# GEOME Documentation

This guide is designed to help you get started with GEOME.

- **Part 1** covers the basics and provides step-by-step instructions for first-time users.
- **Part 2** gives a deeper look into how GEOME is organized and structured.
- **Part 3** presents lower-level technical documentation for advanced users and developers.
- **Part 4** includes a collection of frequently asked questions to help address common concerns.

If you still have questions after reviewing the guide, feel free to reach out at geome.help@gmail.com.

## CONTENTS

© GEOME - 2025

© GEOME - 2025

© GEOME - 2025

© GEOME - 2025

# Part 1: Getting Started With GEOME

## Introduction

GEOME is a web-based database designed to capture and manage metadata for biological samples. It is primarily used in biodiversity inventories, population studies, and environmental metagenomics. GEOME assigns persistent identifiers to all samples and sampling events, and defines a standardized set of metadata attributes to record the *who*, *what*, *where*, and *when* associated with each sample.

The platform provides real-time feedback on data quality during uploads, supports integration with Laboratory Information Management Systems (LIMS), and connects directly with the Biocode LIMS plugin. Additionally, GEOME streamlines data submission to public repositories such as the Sequence Read Archive (SRA) and GenBank's Nucleotide database.

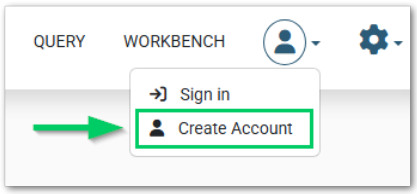
## Quick Start

To get started quickly with GEOME, just follow these two simple steps:

---

**Create an account**

1. Go to [https://geome-db.org/](https://geome-db.org/)
2. Click the *Create Account* button.



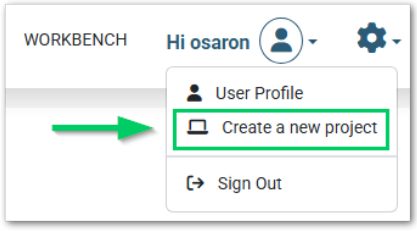

3. A create account window will appear, and you'll be prompted to enter information such as your name, email address, and password.

---

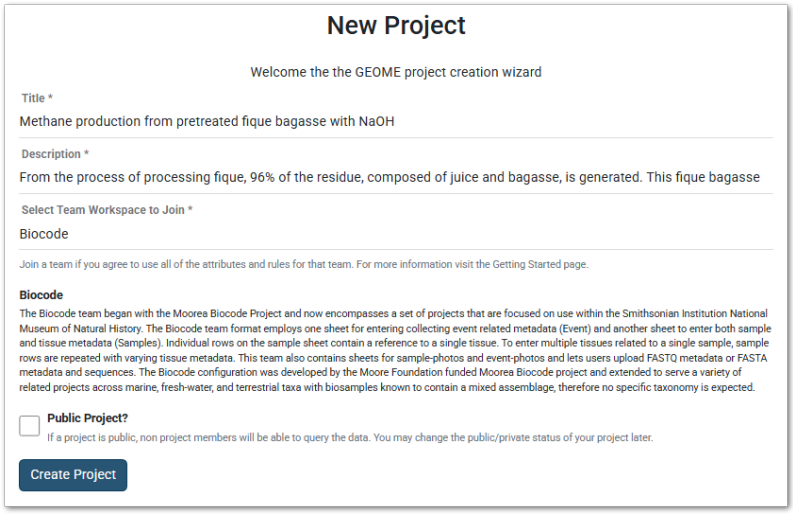
**Create a project**

1. Once you're logged in, click *Create a new project* in the top-right corner of the page.

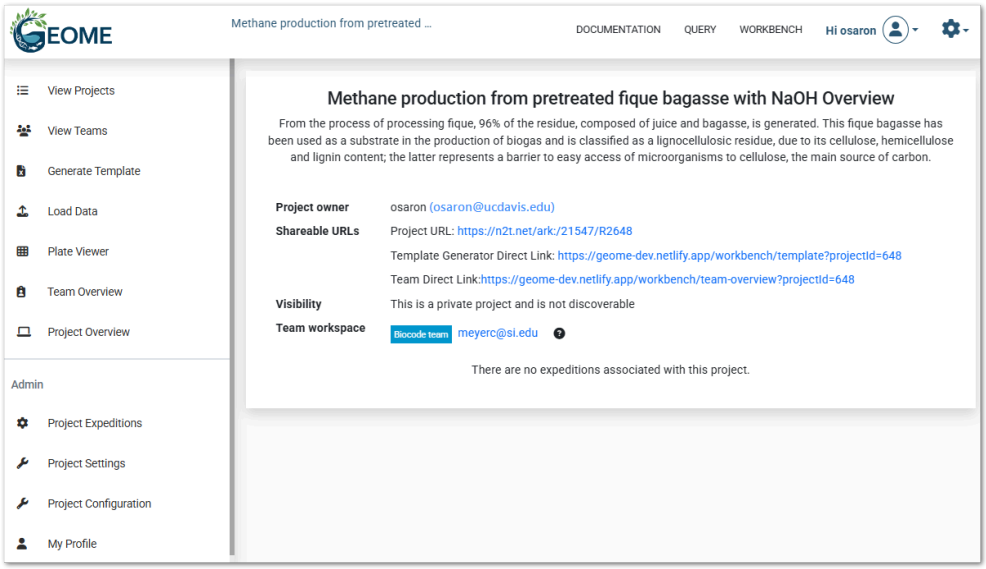
© GEOME - 2025




2. Once the New Project window pops up, fill in the title and description fields. Then, select a **Team Workspace** to join and set the project's visibility as either public or private. Select the *Create Project* button.




3. The GEOME system will create the project and redirect you to the project's overview page.




© GEOME - 2025

## About Teams, Projects, and Expeditions

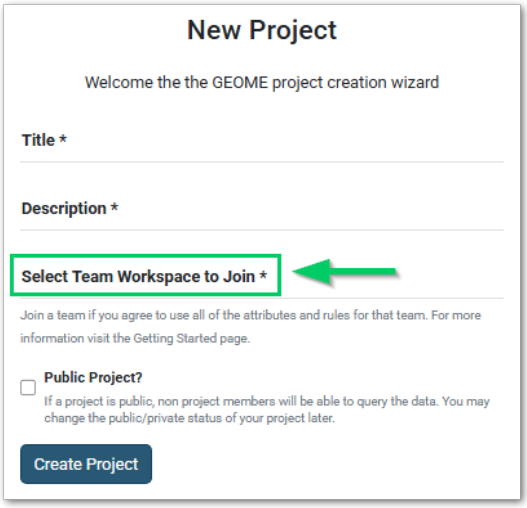
**Teams:**

A **Team** represents a specialized research group with shared settings applicable to all members. Teams enable users to create projects that follow a common set of rules, metadata attributes, and controlled vocabulary terms. When you create a project within a team workspace, you agree to use **ALL** of the team's defined attributes, rules, and controlled vocabularies. The team administrator controls all configuration options.

You can explore existing publicly available teams under the *View Teams* menu in the left-hand navigation. Each team page includes:

- A brief **abstract** describing the team's focus
- The **team owner's contact** information

If you have questions about joining a specific team, you are encouraged to email the team owner directly. In most cases, users are invited to join a team and are already familiar with its rules and structure. However, it is also possible to create a project under a team workspace **without an invitation**, as long as the team is configured to allow this. To create a project within a team, select *"Join team workspace"* during the project creation process, then choose the appropriate team.




**Projects:**

Each **Project** has a single owner who can invite additional members. All project members can create *Expeditions* within the project. Projects serve as containers for related expeditions and define the collaborative space for data collection and management.

**Expeditions:**

Projects are composed of one or more **Expeditions**, each corresponding to a single spreadsheet containing related collecting events, samples, and tissues. All data in GEOME must be entered as part of an expedition. Any project member can create an expedition by uploading a spreadsheet. The expedition creator becomes the expedition owner, with the ability to update or modify expedition data and set its visibility to public or private. The project owner also has the ability to edit expedition metadata across the project.

Expedition identifiers can be set as unique within the expedition or across the entire project. Each expedition is assigned a globally unique and resolvable prefix, known as the expedition root identifier. When a local identifier (enforced as unique within the expedition or project) is appended to this root, it forms a globally unique and resolvable identifier for each collecting event, sample, or tissue. These identifiers are automatically generated by the system and are referred to as BCIDs (Biocode Commons Identifiers).

---

## Main Page

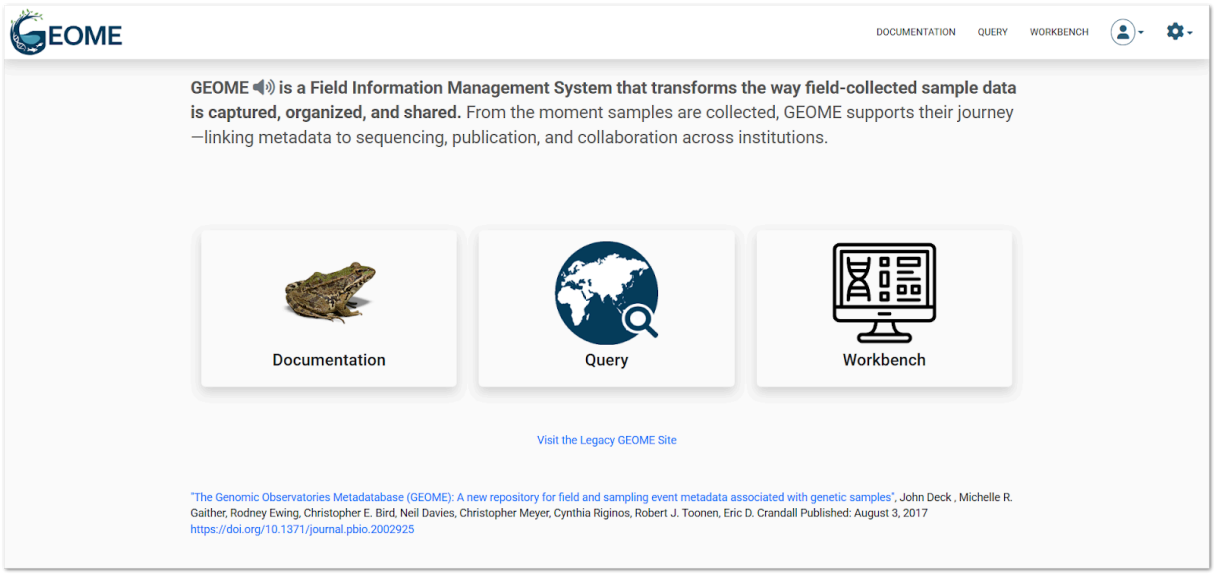
On the GEOME main page, users are presented with three large, clickable tiles that direct them to GEOME's core tools:

- **Documentation:** The documentation opens the downloadable user guide and technical documentation to support onboarding and detailed use.

- **Query:** The GEOME Query page allows users to explore and retrieve metadata records from expeditions hosted on the platform. It provides an intuitive map-based interface for filtering and visualizing samples based on geographic, project, and metadata-specific criteria.

- **Workbench:** The Workbench is your gateway to browsing and managing public and private projects within GEOME. It provides an overview of existing research projects, their associated sample data, and metadata status.

© GEOME - 2025




At the bottom of the page:

- A link to the original GEOME publication is provided for proper citation.
- A link to the legacy GEOME site is included for users accessing older datasets.

### Query

The Query page allows users to explore expedition data by applying entity-level filters, viewing results on a map, and exporting selected datasets. The GEOME **Query interface** filters data based on:

- Geographic information
- Keywords from metadata (e.g., *"Moorea"*)
- Darwin Core terms
- Expedition names
- Any other column defined in the GEOME specification

The results can be viewed in either **Map** or **Table** format by clicking the corresponding buttons in the top-right corner of the interface. To export your results, use the *Download* icon to retrieve the metadata from your query. Following, you will find a table with all the filters you can use:

| Filter | Options |
|---|---|
| Query Entity | `Event`, `Sample`, `Tissue`, `Fastq`, `Sample_Photo`, `Event_Photo`, `Diagnostics` |

| | |
|---|---|
| Team | Dropdown - it filters results by team. |
| Individual Projects | Dropdown - Filter results by a specific project under the selected team |
| Any term | keyword search across all metadata fields |
| **Buttom** | **Purpose** |
| + Add Event Filter | Filter by `eventID`, `country`, `habitat type`, etc. |
| + Add Sample Filter | Filter by `materialSampleID`, `sex`, `family`, etc. |
| + Add Tissue Filter | Filter by `tissueID`, `tissueType`, `tissuePlate`, etc. |
| + Add Sample_Photo Filter | Filter by `photoID`, `originalUrl`, `photographer`, etc. |
| + Add Event_Photo Filter | Filter by `photoID`, `originalUrl`, `photographer`, etc. |
| **Filter** | **Description** |
| Draw Bounding Box | Spatial filter to limit the query to a selected map area.<br>You can also filter by:<br>• is mappable<br>• has coordinateUncertaintyInMeters<br>• has permitInfo<br>• has tissue<br>• has sample photos<br>• has event photos |
| Sequences | `has a FASTA sequence`, `has NCBI Sequence Read Archive Accession Numbers` |
| Select Marker | Dropdown - select a genetic marker (e.g., COI, 16S, etc.) |

> **Tip:** You can check some Query examples here

### Workbench

The Workbench is your gateway to browsing and managing public and private projects within GEOME. It overviews existing research projects, their associated sample data, and metadata status.

