx'. You may also put a link to SRA (e.g. https://www.ncbi.nlm.nih.gov/sra/?term=SAMN09015327) in the 'associatedSequences' column. This will enable GEOME to point to SRA for this record but not from SRA back to GEOME. If you want SRA to point to GEOME data, you will need to start with loading your data into GEOME and then using the GEOME FASTQ data loading service to push data to the SRA.

Can I submit my sequences to GEOME instead of NCBI/GenBank?

GEOME is not a sequence repository, but is designed to work alongside NCBI/GenBank. NCBI/GenBank has a series of quality control steps that are important for ensuring the quality of genetic data. You can upload FASTA files to GEOME, but this should only be in addition to submitting these to a dedicated sequence repository. GEOME does not accept FASTQ files, however users are encouraged to use the FASTQ tool in GEOME for creating an SRA (Sequence Read Archive) submission package for NCBI/GenBank. Using the FASTQ tool in GEOME will ensure that all of the metadata submitted to SRA has gone through our checks and that the proper links between GEOME metadata and SRA are created.

Do I upload my SNP genotypes, or the raw reads I used to get the SNP data?

We encourage you to use the FASTQ tool in GEOME for creating an SRA (Sequence Read Archive) submission package for NCBI/GenBank. GEOME accepts FASTQ submissions because we want to archive unmanipulated sequence data that are free of filtering biases. By doing this we are reducing ascertainment bias (a site that is not a SNP in your dataset may be a SNP when combined with data from a different population). We are also avoiding the subjective choices that go into calling SNPs, such as thresholds for trimming, filtering, coverage and likelihood. Our objective is for future users of your data to be able to make these choices in the context of their own question and dataset. If you would like to deposit your SNP (or microsatellite, alignments, or Amplicon Sequence Variant) data in a permanent open access repository (e.g. Dryad) and link this to sample metadata held in GEOME, you can use the 'derivedGeneticData' fields in GEOME.

What if my genetic data is for microsatellites?

The genetic databasing infrastructure for microsatellite data is not uniform and standardised, but this data is valuable! For any genetic data derived from samples represented in GEOME, that does not have a dedicated repository (e.g. microsatellite genotype files, SNP genotype files, ASV tables), we recommend you deposit the datasets in a permanent open access repository (e.g. Dryad). All of the sample metadata can be uploaded to GEOME and linked to the derived genetic dataset/s using the 'derivedGeneticData' metadata fields.

My genetic data is for a community or environmental sample (e.g. metagenomics, metabarcoding, eDNA) - how can I upload this to GEOME?

GEOME can support metadata for community and environmental DNA samples. In this case, the materialSampleID refers to the physical community or environmental sample, rather than an individual organism. Key fields to use for denoting eDNA samples:

- "basisOfRecord" (property of Sample entity) set to "EnvironmentalDNA"
- Use "environmentalMedium" (property of Event entity) to indicate the type of environmental sample (soil, air, seawater)
- "samplingProtocol" (property of Event entity) field indicates your sampling strategies (e.g. filter size, or other protocols).
- If you have a diversity of organisms in your sample, indicate the highest common-level taxonomic group that you are studying. You may wish to indicate Phylum as "Unknown".

As with all sample types, we encourage you to use the FASTQ tool in GEOME for creating an SRA (Sequence Read Archive) submission package for NCBI/GenBank (i.e. do this before submitting the genetic data to NCBI/GenBank). For any datasets derived from the community or environmental samples represented in GEOME (e.g. ASV or OTU tables), we recommend you deposit the datasets in a permanent open access repository (e.g. Dryad). These derived genetic dataset/s can be linked to the sample metadata in GEOME using the 'derivedGeneticData' fields.

My research is focused on host-symbiont (or host-microbiome, host-parasite, foundation species-community) systems - how can I link the metadata for both biological entities?

There are two options for handling this: 1) creating a single materialSampleID if they are comprised in the same physical sample, or 2) creating two different materialSampleIDs and then linking them. If using the same materialSampleID, you will need to differentiate with unique tissue IDs. Much of their ecological metadata will be the same, but they can each have their own links to their associated raw sequence data or derived genetic data. The 'associatedOrganisms' or 'environmentalMedium' field could additionally be used to name the partnering biological entity. The drawback to this approach is that the taxonomic assignment is made in the materialSampleID so you would only be able to make a single taxonomic assignment. The second approach, creating two materialSampleID's for each organism will let you make two taxonomic assignments and essentially viewing each entity as a separate. In this case, you would fill out the 'associatedOrganisms' field and point to the associated materialSampleID.

What if I have not published the genetic data yet (i.e. it is not on NCBI/GenBank)?

Your sample metadata can be uploaded to GEOME at any time in your project workflow, and it need not include the genetic data, or accession numbers for published genetic data. For next-generation sequencing data, we encourage you to use the FASTQ tool in GEOME for creating an SRA (Sequence Read Archive) submission package for NCBI/GenBank (i.e. do this before submitting the genetic data to NCBI/GenBank). When uploading your data to GEOME you have the option to keep your data private for a period of time. We ask that authors limit private data on GEOME for a period of 2 years and then make it public or remove it. If you make your data private, you can still make your project metadata "discoverable" using the discoverable switch under "Project Settings".

What if I have a tissue sample I am willing to share, but I have no genetic data attached to that sample (yet)?

Fill out the metadata template as completely as possible, including the 'Tissue details' fields. To indicate that you have a sample/tissue available write 'available' in the 'tissueRemarks' field.

What if I do not have any sample/tissue corresponding to that genetic data anymore?

We would still like to know the details of the sample and the sampling event that correspond to your genetic data. Fill out the metadata template as completely as possible, and pay extra attention to the 'tissueRemarks' field, you should write 'unavailable'.

How do I represent pooled RADSeq data?

Insert the count of the number of pooled individuals in the individualCount field. Any materialSampleID with individualCount > 1 will be considered to have been pooled. If individuals were collected across a spatial range, include this as coordinateUncertaintyInMeters.