**Public Projects List:** This section displays a searchable table of publicly available projects in GEOME. Each entry shows:

- Project Title (clickable link)
- Last Activity Date
- Sample Count

Projects may also show a team label indicating the team they belong to (e.g., AmphibiaWeb, FuTRES).

**Sidebar Tools:** On the left-hand menu, you can:

- View all your projects and teams
- Generate metadata templates
- Validate and load data
- Navigate to your Project Overview page

Use this page to choose a project for data upload, exploration, or metadata management.

---

## Generate Template

Sample metadata is recorded on an Excel Spreadsheet, and you can create and customize your own templates under **Workbench** > **Generate Template**.

**Tip:** Make sure you have selected a project

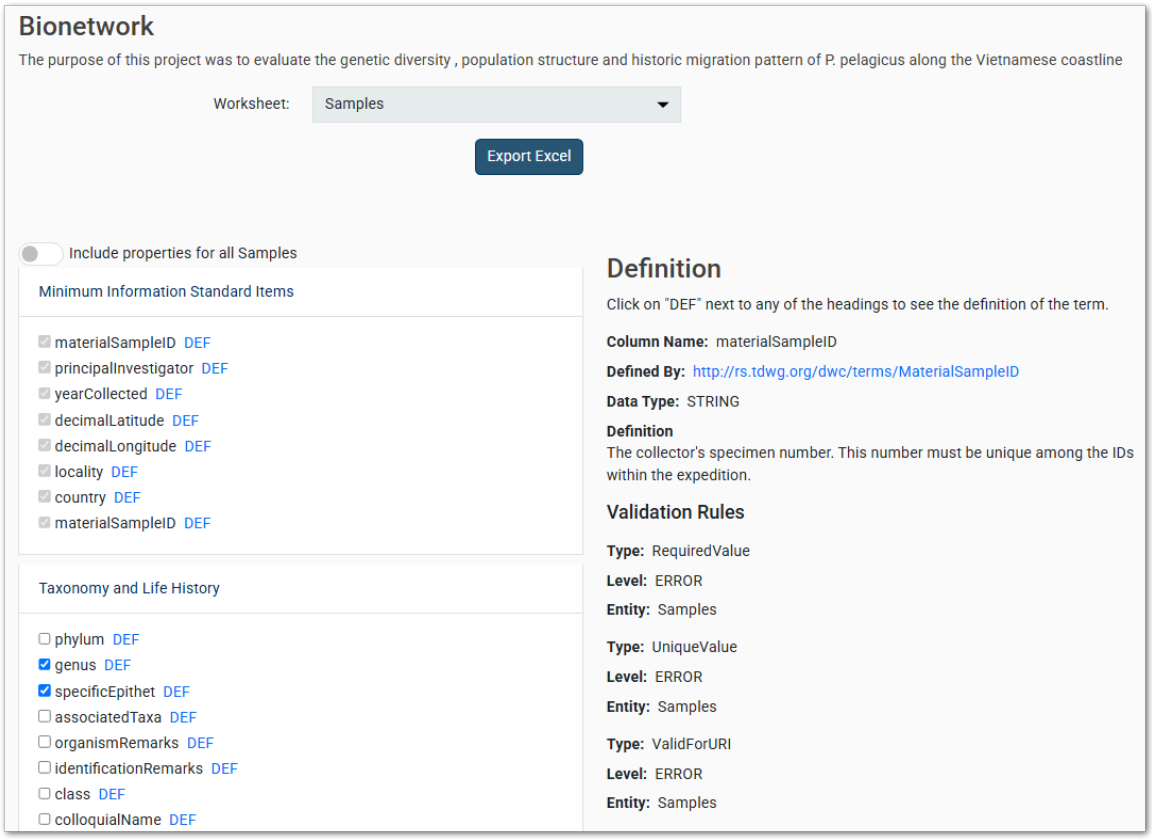

The purpose of the Generate Template page is to create a spreadsheet that can be used for data ingestion and validation via the **Load Data** section. The generated template includes all mandatory fields required by your project configuration, along with any optional fields you choose to include. It is formatted to ensure compatibility with GEOME's validation system.

© GEOME - 2025

On the **Generate Template** page, you can choose which columns to include in your spreadsheet. Click the DEF link next to each column name to view its definition.

- Mandatory columns are pre-checked and shown in **grey**—these cannot be unchecked.
- Suggested columns are pre-checked and shown in **blue**—these can be unchecked if not needed.

After selecting the columns you want to include, click the *Export Excel* button to download a spreadsheet. You can then enter your sample metadata directly into this file, which will be ready for upload through the **Load Data** interface. In addition, some data validation rules (e.g., field formats or controlled vocabularies) are embedded directly into the Excel template to help you enter valid data before uploading.

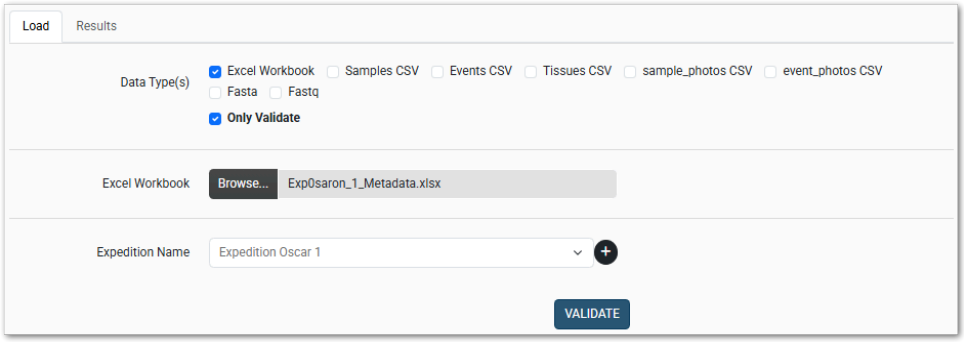



## Validate and Load Data

**Important Tip:** GEOME is a case-sensitive application. All sample, tissue, and event identifiers are case sensitive as well as URLs that are used to reference landing pages for these identifiers. Whenever working with spreadsheets, FASTA, or FASTQ files pay attention to case.

The Validate and Load Data option can be found under **Workbench** > **Load Data**. The first step is data validation, which checks that the metadata you entered conforms to the specific rules of your team and project. It provides immediate feedback on any issues that need to be corrected before your data can be accepted. To do this,

1. Select the data type of your file (excel, samples csv, FASTA data csv or FASTQ data csv).
2. Check the option **Only validate**.




3. Browse for the file in your system.
4. Select an existing expedition from the drop-down list or create a new one.
5. Click on **Validate** button. The website will automatically take you to the Results tab, where you can review any warnings or errors.

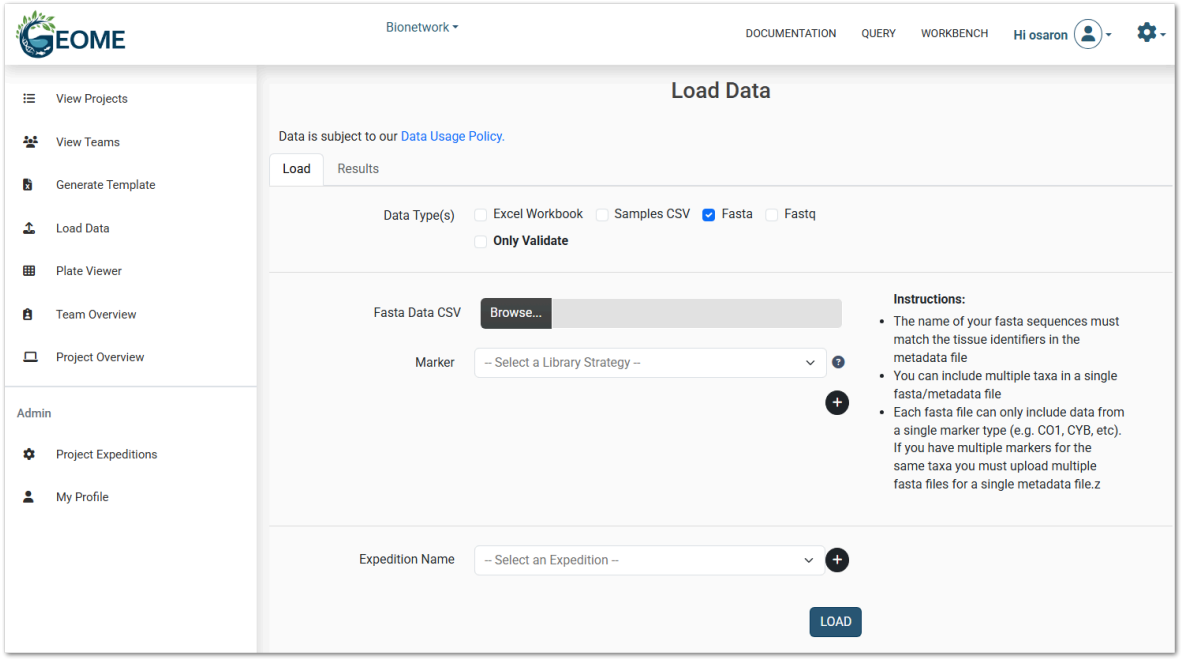
Use the Browse button to browse for your file and select the "Validate" button. After data validation, you can Upload your dataset and include just the metadata or include FASTA or FASTQ metadata.

### FASTA Upload Example

GEOME lets you upload FASTA files containing barcode data elements for integration with your metadata. Before uploading, you must either create a new expedition or select an existing one to associate your data with. Next, select your FIMS metadata file, the corresponding FASTA file, and a marker name (e.g., CO1, CYB). You'll also be required to confirm that you have visually verified sample locations using the map at the bottom of the page.

> ⚠ **Note:** The sequence identifiers in your FASTA file must exactly match the tissue identifiers in your metadata file.

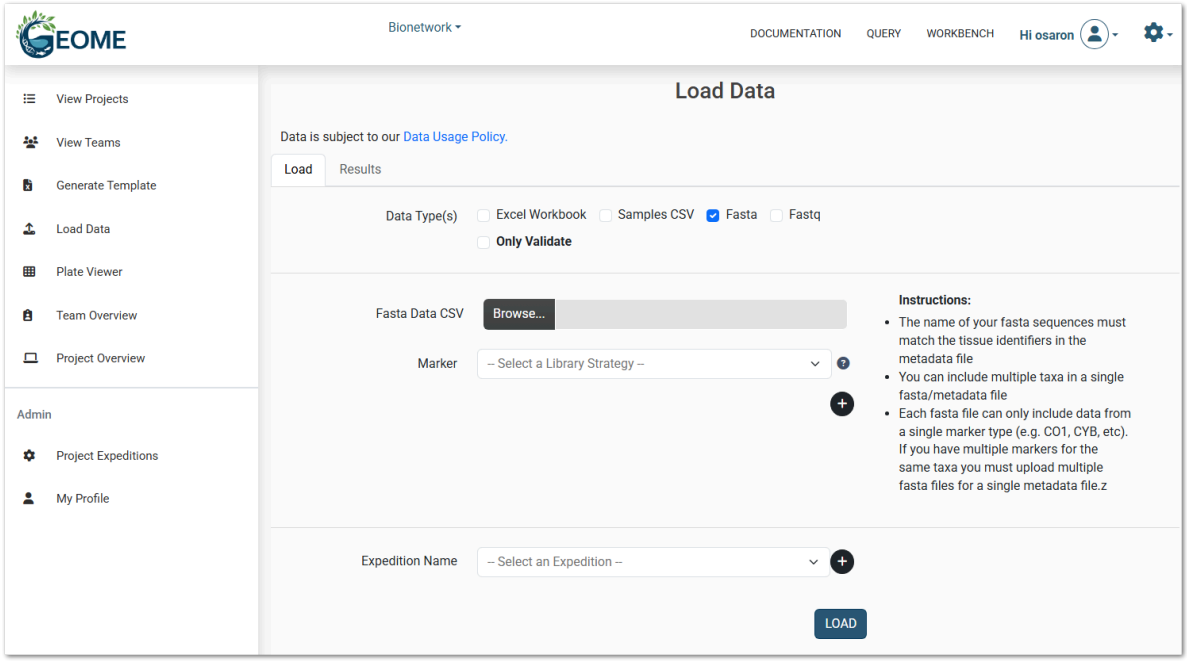
Each FASTA file should include sequences for only one marker type. If you have sequences from multiple markers for the same samples, upload separate FASTA files for each marker using the "+" button next to the marker selection field.




## FASTQ Uploading and Preparation for Submission to NCBI's Sequence Read Archive (SRA)

For teams that support the FASTQ data, GEOME allows you to bundle samples and tissue metadata with local FASTQ filenames to prepare a submission package for the NCBI Sequence Read Archive (SRA). Using GEOME for processing your metadata ensures that your submission to SRA aligns with community standards and maintains compatibility with other projects in your GEOME Team. GEOME's data validation processes help catch common metadata errors early in the workflow, avoiding problems during NCBI submission.

© GEOME - 2025




When uploading your data:

1. Check the "**FASTA Metadata**" box during file upload
2. You will be prompted to upload a **FASTQ Metadata file**, which is a plain text file listing all FASTQ filenames located on your local file system.

This file tells GEOME how to match each FASTQ read file with the appropriate tissue metadata.

### FASTQ Metadata File Requirements

To ensure successful validation, follow these rules when preparing your FASTQ metadata file:

- Each line should contain **only one (1) FASTQ filename**.
- Each filename should contain data from a **single individual**.
- Each filename **must begin with the tissueID** listed in your `Tissues` metadata sheet.

### FASTQ Filename Examples

For a tissue ID of sample1:

- **Single-end** filename
  - sample1.fastq.gz
  - sample1.fq.gz

- **Paired-end** filenames must have either a (-1 or -2) or (-F or -R).
  - sample1-1.fq.gz
  - sample1-2.fq.gz

© GEOME - 2025

- sample1-F.fastq.gz
- sample1-R.fastq.gz

⚠ **Note:** For the DIPNet team, `materialSampleID` and `tissueID` are synonymous. Other teams may assign multiple tissueIDs to a single sample, GEOME always matches by `tissueID`.

### After the Upload Process

Once validated, GEOME generates two submission files:

- BioSample attributes file
- SRA metadata file

These files are available for download and come with step-by-step instructions to help you finalize your submission on NCBI's SRA portal.

⚠ **Remember**

The FASTQ sequence files themselves **are not uploaded to GEOME**, only the metadata is stored. Also, if you receive an error related to filename encoding, your filenames may need to be converted to ASCII. See this issue for help: GEOME Issue #41.

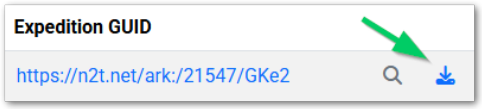
## Fastq Short Read Archive Upload

This section reviews the submission process into the Short Read Archive after you have uploaded your data into GEOME and successfully downloaded the bundle that you can use to submit to the SRA.

### Step 1: Download the Submission Bundle from GEOME (if you have not already done so)

After successfully uploading and validating your data in GEOME:

1. Go to your Project > Project Overview.
2. Select the Download arrow, on the right side of the expedition that contains the FASTQ metadata.




3. Select the **FASTQ SRA Metadata** bundle
4. Download and extract (unzip) the compressed file on your computer.

This bundle includes two essential `.tsv` files:

- "`bioSample-attributes.tsv`"
- "`sra-metadata.tsv`"

> You'll upload both to NCBI in the next steps.

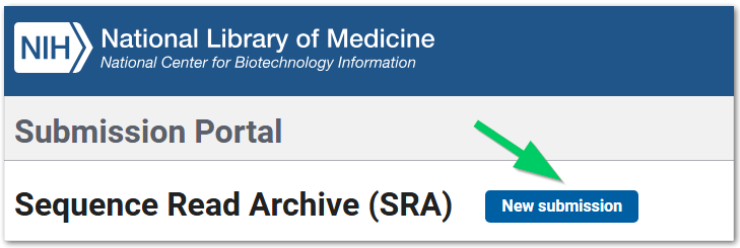
### Step 2: Upload Your Data to NCBI SRA

First of all, you need to set up you NCBI Account

- If you don't have one, create an NCBI account
- If you do, log in at the top-right of NCBI

### Step 3: Start the Submission

1. Go to: https://submit.ncbi.nlm.nih.gov/subs/sra/
2. Click **New Submission**




3. For detailed instructions, see: https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/

### Step 4: Complete the Submission Steps

| Submission Step | Description |
|---|---|
| **1.** Submitter Info | Confirm your contact details |
| **2.** General Info | • Have a **BioProject**? → Most will answer **No**<br>• Have **BioSamples**? → Most will answer **No**<br><br>Answering "**No**" lets SRA create these records for you: ncbi.nlm.nih.govbioinformaticsworkbook.orgncbi.nlm.nih.gov+8ncbi.nlm.nih.gov+8biostars.org+8 |

| | |
|---|---|
| **3.** Project Info | • **Project Title:** a clear short name (e.g., DIPNet Reef Fish Sequencing)<br>• **Description:** e.g., "RADSeq data for reef fish collections"<br>• **Relevance/Grants:** select as applicable. |
| **4.** BioSample Type | Select an appropriate package:<br>• Invertebrates<br>• model organism or animal sample<br>biostars.org+11ncbi.nlm.nih.gov+11docs.hpc.oregonstate.edu+11biostars.org+1bioinformaticsworkbook.org+1ncbi.nlm.nih.gov+1ncbi.nlm.nih.gov+1 |
| **5.** BioSample Attributes | Choose the **Upload a tab-delimited file** option. Then upload your `bioSample-attributes.tsv` (Downloaded from GEOME) |
| **6.** SRA Metadata | Choose the **Upload a tab-delimited file** option. Then upload your `sra-metadata.tsv` (Downloaded from GEOME). |
| **7.** Upload Data Files | ⚠ You'll be prompted to install **Aspera Connect** to make a fast transfer<br>• Choose **Fastq** or **Directory** upload, and locate your FASTQ data folder<br>• Use Aspera or a web browser to upload .fastq, .fq.gz, .tar.gz, or .zip files |
| **8.** Review & Submit | Verify all info under the **Overview** tab. Then submit it to receive a Submission ID (e.g., `SUB123456`). |

⚠ **Note:** You don't need to manually create BioSamples if GEOME has already generated the BioSample attributes file; NCBI will create them during the upload process.

### Step 5: After Submission

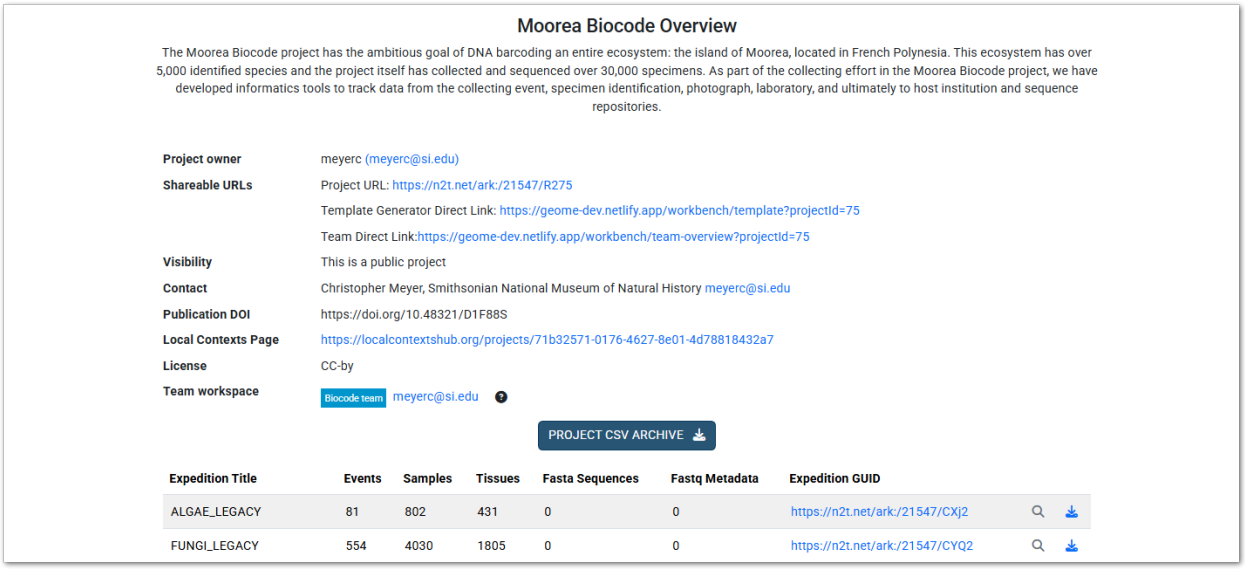
- Track progress under **My Submissions** or via your personalized link, which would look similar to this (**replace SUB#**): https://submit.ncbi.nlm.nih.gov/subs/sra/**SUB#**/overview
- SRA staff will review and process your submission. You'll receive **accession numbers** (e.g., SRR, SRX, SAMN) and completion status or error updates via email.

### Troubleshooting Tips

| Issue | Resolution |
|---|---|
| BioSample duplication | Use **one BioSample** per physical sample. Multiple FASTQ may map to different Experiments within the same BioSample: https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/ |
| File errors | Make sure FASTQ filenames match your GEOME `tissueID` values exactly |
| The file size is too large | Keep files **under 100 GB**. For > 5 TB files, split uploads across batches: https://www.ncbi.nlm.nih.gov/sra/docs/submitportal/ |
| Header mismatch | Ensure `.tsv` headers exactly match the required SRA formats |
| Unique Sample Names | Unique sample names are required, no duplicates, biostars.org |

---

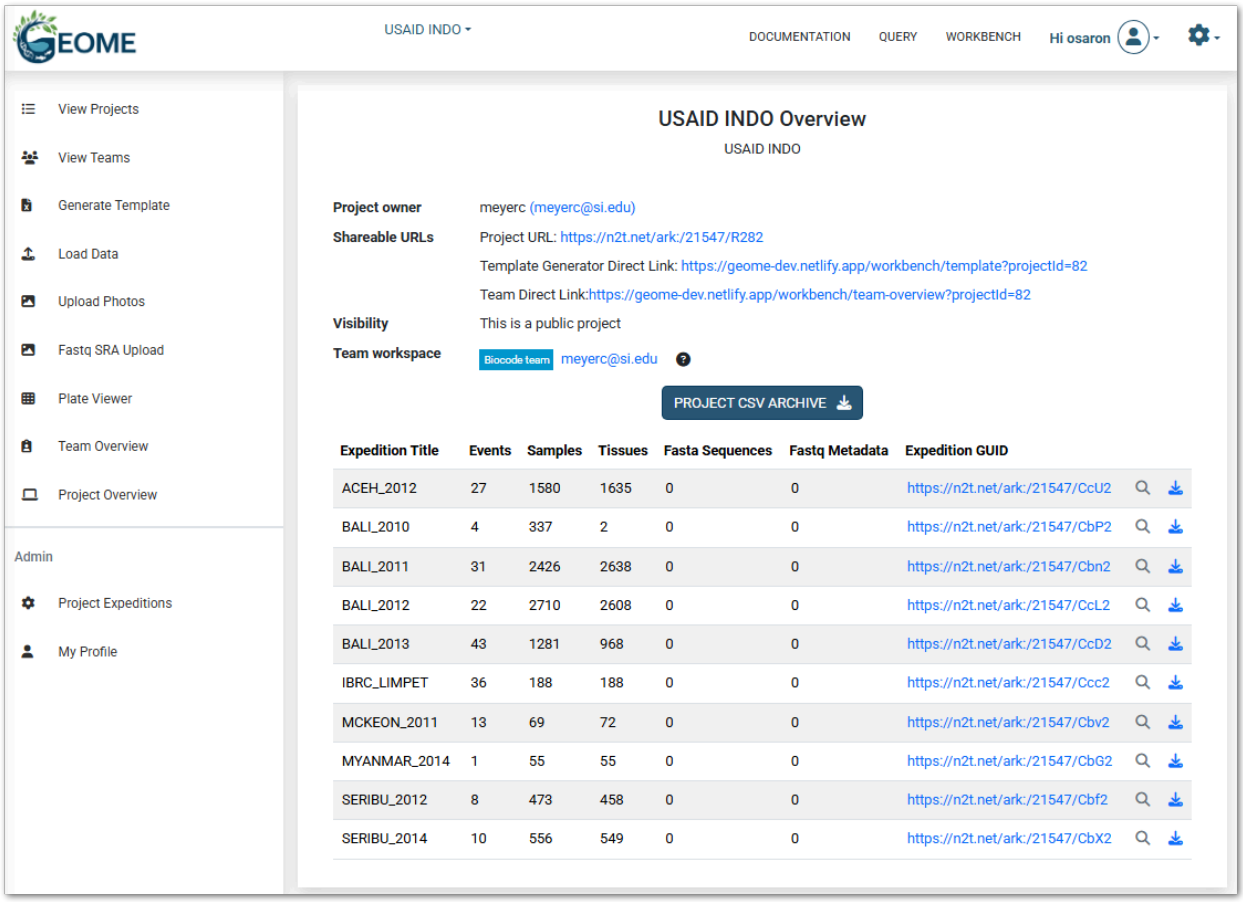
## Project Overview




## 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 EXCEL, CSV, FASTA, or FASTQ formatted metadata.

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.




## Upload Photos

GEOME supports two methods for attaching photos to your records. Photos can be linked to either samples or events, depending on the selected Photo Entity (Sample Photo or Event Photo).

1. **Reference Online Images (via CSV)**
   If your images are already hosted online, you can upload a CSV file containing the URLs and associated metadata. To update or remove existing photo metadata, you must reload the expedition it belongs to.

- Go to **Workbench** > **Load Data**, then select `sample_photos_csv` or `event_photos_csv`.
- Include the following fields in your CSV:
  - `photoID` (optional but recommended for future updates)
  - `originalUrl` (required – link to the image online)
  - `expeditionCode` (required if referencing multiple expeditions)
  - Other optional metadata fields (see Generate Template > Photo Entity)

© GEOME - 2025

2. **Upload Photos from Your Computer**
   You can upload a zipped directory (“**.zip**” **extension is required**) of image files, max **2GB per upload**. The JPEG format is most reliable. There are two ways to match photos with records:

- *Filename-based Upload*, name each file using:
  ```
  materialSampleID+photoIdentifier.ext
  ```
  No spaces, dashes, or special characters in the photoIdentifier.

- *Metadata-based Upload*, Include a `metadata.csv` file in the zip with these required fields:
  - `materialSampleID` or `eventID` (to link the photo to the correct record)
  - `fileName` (matching the image file name)
  - `expeditionCode` (required if spanning multiple expeditions)
  - Additional photo metadata fields as needed

To upload photos:

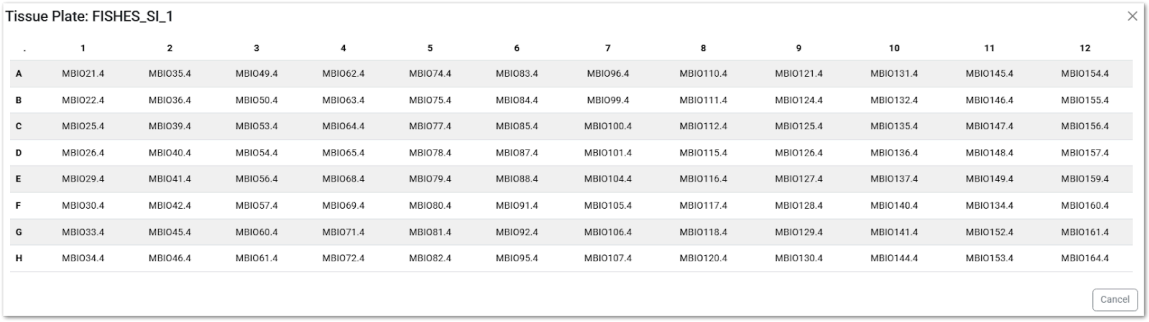
1. Select the appropriate Photo Entity ("*Sample Photo*" or "*Event Photo*")
2. (Optional) Check Auto-Generate IDs to let GEOME create photo identifiers automatically
3. Click Choose File and select your zipped image folder
4. Click Upload

> ⚠️ **Note:** Photo uploads may take several minutes to fully process, depending on file size and server load.

**Plate Viewer**

The plate viewer allows you to visualize the contents of a tissue plate as they are arranged in a 96-well format. This tool helps inspect sample distribution and verify that tissue identifiers are correctly organized. To do this, follow the next steps:

1. Select a **Tissue Plate** from the dropdown list
2. GEOME will automatically display the plate in a grid format (A-H by 1-12), showing the tissue IDs assigned to each well. Here’s a preview of a plate:




| . | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | MBIO21.4 | MBIO35.4 | MBIO49.4 | MBIO62.4 | MBIO74.4 | MBIO83.4 | MBIO96.4 | MBIO110.4 | MBIO121.4 | MBIO131.4 | MBIO145.4 | MBIO154.4 |
| B | MBIO22.4 | MBIO36.4 | MBIO50.4 | MBIO63.4 | MBIO75.4 | MBIO84.4 | MBIO99.4 | MBIO111.4 | MBIO124.4 | MBIO132.4 | MBIO146.4 | MBIO155.4 |
| C | MBIO25.4 | MBIO39.4 | MBIO53.4 | MBIO64.4 | MBIO77.4 | MBIO85.4 | MBIO100.4 | MBIO112.4 | MBIO125.4 | MBIO135.4 | MBIO147.4 | MBIO156.4 |
| D | MBIO26.4 | MBIO40.4 | MBIO54.4 | MBIO65.4 | MBIO78.4 | MBIO87.4 | MBIO101.4 | MBIO115.4 | MBIO126.4 | MBIO136.4 | MBIO148.4 | MBIO157.4 |
| E | MBIO29.4 | MBIO41.4 | MBIO56.4 | MBIO68.4 | MBIO79.4 | MBIO88.4 | MBIO104.4 | MBIO116.4 | MBIO127.4 | MBIO137.4 | MBIO149.4 | MBIO159.4 |
| F | MBIO30.4 | MBIO42.4 | MBIO57.4 | MBIO69.4 | MBIO80.4 | MBIO91.4 | MBIO105.4 | MBIO117.4 | MBIO128.4 | MBIO140.4 | MBIO134.4 | MBIO160.4 |
| G | MBIO33.4 | MBIO45.4 | MBIO60.4 | MBIO71.4 | MBIO81.4 | MBIO92.4 | MBIO106.4 | MBIO118.4 | MBIO129.4 | MBIO141.4 | MBIO152.4 | MBIO161.4 |
| H | MBIO34.4 | MBIO46.4 | MBIO61.4 | MBIO72.4 | MBIO82.4 | MBIO95.4 | MBIO107.4 | MBIO120.4 | MBIO130.4 | MBIO144.4 | MBIO153.4 | MBIO164.4 |

3. You can also click **New Plate** to create additional plates if needed

> **Note:** This visualization is particularly useful for validating plate layout before sequencing or downstream processing.

### Team Overview

The Team Overview module provides a summary of the team's purpose, structure, and associated projects. It outlines the metadata configuration and data entry format used by the team, as well as any special considerations for sample, tissue, and sequencing metadata. At the top of the page, you'll find a brief team description explaining its origin, metadata approach, and scope of supported taxa or projects. Also, you will find the administrator and Contact information for team-related inquiries.

Below this, a Projects table lists all projects created under the team, with the following columns:

- **Title** – Linked project name
- **Samples** – Total number of samples within the project
- **Project Owner** – User who owns or manages the project
- **Latest Activity** – Most recent update to the project

This module is helpful for exploring ongoing research efforts, finding relevant contacts, or reviewing the data structure followed by your team.

---

## Query Examples

### Event Query

1. In the *Query Entity* filter, select **Event**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `Events` will show up on the map. You will be able to visualize them by clicking on the different location icons.


### Sample Query

1. In the *Query Entity* filter, select **Sample**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `Samples` will show up on the map. You will be able to visualize them by clicking on the different location icons.

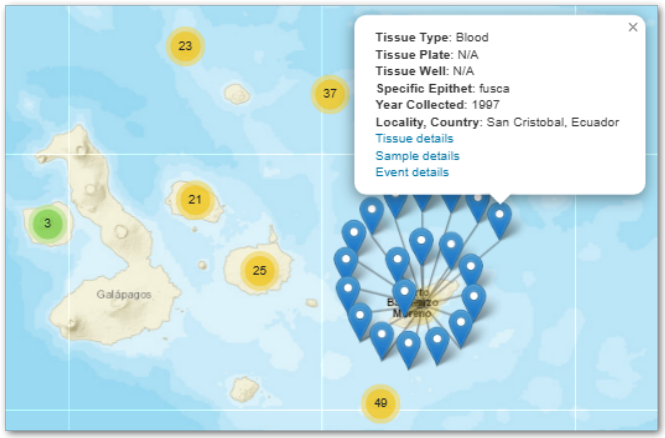
© GEOME - 2025


## Tissue Query

1. In the *Query Entity* filter, select **Tissue**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `Tissues` will show up on the map. You will be able to visualize them by clicking on the different location icons.

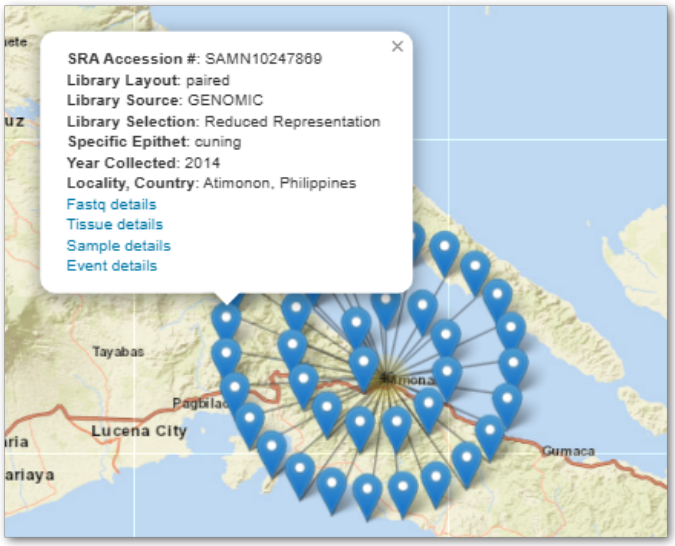


### Fastq Query

1. In the *Query Entity* filter, select **Fastq**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `Fastq` documents will show up on the map. You will be able to visualize them by clicking on the different location icons.

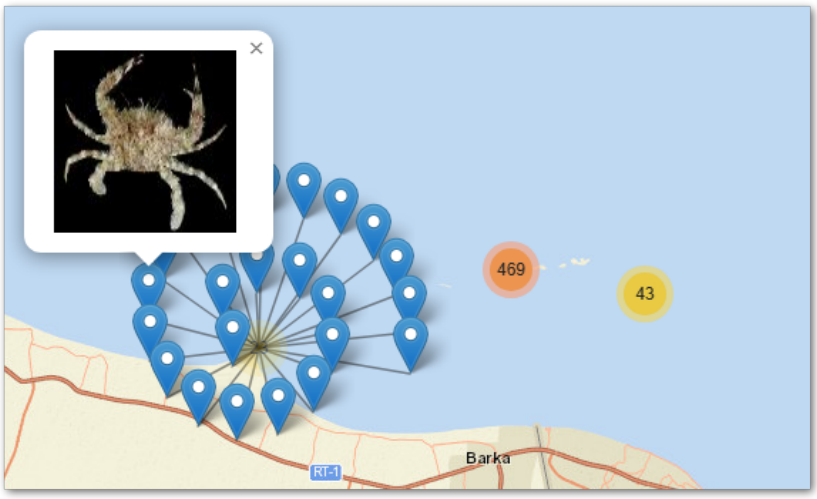



### Sample/Event Photo Query

1. In the *Query Entity* filter, select **Sample Photo** or **Event Photo**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `sample/event photos` will show up on the map. You will be able to pre-visualize them by clicking on the different location icons.




5. Click on the photo to check detailed information about it. The following image represents a biological specimen (Charybdis crab) associated with the sample. On the right-hand side, a metadata panel lists key details:

    - **Filename and photoID:** Unique identifiers for the image
    - **materialSampleID:** `BOMAN_3769`, linking the photo to its parent biological sample.
    - **originalUrl:** Direct link to the image source (e.g., a hosted CDN)

You will also see additional fields like hasScale, photographer, and qualityScore (shown as N/A when not supplied).

> **Tip:** Click on the `materialSampleID` link at the bottom to check more info about the sample.

## Sample_Photo (490c4fa3_25cc_4871_93a3_55f0b77207ae)

| | |
|---|---|
| filename : | 490c4fa3-25cc-4871-... |
| hasScale : | N/A |
| materialSampleID : | BOMAN_3769 |
| originalUrl : | https://cdn.floridamu... |
| photoID : | 490c4fa3_25cc_4871... |
| photographer : | N/A |
| processed : | true |
| qualityScore : | N/A |

| | | | |
|---|---|---|---|
| **img1024:** | 1024 pixel wide image | **expeditionCode:** | NSF_OMAN |
| **img128:** | 128 pixel wide image | **project:** | Meyer LAB |
| **img512:** | 512 pixel wide image | | |

### Parent Sample

**materialSampleID:** BOMAN_3769 **scientificName:** Charybdis **lowestTaxonRank:** N/A **catalogNumber:** N/A

## Sample (BOMAN_3769)

| | |
|---|---|
| bcid : | https://n2t.net/ark:/21 |
| genus : | N/A |
| materialSampleID : | BOMAN_3769 |
| specificEpithet : | N/A |


| | | | |
|---|---|---|---|
| **eventID:** | OMAN_029 | **expeditionCode:** | NSF_OMAN |
| **institutionCode:** | University of Florida | **phylum:** | Arthropoda |
| **sampleEnteredBy:** | Abby Uehling | **project:** | Meyer LAB |
| **scientificName:** | Charybdis | | |

### Parent Event

**eventID:** OMAN_029 **yearCollected:** 2020 **country:** Oman **decimalLatitude:** 23.7823569 **decimalLongitude:** 57.7947008

### Child Entities

**Sample_Photo** **photoID:** 490c4fa3_25cc_4871_93a3_55f0b77207ae **qualityScore:** N/A **hasScale:** N/A

**Tissue** **tissueID:** BOMAN_3769.1 **tissuePlate:** FLMNH_228 **tissueWell:** C11 **fromTissue:** N/A **tissueInstitution:** N/A **tissueType:** N/A

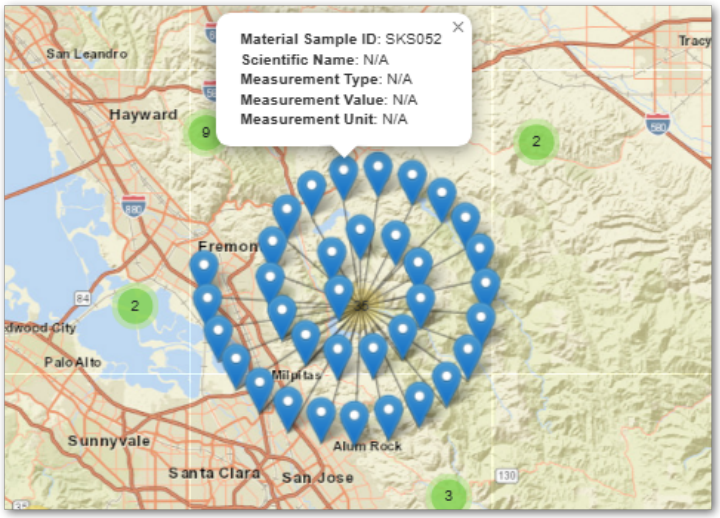
© GEOME - 2025

### Diagnostics Query

1. In the *Query Entity* filter, select **Diagnostics**
2. Click on the **search** button at the bottom of the page
3. On the map, click on the highlighted location you want to visualize.

> ⚠ **Note:** The website will automatically zoom if there's more than one location close to the place selected.

4. Once you reach the final location, all the `diagnostics` will show up on the map. You will be able to visualize them by clicking on the different location icons.




---

## 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.

---

## 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 the 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 next-gen 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 generated:

- `bioSample-attributes.tsv`

- `sra-metadata.tsv`

These files are required for uploading your data to the NCBI SRA and will significantly streamline the submission process!

**Tip:** If you don't already have an NCBI account, create one here.
If you already have an account, sign in using the Sign In tab at the top-right corner of the page.

**Step-by-Step Submission Process**

**Step 1:** Submitter
Enter your personal information as requested.

**Step 2:** General Info
You'll be asked two questions:

- *Did you already register a BioProject for this dataset?*
- *Did you already register BioSamples for this dataset?*

In most cases, the answer to both is **No**. The following steps assume you answered "No."

**Step 3:** Project Info
Fill in project information, for example:

- **Project Title:** `Acanthurus_reversus_RADSeq_data`
- **Description:** `RADSeq data for the reef fish Acanthurus reversus`
- **Relevance:** `Evolution`
- **Part of a larger initiative?** Most likely **No**
- **External Links / Grants:** Add if applicable

**Step 4:** Biosample Type
Select your sample type. Most users will choose:

- "Invertebrates"
- or "Model organism or animal sample" (for vertebrates)

**Step 5:** Biosample Attributes

Upload the `bioSample-attributes.tsv` file generated by GEOME.

- **Validator Warnings:**
  You may receive warning messages from the SRA validator.

    - **Errors** must be fixed.
    - Some **warnings** may be safely ignored (e.g., locality warnings for marine samples based on nearby terrestrial locations).

**Step 6:** SRA Metadata
Select **Upload a file**, then upload the `sra-metadata.tsv` file produced by GEOME.

**Step 7:** Files
Follow the instructions to upload your sequencing data:

- Download and install **Aspera Connect** (required for faster uploads).
- Click **Choose Files**, select your zipped data folder.
- Aspera will launch automatically to handle the upload.

**Step 8:** Overview & Submit
Review your submission details, then click **Submit** to finalize the process.

© GEOME - 2025

# Part 2: Understanding How GEOME Works

## Minimum information requirements

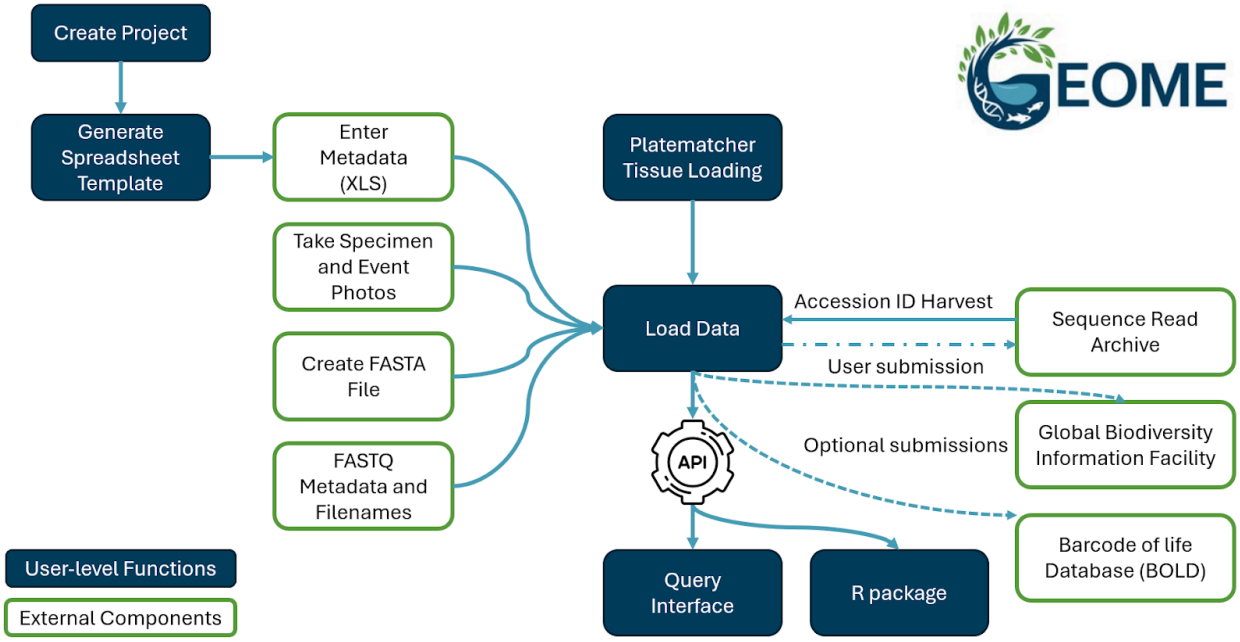
GEOME requires the following fields to be entered for ALL projects:

- `materialSampleID`
- `yearCollected`
- `country`
- `locality`

Each **Team** can define additional field requirements based on its specific needs. The following table outlines the minimum required information for both teams and individual 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: Managing information in the GEOME environment




## The GEOME R Package

The **GEOME R package** is used to retrieve data from GEOME for analysis. Since it is no longer available on CRAN, you can install it using the GitHub installer. Please visit our page GEOME R package for the latest installation instructions and code examples.

## The Biocode LIMS Plugin

All data uploaded to GEOME can be accessed and managed through a laboratory information management system (LIMS) using a custom-built LIMS plugin designed for the **Geneious environment**. The purpose of the LIMS tool is to help manage lab and sequence analysis workflows. To learn how to use the plugin, please refer to the LIMS plugin wiki.

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

© GEOME - 2025

## 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**, which ran from 2006 to 2011 and was funded by the Moore Foundation. The project aimed to create a comprehensive biotic inventory of a single tropical island, involving six teams, 50 researchers, and thousands of collecting events. To support data collection, the Moorea Biocode Field Information Management System (FIMS1) was developed. The tool (written in Perl and Java) allowed researchers to ingest spreadsheet data with built-in validation. FIMS1 remained in use from 2006 to 2018 and introduced several tools, including:

- **Plate Matcher** – for mapping tissues to 96-well plates
- **bioValidator** – for loading and validating spreadsheets
- **A web interface** – for managing data

From 2012 to 2015, the National Science Foundation funded the BiSciCol (Biological Sciences Collections) project. Initially, BiSciCol focused on persistent identifiers, ontologies, and semantic web tools to connect biodiversity data, where its main use case was linking events, samples, and tissues across different systems using linked data technology. It was achieved by integrating the Moorea Biocode data with member institution databases. This project led to several key developments:

- The BiSciCol triplifier
- The Biological Collections Ontology (BCO)
- A clearer understanding of persistent identifiers
- The launch of FIMS2, also known as the BiSciCol FIMS

FIMS2 introduced the use of ARK identifiers (via California Digital Library's EZID system) for events, samples, and tissues. It also utilized a Fuseki triplestore with metadata configurations defined through an XML file, stored and managed separately by each project.

Between 2014 and 2017, FIMS2 began supporting the Diversity of the IndoPacific Network (DIPNet), building a unified metadata repository for the IndoPacific region. During this time, the system gained the ability to upload **FASTA** and **FASTQ** files, and the name GEOME was introduced, eventually becoming the brand name for what would become **FIMS3**.

Around the same period, a separate FIMS system for the Smithsonian National Museum of Natural History (NMNH) was developed to create a centralized field data ingestion system. Although this version (NMNH FIMS) was later discontinued, Smithsonian users continued to work with FIMS1 and FIMS2 for various programs, including the Barcode of Wildlife Project and the Global Genome Initiative.

In 2016, John Deck and RJ Ewing began developing a redesigned system, FIMS3, now known as **GEOME**. This new platform was built with a PostgreSQL backend for managing project features and used JSON metadata objects to store data and configurations. It incorporated key capabilities from earlier systems:

- From FIMS1: photo uploads, specimen/tissue/event pages, and plate-matching tools
- From FIMS2: persistent ID generation and flexible configuration using non-relational metadata structures

FIMS3 also introduced the concept of networks—collections of projects governed by shared configuration templates—and a more user-friendly interface for project creation.

The goal of GEOME is to unify the features of FIMS1 and FIMS2 into a modern, scalable system. The original FIMS1 and FIMS2 installations were phased out by Spring 2019.

**GEOME** was hosted at the University of Florida until 2018, then at the University of Arizona (CyVerse) from 2019 to 2025, and has been hosted at Indiana University (Jetstream2) since 2025.

---

## 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 the 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 a 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 aspect of GEOME is converting local identifiers—those you create and use within your own research—into globally unique, resolvable identifiers. This is achieved by appending your local identifier to a unique root generated for each resource within every expedition. Examples of local identifiers include:

- “Grinnell1213”
- “MooreaEvent2”
- “MBIO56_1”

These identifiers are transformed into globally unique formats to ensure consistency, traceability, and interoperability across systems. Each identifier minted in GEOME is resolvable via HTTP using the California Digital Library’s Name-to-Thing (N2T) resolver. Because the resolver is sensitive to certain characters, GEOME enforces a restricted set of allowable characters for local identifiers. These character rules are automatically validated during data upload. If an invalid character is used, you will receive an error message. The following are the allowed local identifier characters:

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

© GEOME - 2025

| VALID local identifiers | INVALID local identifiers |
|---|---|
| MVZ:Herp:1234 | MVZ-Herp-1234 |
| Grinnell (1234) | Grinnell//Alexander 1234 |

Once your data is uploaded and made public, your globally unique and resolvable identifiers will appear in the Query page under the BCID column. The following example illustrates how to convert an identifier into a resolvable URL:

| Identifier | Prefix | Resolvable |
|---|---|---|
| ark:/21547/R2MBIO564 | http://n2t.net/ark: | https://n2t.net/ark:/21547/CXs2MBIO564 |

In the example, **MBIO564** is your locally unique identifier, which becomes part of the globally resolvable identifier through the GEOME system and the California Digital Library's Name-to-Thing (N2T) service.

## GEOME Queries

GEOME supports a custom SQL-like query syntax designed to help you efficiently search for the data you need. This syntax complements the Swagger Application Programming Interface (**API**) documentation.

### Basic Query Behavior

By default, query terms are evaluated across **all columns** within a project. To target a specific column, use the format `columnName:query`. This is called the full text search query:

- A query like `japan` will return results where any column contains the word "japan".
- A query like `column1:japan` will return results where only column1 contains the word "japan".

### Logical Operators and Grouping

You can construct more advanced queries using the following logical operators:

- `AND`
- `OR`
- `NOT`

You can also use parentheses `()` to group terms and control the order of operations.

| Query | Description |
|---|---|
| `_expeditions_:myExpedition AND NOT japan` | This query returns all records from the expedition named `myExpedition` that do not contain the word `japan` in any field. |

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 (FTS)

Full text search is the default query method. It searches across a tokenized version of the uploaded data. Here are some syntax examples:

- `col1:value` – searches for `value` within `col1`
- `col1:val*` – searches `col1` for words starting with `val` (prefix match)
- `value` – searches for `value` across **all columns**

### Comparison Queries

Comparison queries allow you to filter based on value relationships (**between 2 values**). The following operators are supported:

| Operator | Meaning |
| --- | --- |
| `=` | Equals |
| `<>` | Not equals |
| `>` | Greater than |
| `>=` | Greater than or equal to |
| `<` | Less than |
| `<=` | Less than or equal to |

⚠ **Note:** For accurate comparison results when using `<`, `<=`, `>`, or `>=`, the attribute's `dataType` should be one of the following:

- "`Integer`"
- "`Float`"
- "`Date`"
- "`Datetime`"
- "`Time`"

These data types can be configured in your project settings. If you're unsure how to do this, contact your project administrator.

### Project Query

Use this query to filter results based on the **project(s)** to which the records belong. Syntax Examples:

- `_projects_:1` returns all records uploaded under **Project 1**.
- `_projects_:[1, 2]` returns all records uploaded under **Project 1** or **Project 2**.

### Expedition Query

This query filters results by expedition name(s).

⚠ **Note:** Since expeditions are only unique within a project, it's recommended to combine this query with a project filter to avoid ambiguity.

Syntax Examples:

- `_expeditions_:myExpedition` returns all records under the expedition named **myExpedition**.
- `_expeditions_:[myExpedition1, myExpedition2]` returns all records under either **myExpedition1** or **myExpedition2**.

### _exist_ Query

This query returns results where a specific column **has a non-empty value**. Syntax Examples:

- `_exists_:column1` returns all records where **column1** contains a value.
- `_exists_:[column1, column2]` returns records where either **column1** or **column2** contains a value.

### like Query

Performs a SQL `ILIKE` query (case-insensitive match) using a pattern. Syntax Example:

- `col1:"%value"`
  Matches any value in **col1** that ends with `"value"`
  Equivalent to: `col1 ILIKE '%value'`

### phrase Query

Performs a SQL `ILIKE` query (case-insensitive match) using a pattern. Syntax Example:

- `col1:"some value"`
  Matches any value in **col1** that ends with `"some value"`
  Equivalent to: `col1 ILIKE '%some value%'`

### Range Query

A shorthand for a comparison query that defines a numeric, date, or time range.

> ⚠ **Note:** To ensure correct behavior, the `dataType` of the attribute must be set to one of the following:
>
> - "`Integer`"
> - "`Float`"
> - "`Date`"

- "`Datetime`"
- "`Time`"

This is set in your project configuration. Contact your project administrator if needed. Syntax Examples:

| Query | Meaning |
|---|---|
| `col1:[1 TO 10]` | `col1 >= 1 AND col1 <= 10` |
| `col1:[1 TO 10}` | `col1 >= 1 AND col1 < 10` |
| `col1:{1 TO 10}` | `col1 > 1 AND col1 < 10` |
| `col1:{* TO 100]` | `col1 <= 100` (no lower bound) |

### Select Query

The `_select_` query is used to include related parent or child entities in the query results. The value must be the `conceptAlias` of the related entity you wish to include. The selected entities must be related to the query entity, but do not need to be directly related. For example, if you're querying a parent entity, you can also select its grandchildren or grandparents.

⚠ **Note:** `_select_` queries:
- Should not be combined with `AND`, `OR`, or `NOT` keywords.
- Must stand alone in the query expression.

Examples:
- `_select_:parentEntity` returns results for the queried entity plus the related `parentEntity`.
- `_select_:[parentEntity, grandParentEntity]` returns results including the queried entity, its **parent**, and its **grandparent** entities.

### Tokenization

GEOME applies a tokenization process to all text fields before they are indexed. This process:

- Breaks down text into individual tokens (words and numbers)
- Converts words into their normalized (root) forms for improved search matching

Example:

- Input: `"many donkeys"`
- Tokens: `["many", "donkey"]`

This allows full-text search queries to match on the root form of words, improving the accuracy and flexibility of your searches. For more information, you can view the psql tokenization.

---

## Installation

This section is for developers or system administrators wishing to install GEOME on their own server.

### Overview

GEOME is composed of a core set of Java classes and RESTful services. Developers would be able to:

- Use the publicly hosted BCID service (with built-in EZID minting), or
- Host their own instance of BCID, which requires setting up an EZID account (purchase required)

### Required Components

To deploy an instance of GEOME (FIMS), you will need the following components:

- A Unix-based server
- A Java servlet container, such as:
    - Apache Tomcat
    - Glassfish
    - Jetty
- Connection to a BCID service (public or self-hosted)

### Installation & Build Steps

You can install GEOME and migrate an existing installation by following these steps:

1. Clone the Repositories, `bcid` and `geome-db` from GitHub.

2. Install Prerequisites.

- PostgreSQL
- Jetty 9
- Java 8

3. Copy `biocode-fims.template` to `biocode-fims.props` in the root directory. Update the environment-specific properties files in: `src/main/environment/production`

4. Use the provided Gradle build file to build and deploy: `gradle war deploy` `build/libs/geome-db.war` Repeat the process for both `bcid` and `geome-db`.

5. To generate OpenAPI documentation, run the following Gradle task: `gradle resolve`.

---

## Configuration files

GEOME uses JSON-based configuration files at both the network and project levels to define structure, behavior, and validation rules for data entry and querying.

### Network Configuration

The network-level configuration file acts as the foundation for consistency across projects within a network, and defines:

- Global network rules for all GEOME projects
- All available entities and data properties

### Project Configuration

Each project includes its own configuration file that complements the GEOME network configuration file, and defines:

- Custom resources (e.g., specimens, tissues, events)

- Attributes (columns/fields)
- Validation rules
- Relationships between entities

These project-level configurations allow for flexibility while maintaining compatibility with network-wide standards.

---

## Attributes

### Data Types

Each attribute in a project can be assigned a **dataType** to enable additional validation and standardized formatting. This improves data quality and ensures compatibility for querying and analysis. The following table presents the supported dataType values:

| Data Type | Description | Notes |
|---|---|---|
| String | The default type for text-based data | Used if no type is specified |
| Integer | Whole numbers only | Enables numeric comparisons |
| Float | Decimal numbers | |
| Date | Calendar date | Requires a `dataFormat` |
| Time | Time of day | Requires a `dataFormat` |
| Datetime | Date and time combined | Requires a `dataFormat` |

⚠ For `Date`, `Time`, and `Datetime`, a `dataFormat` must be specified (e.g., `yyyy-MM-dd` for dates).

### Record

In GEOME, FIMS validation and upload revolve around the concept of a *Record*. A Record represents a single instance of an Entity (e.g., a sample, tissue, or event) and is typically structured as a key-value (k:v) map where:

- The key is the `columnUri` (The value is the data associated with that field)
- The DataReader implementation is responsible for mapping `columnName -> columnUri` values when creating an instance of a `Record`.

Each type of Record has a corresponding **RecordValidator** implementation to validate its contents. The default and most commonly used record type is the `GenericRecord`, which is validated strictly against the project configuration with no additional built-in validation logic. At the moment, GEOME supports the following types of Records:

| Record Type | Description |
|---|---|
| GenericRecord | The default record used for most FIMS entities |
| FastaRecord | Specialized for FASTA file validation and handling |
| FastqRecord | Handles FASTQ-formatted sequencing data |
| PhotoRecord | Used for image metadata and photo uploads |

### RecordSet

A RecordSet is a collection of `Record` instances belonging to a specific `Entity`.

### DataSet

A Dataset is a collection of `RecordSet` instances that represents a complete upload session or data package.

- If a `Dataset` contains a **child** entity's `RecordSet`, it must also include the associated **parent** entity's `RecordSet`.
- Use the `DatasetBuilder` class to create valid `Dataset` objects that respect `entity` hierarchies and parent-child relationships.

### Data Readers

DataReader implementations are responsible for parsing specific file formats and converting them into one or more `RecordSet` instances. When a file is uploaded for validation:

- It is passed to the `DataReaderFactory`
- The factory selects the appropriate `DataReader` based on the file extension

- The selected `DataReader` converts the file into a `RecordSet`. The `handlesExtension()` method in each `DataReader` must return true for supported extensions (e.g., .csv, .fasta).

> ⚠ **Note:** If multiple `DataReader` implementations handle the same file extension, **only one** (1) can be enabled at a time. This limitation may be removed in future versions.

#### Entity

Custom Entity types can be created by setting a subclass to the base `Entity` class. Subclasses must be located in the `biocode.fims.digester` package, for proper polymorphic serialization/deserialization via Jackson. Some `Entity` features are:

- Customize and fix specific components of an entity
- Add custom validation logic to be executed during `ProjectConfig` updates
- Ensure entity definitions are complete and not missing pertinent information

---

## REST Services

FIMS REST Services are available at: https://api.geome-db.org/apidocs/

---

## Versioning

FIMS REST services are now versioned, allowing clients to target specific versions of the API to ensure compatibility and stability. The default version is v1, and we are supporting **V1** and **V1.1**.

> ⚠ **Note:** You can specify the version in one of two ways:
> 1. `Api-Version: {version}` or via the URL
> 2. `http://biscicol.org/biocode-fims/rest/{version}/...`
>
> So, if you want to access via URL for the version **v1.1**, it would look like this:
> `http://biscicol.org/biocode-fims/rest/v1.1/...`

If no version is specified, the system will use v1 by default. More details about version-specific resources and changes will be provided in the upcoming documentation.

---

## User Accounts

User accounts are not required to look up or resolve BCIDs, but are required to:

- Work with projects and expeditions
- Create new BCIDs
- Upload or manage data in GEOME

This section outlines how to obtain and manage user access within the Biocode-FIMS system.

### Account Creation

User accounts can be created in one of two ways:

- By the Biocode-FIMS instance owner
- By project administrators

If you need access to a specific project or permission to create expeditions, contact your project administrator. They can authorize any existing user in the system.

> 🔗 Learn more about authentication and authorization in the OAuth2 documentation.

### Project Administrators

Each project has one designated project administrator, assigned by the Biocode-FIMS instance owner. Project administrators have the following permissions:

- Add or remove users from the project
- Create new users
- Assign expedition creation rights
- Set the location of the validation XML file
- Define and update the project abstract

## curl Examples

Plenty of curl examples are available at our Swagger Application Programming Interface documentation at: https://api.geome-db.org/apidocs/

## oauth2

GEOME uses the OAuth2 protocol to provide secure access to protected resources. This section explains how to register your app, obtain access tokens, and make authenticated requests.

App Registration
All developers must register their application to access the authentication system.

Contact the system administrator to register your app. You will be issued a:

- `client_id` — Public identifier for your app
- `client_secret` — Private key used to authenticate your app (keep this secure)

### Authorization

#### Request Authorization Code

Your client application must make the following `GET` request:

GET /id/authenticationService/oauth/authorize

Required Query Parameters:

| Parameter | Description |
|---|---|
| `client_id` | Your registered application's client ID |
| `redirect_uri` | The URI where the authorization code should be sent |

| | |
|---|---|
| `state` | *(Optional)* Will be returned unchanged in response |

Response Parameters:

| Parameter | Description |
|---|---|
| `code` | One-time, 20-character authorization code (expires in 10 minutes) |
| `state` | Returned only if it was included in the request |

### Access Token

#### Exchange Code for Access Token

Your client application must make the following `POST` request:

`POST` /id/authenticationService/oauth/access_token

Required Query Parameters:

| Parameter | Description |
|---|---|
| `client_id` | Your registered application's client ID |
| `client_secret` | Your registered client secret |
| `code` | The authorization code from the previous step |
| `redirect_uri` | Must match the URI used during the authorization request |
| `state` | *(Optional)* Will be returned in the response |
| `grant_type` | *(Optional)* Set to password for direct username/password login |
| `password` | *(Optional)* Required if grant_type=password |
| `username` | *(Optional)* Required if grant_type=password |

Response Parameters:

**JSON Response**

```
{
  "access_token": "abc123xyz456...",
  "refresh_token": "def456uvw789...",
  "token_type": "bearer",
```

```
  "expires_in": 3600,
  "state": "your_state_here"
}
```

### Refresh Token

#### Refreshing an Access Token

Your client application must make the following `POST` request:

POST /id/authenticationService/oauth/refresh

Required Query Parameters:

| Parameter | Description |
|---|---|
| `client_id` | Your registered application's client ID |
| `client_secret` | Your registered client secret |
| `refresh_token` | The refresh token previously issued |

⚠ **Note:** The refresh token must be **less than 24 hours old**. If correct, a new access token and refresh token will be issued. The old refresh token becomes invalid.

Response Parameters:

**JSON Response**

```
{
  "access_token": "new_access_token",
  "refresh_token": "new_refresh_token",
  "token_type": "bearer",
  "expires_in": 3600
}
```

### API Access

To access protected REST services, add your `access_token` to the request URL:

GET /id/userService/profile

Profile Response:

**JSON Response**


```
{
  "firstName": "John",
  "lastName": "Doe",
  "email": "j.doe@example.com",
  "institution": "Your University",
  "userId": 12345,
  "username": "jdoe",
  "projectAdmin": true,
  "hasSetPassword": true
}
```


You can use this token with any GEOME REST endpoint by appending: `?access_token=your_access_token` to the request URL, in order to access the service. Here's an example of how a request would look:

GET https://geome-api.org/id/userService/profile?access_token=abc123xyz456

⚠ **Note:** In the previous example, just replace abc123xyz456 with your actual access token.

---

## Resolution System

The diagram below illustrates how BCIDs (Biocode Identifiers) interact with local identifiers, the web, and EZID's Name-to-Thing (N2T) resolution service.

### How It Works

1. A field researcher uses a local identifier, such as MBIO56, during sample collection.

2. This data is uploaded to FIMS, where it is:

    - Linked to a specific resource category (e.g., R2)
    - Assigned a globally unique identifier following the ARK scheme

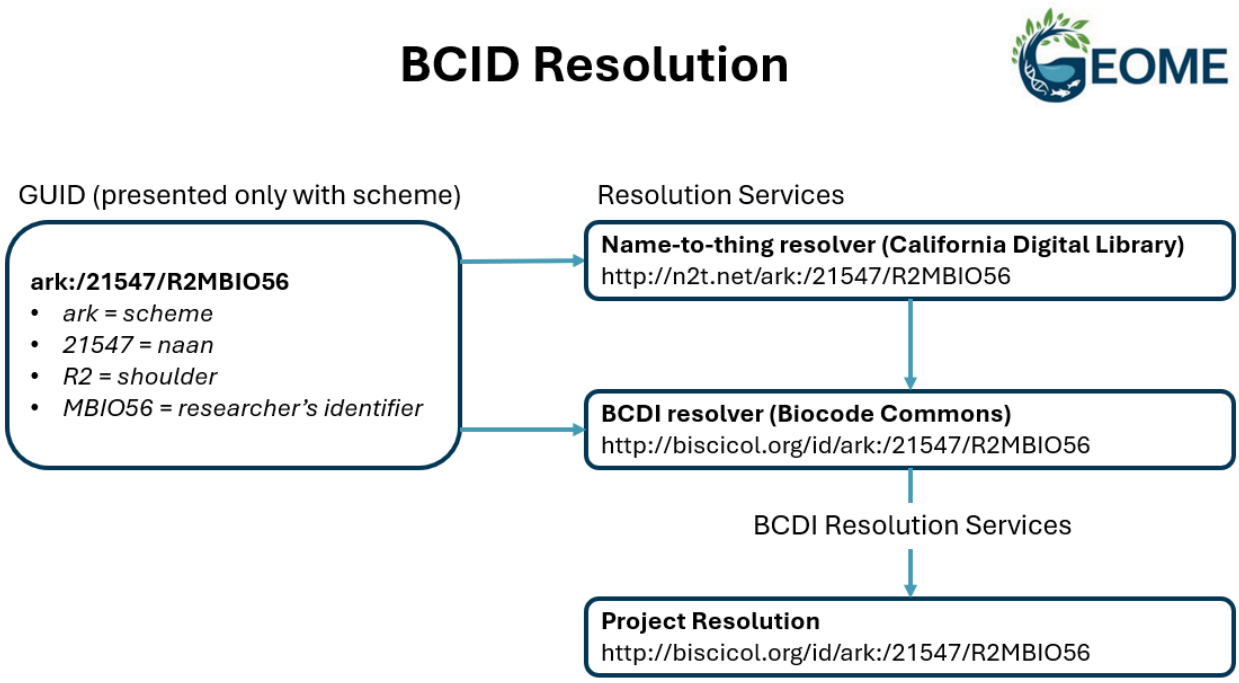
3. GEOME's FIMS system is registered under the ARK namespace with a Name Assigning Authority Number (NAAN) of 21547.

4. The final BCID looks like this: `ark:/21547/R2MBIO56`

5. When a user attempts to resolve this identifier (e.g., by entering it in a browser or linking to it):

    - The N2T resolver (https://n2t.net/ark:/21547/R2MBIO56) redirects the request
    - The request is handled by the BCID resolution service, which delivers the associated metadata or resource

> ⚠ **Note:** This resolution architecture ensures that even locally generated identifiers can be transformed into globally unique, persistent, and resolvable identifiers on the web.

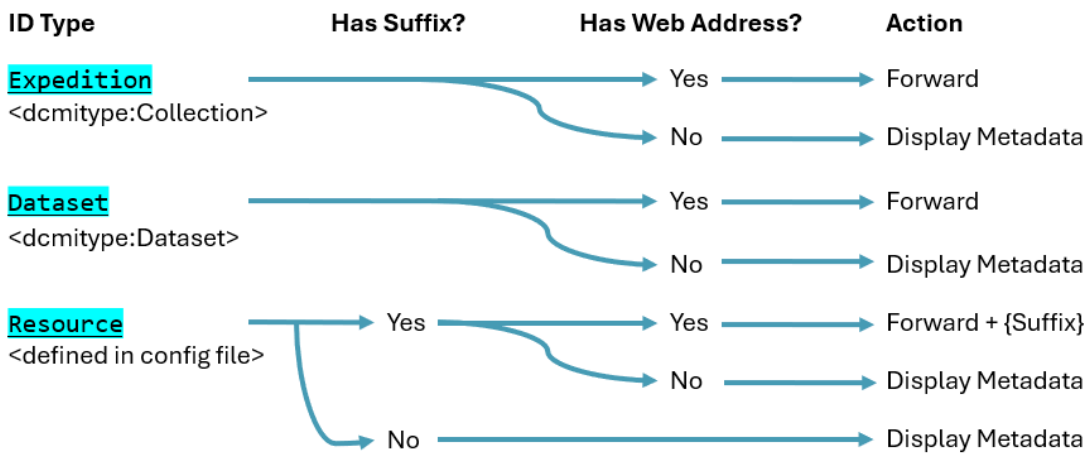



The following chart illustrates how BCID resolution functions for expeditions, datasets, and resources within the FIMS system. Each action results in either forwarding to a target URL or displaying metadata.
Forwarding behavior is determined by one of two factors:

- A target web address defined in the database, or
- A default redirect rule is specified in the project's configuration file (if no target is set)

© GEOME - 2025

# BCID Resolution Services




**Forward Logic (BCID Resolution)**

```
// If a target web address is defined in the database, use it
if (bcid.webaddress != null) return bcid.webaddress;  // From database
else {
   if (ID Type = Expedition) return <metadataParam.expeditionForwardingAddress>{ark};
   else if (ID Type != Dataset) return metadataParam.conceptForwardingAddress/{ark}/{suffix};
   else return "Display Metadata Address" // Uses Apache strSubstituor
   }
```

---

## Types of Identifiers

GEOME's FIMS system uses a centralized minting service to assign persistent identifiers for three entity types:

- Expeditions
- Datasets
- Resources

Each FIMS installation must use its own Name Assigning Authority Number (**NAAN**) and be registered with the California Digital Library's EZID service to mint Archival Resource Keys (**ARKs**).

© GEOME - 2025

**Expedition Identifiers**
*Resource Type:* http://purl.org/dc/dcmitype/Collection
*Mutable:* Represents the most current version of an expedition's spreadsheet
*Metadata includes:*

- `expeditionCode`
- `expeditionTitle`
- `userId` – Creator of the expedition
- `ts` – When was it loaded
- `projectId` – Associated project
- `public` – Public visibility flag

**Dataset Identifiers**
*Resource Type:* http://purl.org/dc/dcmitype/Dataset
*Inmutable:* Represents a fixed snapshot of data within an expedition
Belongs to a specific expedition
*Metadata includes:*

- `webAddress` – URL where the dataset is stored in its native format (varies by installation)
- `userId` – Uploader of the dataset
- `doi` – Optional DOI, in addition to the ARK

**Resource Identifiers**
*Resource Type:* Defined per project in the configuration file
*Properties:*

- Belong to an expedition
- Multiple resource types may be defined per expedition (e.g., samples, events)
- Support suffix-passthrough: a locally unique suffix can be appended to the root identifier to uniquely resolve individual records (e.g., material samples)

For example, if a "Material Sample" resource identifier is created for an expedition containing 1000 samples, 1000 resolvable identifiers are formed by appending unique suffixes to the root resource ARK.

> ⚠ **Note:** A resource identifier + local primary key from the most recent dataset = a globally unique identifier for that resource instance.

# Part 4: Frequently Asked Questions

## Getting Started Questions

**What is GEOME?**

The **Genomic Observatories Metadatabase** (GEOME) is a web-based database for capturing and managing metadata for biological samples. You can access it at www.geome-db.org.

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

To simplify future uploads and maintain consistency, we recommend using the GEOME **metadata template** as your standard spreadsheet for recording sample information during both field and laboratory work. The template can be customized as much as you like (e.g., change column order, add extra fields), as long as all mandatory fields required by GEOME remain unchanged and are properly filled in. When you're ready to upload:

1. Save your file as either:
    - An **Excel** (.xlsx) file using the structure provided under "Generate Template", or
    - A **CSV** file with a single row of column headers
2. During upload, make sure to select the correct file format option in the GEOME interface.

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

When publishing data uploaded to GEOME for the first time, you should include the GUID (Globally Unique Identifier) for your expedition or dataset in the data accessibility section of your publication. The **GUID** is a persistent, resolvable identifier for your metadata and should be:

- Acknowledged in the original publication of the dataset
- Cited when any part of the dataset is reused or downloaded

> **Example:**

> 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*])

You can adapt the citation for other data types, such as Sanger Sequences (FASTA) or microsatellite datasets. To find the GUID related to your own dataset:

1. Go to **Workbench** > **My Expeditions**
2. Locate your dataset
3. The **GUID** is listed next to the label "**Identifier**"

**How do I cite, or acknowledge use of GEOME?**

If you use GEOME for data storage, access, or publication, please cite the original publication as follows:

> Deck, J., Gaither, M.R., Ewing, R., Bird, C.E., Davies, N., Meyer, C., Riginos, C., Toonen, R.J., & Crandall, E.D. (2017). *The Genomic Observatories Metadatabase (GEOME): A new repository for field and sampling event metadata associated with genetic samples*. PLOS ONE, 12(8), e0182006. https://doi.org/10.1371/journal.pbio.2002925

Including this citation in your work helps support the continued development of the GEOME platform.

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

To reference a dataset in your publication, project documentation, or within GEOME, use its **GUID** (Globally Unique Identifier). The GUID is a persistent identifier for expedition-level metadata and should be:

- Acknowledged in the original publication of the dataset
- Cited when any part of the dataset is reused or downloaded

To find the GUID related to your GUID, for a Single Dataset:

1. Go to **Workbench** > **My Expeditions**
2. Locate the relevant dataset
3. The **GUID** is listed next to the label "**Identifier**"

For Multiple Datasets in a project:

1. From the **Workbench**, open **Project Overview** for your selected project
2. You'll see a table under "**Expedition Title**"
3. Each expedition listed will have a **GUID** associated

> **Example** GUI format of different expeditions of the same project:
> https://n2t.net/ark:/21547/CXt2
> https://n2t.net/ark:/21547/CZq2
> https://n2t.net/ark:/21547/CVd2

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

Once uploaded, your data becomes permanently accessible through the GEOME web interface via the Query function. You can search your metadata using:

- Geographic bounding boxes
- Search terms from your uploaded metadata (e.g., species, location, expedition name)

Each dataset is saved as a unique expedition, identified by a globally unique persistent identifier (GUID). Additionally, each sample (`materialSampleID`) receives a **BCID** (Biocode Identifier), which is also globally unique and resolvable. To find the GUID related to your GUID, for a Single Dataset:

1. Go to **Workbench** > **My Expeditions**
2. Locate the relevant dataset
3. The **GUID** is listed next to the label "**Identifier**"

For Multiple Datasets in a project:

4. From the **Workbench**, open **Project Overview** for your selected project
5. You'll see a table under "**Expedition Title**"
6. Each expedition listed will have a **GUID** associated

> **Example** GUI format of different expeditions of the same project:

https://n2t.net/ark:/21547/CXt2
https://n2t.net/ark:/21547/CZq2
https://n2t.net/ark:/21547/CVd2

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

GEOME is designed to store and manage **metadata associated with genetic sequences**, not the genetic data itself. When you perform a query in GEOME:

- You will receive **metadata record**s that describe the samples and sequences, including information like taxonomy, collection location, and associated identifiers.
- Accession numbers for sequence data housed in external repositories (e.g., NCBI/GenBank) are included in this metadata.

To access the actual genetic data (e.g., FASTQ or FASTA files), you must:

1. Use the accession numbers provided in the metadata to retrieve sequences directly from NCBI/GenBank.
2. In some cases, FASTA files may be downloadable directly from GEOME if the data contributor opted to upload them along with the metadata (this is optional).

**Summary:** GEOME helps you discover and describe genetic data; it does not store all sequence files unless explicitly provided by the contributor

**What are the advantages of depositing metadata in GEOME?**

Depositing metadata in GEOME provides multiple key benefits:

- **Streamlined Data Submission**; GEOME simplifies the process of depositing raw genetic data to the INSDC's Sequence Read Archive (SRA) while ensuring persistent links to ecological and sample metadata.
- **FAIR Data Principles**; GEOME supports Findable, Accessible, Interoperable, and Reusable (FAIR) data practices by maintaining standards-compliant metadata and globally unique identifiers.

- **Collaboration-Friendly**; Designed to support large collaborative groups, GEOME provides shared metadata templates, project-level organization, and controlled access to expeditions and datasets.
- **Batch Data Retrieval**; Expedites the batch retrieval of genetic sequence data from repositories like NCBI SRA using linked accession numbers stored with your metadata.

> These advantages and several others are fully described in:
> Deck et al. (2017) - *The Genomic Observatories Metadatabase (GEOME)*
> Riginos et al. (2020) - *The importance of metadata and linking genetic data to ecological context*

**What is derived data?**

Derived data are databases generated through analytical processing or bioinformatic pipelines from original raw genomic reads. GEOME is particularly interested in the following types of derived genetic data:

- SNP calls in formats such as:
    - VCF (Variant Call Format)
    - Genepop
    - Structure
- Microsatellite datasets or other genetic datasets created from the same biological samples.

We are especially interested in URLs (permanent links) to databases hosted in external repositories such as Dryad or others.

> **Note:** If you are submitting multiple sets of derived data, please refer to the instructions below for proper formatting and metadata linkage.

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

The purpose and capabilities of the Genomic Observatories Metadatabase (GEOME) are described in publications by Deck et al. (2017) and Riginos et al. (2020). To explore GEOME further, please visit www.geome-db.org, where you can browse its features, request to join an existing *project* or *team*, or even start your own.

We welcome feedback on how GEOME can better support your research needs. If you find value in the platform, consider promoting its use among your colleagues, students, or authors of manuscripts you review or handle as an editor. Your engagement helps expand the community and strengthens data sharing practices across the genomics research landscape.

---

## Technical Questions

**How can I update multiple expeditions at once?**

Yes, GEOME allows you to update multiple expeditions simultaneously by using the **Project CSV Archive** feature. The process starts by going to *Project Overview* and clicking the "PROJECT CSV ARCHIVE" button. This will download a zip archive containing all the project's data to your local computer.

- **Make a backup** of the original archive before making changes, in case you need to revert.
- **Extract a working copy** of the archive. Inside, you will find one or more CSV files – each representing a worksheet or entity used to upload data.

Open the CSV files and apply your updates. Be sure to retain the following columns, as they are essential for mapping each record correctly:

- `bcid`
- `expeditionCode`
- `projectId`

Once your edits are complete:

1. Go to **Load Data** in the GEOME interface.
2. Check the boxes for each entity (CSV file) you've updated.
3. In the expedition selector, choose "Multiple Expeditions".
4. Upload your files and click **LOAD** to apply the updates.

This process ensures that all your modified records are correctly reassigned to their respective expeditions.

### Using a pre-existing materialSampleID formed as a URI

In GEOME, the `materialSampleID` is a researcher-assigned field identifier used to uniquely identify each biological sample. It forms the basis of a persistent URI, created by appending the `materialSampleID` to a designated URI root. To ensure proper resolution and URI safety, the `materialSampleID` field can only include the following characters:

> [a–z, A–Z, 0–9, +, =, :, ., _, (, ), ~, *]
>
> ⚠️ **Note:** Characters commonly found in URIs – such as slashes (`/`) and dashes (`-`) – are not allowed in the `materialSampleID` field.

If your existing `materialSampleID` is already formatted as a URI, it won't be accepted by GEOME due to character restrictions. Instead, follow this approach:

1. Insert your full URI into the `voucherURI` field:

   > **Example:** http://n2t.net/ark:/65665/34f224976-105e-4392-9d3d-4a4f1f22f048

2. Use the URI suffix – or mint a new, GEOME compatible value – as the `materialSampleID`.

   > **Example:** 34f224976105e43929d3d4a4f1f22f048

This ensures that your original identifier is preserved via `voucherURI`, while maintaining compliance with GEOME's formatting rules for `materialSampleID`.

### How do I delete an existing metadata record in GEOME?

If you need to delete a metadata record – even for just a single sample – the simplest and most reliable method is to download the entire expedition, make your edits locally, and then replace the existing data. Here's a step-by-step guide on how to do it:

1. Go to the expedition containing the record you want to delete.
2. Download the full expedition dataset (CSV or Excel format).
3. Open the file and remove the row(s) corresponding to the sample(s) you wish to delete.
4. Return to GEOME, go to **Load Data**, and upload the modified file.

5. Be sure to check the box labeled "*Replace expedition data*" before submitting.

This will overwrite the expedition with your updated file, effectively removing any deleted records from the system.

**How to update data, and what does the "Replace Expedition Data" box mean?**

To update metadata in GEOME, you should begin by downloading the current version of your expedition's data. From the *Project Overview*, select the expedition you wish to update and download the *Excel Workbook*. Make all necessary edits directly in this file – whether you're correcting entries, adding new samples, or preparing to delete existing ones. Then go to **Load Data** in your Workbench and ensure you're submitting the metadata to the same project as before. When uploading your revised metadata:

- If you're deleting records, check the "*Replace Expedition Data*" box. This option tells GEOME to replace the entire existing dataset with your updated file.

  > ⚠️ **Note:** When using this option, you must retain all original columns from the downloaded file to avoid data loss or validation errors.

- If you're only updating or adding records (not deleting any), you can leave the "*Replace Expedition Data*" box **unchecked**. This will append or update records without overwriting the whole expedition.

Next, be sure to select the same Expedition name you used previously, and proceed to upload the updated file. This process applies whether you're modifying sample metadata or uploading additional file types like **FASTA** or **FASTQ** for submission to SRA.

**What do 'concatenated and separated' and 'delimited list' mean?**

In GEOME, when you have multiple values for a single metadata field, they need to be entered in a way that allows the system to recognize and separate each entry programmatically. This is done using a pipe

symbol (|) as a delimiter between entries. For example, in the field `collectorList`, if your sample was collected by both Charles Darwin and Alfred Wallace, you should enter:

> 'Charles Darwin | Alfred Wallace'

This format is called a delimited list, where values are concatenated (joined) but separated by a consistent symbol (in this case, |) to distinguish them as individual entries.

**What if I have multiple values for a particular metadata field, like `associatedReferences`, `permitInformation`, `environmental_medium`, or `derivedDataXXX`?**

For most metadata fields – except those related to derived data – you can enter multiple values in a single cell using the pipe-delimited format:

```
Reference A | Reference B | Reference C
```

This concept applies to fields like:

- `associatedReferences`
- `permitInformation`
- `environmental_medium`

Unlike the previous fields, `derivedDataXXX` datasets require separate rows for each dataset. Here's how to structure them:

1. Fill in all metadata fields for your sample as you normally would.
2. Highlight and copy the entire row for that sample and paste it below the original.
3. In the copied row, update **ONLY** the `derivedDataXXX` fields to point to the second derived dataset.
4. Repeat this process for each additional derived dataset you need to associate with that sample.

This ensures that each derived dataset is linked properly while preserving the rest of the sample metadata.

**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 your samples were collected from sensitive locations, such as a protected habitat or a market where the precise origin is unknown, you can still contribute meaningful spatial metadata without disclosing exact coordinates.

For example, if a sample was purchased freshly killed at a market (e.g., a fish market), you may use the location of the market as the sampling site and set the field `coordinateUncertaintyInMeters` to 100,000 (i.e., 100 km). This indicates that the actual collection site may be up to 100 km from the reported coordinates.

Similarly, if the exact sampling location should be kept confidential, for reasons such as conservation or respect for local custodians, you are encouraged to:

- Provide proximal (approximate) coordinates
- Round the coordinates to reduce precision
- Set `coordinateUncertaintyInMeters` to 100,000 to reflect reduced spatial accuracy

This approach enables you to maintain the **scientific value** of your data by contributing general spatial context while ensuring that sensitive locations remain protected.

---

## 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 update or replace existing sequence data by reloading your **FASTA file** using the **Load Data** function in GEOME. To do this, you should:

1. Prepare a new FASTA file containing the updated sequence data.
2. Ensure that the FASTA file references the same tissue identifiers (`tissueID`) that were previously uploaded.
3. Go to **Load Data** in your Workbench.
4. Upload the new FASTA file.
5. When prompted, select the same **genetic marker** you used previously.

This process will overwrite the existing sequence data for that marker with the new file contents, similar to how expedition metadata can be replaced.

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

While it is best practice first to upload your metadata to GEOME and then push your data to the SRA, you can still link your metadata in GEOME to existing sequence data already submitted to the **Sequence Read Archive** (SRA). To do this, follow these steps:

1. Use a metadata template that includes the `Tissue` entity. This entity is essential for associating your metadata with sample-level sequence data in the SRA.
2. Add a `bioSampleAccession` field to the tissue metadata sheet. This field will store the BioSample accession number (e.g., `SAMDXXXXXXXX`).
3. Match each bioSampleAccession to the correct `tissueID`. Each tissueID should correspond to a `materialSampleID` you've defined.

   > **Note:** If a single sample has multiple BioSample accessions, multiple tissue records will be created in GEOME during upload.

4. Upload the completed metadata file to GEOME. The system will recognize the `bioSampleAccession` values and link your metadata to the appropriate records in the SRA.

This approach allows you to maintain synchronized, well-structured metadata in GEOME – even for sequences already submitted to external repositories.

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

To access detailed SRA metadata – such as library strategy, sequencing platform, and read type – you can use the rentrez package in R https://github.com/ropensci/rentrez. This package allows you to query NCBI databases programmatically and link SRA metadata with records stored in GEOME.

> For full documentation about the feature, visit: https://docs.ropensci.org/rentrez/

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

To cite individual `materialSamples`, `tissues`, or `events` retrieved from GEOME, use their ARK identifier in combination with the Name-to-Thing (N2T) resolver. This ensures your citation is persistent and globally resolvable. For example, to cite a sample in GEOME:

> **Example:** https://n2t.net/ark:/21547/CVJ2BMOO_00004

You can use the link format in publications, data papers, or reports when citing material samples or field events documented in GEOME. This format structure includes:

- `n2t.net`: the resolution services that have a persistence mission
- `ark`: the open identifier scheme
- `21547`: the Name Assigning Authority Number (NAAN) for GEOME
- `CVJ2`: the expedition identifier
- `BMOO_00004`: the locally unique suffix for the sample

For genetic data that has been accessioned into public repositories such as **NCBI**, cite the corresponding identifiers directly from the repository. The following link refers to the BioSample record associated with the genetic data, allowing readers to access both the sequence and its metadata:

> **Example:** https://www.ncbi.nlm.nih.gov/biosample/SAMN10240383

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

If your sequence data has already been submitted to NCBI, you can still integrate it into GEOME by referencing the appropriate accession numbers in your metadata template. For data in the **Nucleotide database**, add the accession number(s) to the `associatedSequences` column in your sample sheet:

> **Example:** `MT123456` or `AY987654`

For **next-generation sequencing** data already submitted to the Sequence Read Archive (SRA):

- Insert the BioSample accession number (e.g., SAMNxxxxxxxxx) into the `biosampleAccession` field in the Tissue metadata sheet.
- Optionally, you can also include a direct link to SRA in the `associatedSequences` column:

> **Example:** https://www.ncbi.nlm.nih.gov/sra/?term=SAMN09015327

This approach allows GEOME to reference your existing data in the SRA or GenBank, but only in one direction – from GEOME to NCBI. If you want SRA to point back to GEOME, you must first upload your metadata to GEOME, then use the GEOME FASTQ loading service to push data to the SRA. This ensures proper linkage and synchronization between the two systems.

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

No, GEOME is not a sequence repository. It is designed to complement public repositories like NCBI/GenBank, not replace them. NCBI and GenBank apply important quality control checks to genetic data submissions, which are essential for maintaining data integrity and usability across the scientific community. You may upload **FASTA files** to GEOME to associate sequences with metadata, but this should be considered supplemental and not a substitute for submitting your data to a dedicated sequence repository.

> ⚠️ **Note:** GEOME does not accept **FASTQ** files directly.

However, GEOME offers a FASTQ submission tool that helps you prepare your data for the NCBI Sequence Read Archive (SRA). This tool:

- Ensures your metadata is complete and validated
- Packages your data correctly for SRA submission
- Creates proper links between your GEOME metadata and the sequence data in NCBI

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

You should upload the **raw reads** (e.g., FASTQ files), not just the SNP genotypes. GEOME encourages users to use its **FASTQ** submission tool to prepare a submission package for the NCBI Sequence Read Archive (SRA). By submitting unfiltered, unmanipulated sequence data, GEOME helps ensure that:

- **Ascertainment bias is reduced**, allowing for broader future use of your data (a site not called as an SNP in your dataset may still be useful in another context).
- **Subjective filtering decisions** (e.g., coverage thresholds, trimming, or genotype likelihoods) are not locked into the dataset, enabling downstream users to apply their own analysis choices relevant to their research questions.

This process preserves the raw data so future researchers can reanalyze it alongside other datasets using different parameters or tools. If you also wish to share derived data – such as SNP genotypes, microsatellite data, alignments, or Amplicon Sequence Variants (ASVs) – you can deposit these in a public open-access repository (e.g., Dryad) and link them to your GEOME records using the `derivedGeneticData` fields.

**What if my genetic data is for microsatellites?**

While there is currently no standardized, centralized repository for microsatellite genotype data, this information is still highly valuable and can be preserved and linked through GEOME. For **microsatellite data**, as well as other derived datasets that lack a dedicated repository (e.g., SNP genotype files, ASV tables):

- We recommend depositing your dataset in a permanent open-access repository, such as Dryad or Zenodo.
- You can then upload the sample metadata to GEOME and use the `derivedGeneticData` fields to link each sample to its corresponding dataset in the external repository. This will make your data discoverable through GEOME while supporting FAIR data practices by keeping your sample data findable, accessible, and reusable.

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

GEOME fully supports community and environmental DNA (eDNA) samples. In these cases, the `materialSampleID` should represent the environmental or community sample itself, rather than an individual organism. To ensure your eDNA samples are properly formatted, follow these guidelines:

- `basisOfRecord` (Sample entity): Set this to "*EnvironmentalDNA*" to indicate the sample type.
- `environmentalMedium` (Event entity): Use this to describe the environment from which the sample was taken (e.g., soil, seawater, air).
- `samplingProtocol` (Even entity): Specify your collection method (e.g., filter size, extraction protocol, or sequencing workflow).
- `scientificName` or `taxonID`: If your sample includes a broad range of organisms, list the highest common-level taxonomic group. If unknown, you may use "Unknown" at the phylum level or higher.

As with all sample types in GEOME, we recommend using the **FASTQ submission tool** in GEOME to prepare a metadata-validated submission package for the NCBI Sequence Read Archive (SRA). This should be done before submitting your sequence data to NCBI/GenBank to ensure that proper links between GEOME metadata and genetic data are established.

For derived datasets (e.g., ASV or OTU tables), deposit them in a permanent open-access repository such as Dryad or Zenodo. These derived genetic datasets 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?**

GEOME supports flexible metadata structures for complex biological systems such as host-symbiont, host-microbiome, host-parasite, or foundation species-community relationships. You have two (2) main options depending on your sampling approach and how you wish to manage taxonomic assignments.

1. Use a single `materialSampleID`

   Use this option when the host and symbiont (or other co-sampled organisms) are physically part of the same collected sample.

- Assign one `materialSampleID` for the shared sample
- Differentiate the components using distinct `tissueID`s (e.g., one for host tissue, one for symbiont DNA)
- Link each tissue to its own sequence data or derived genetic dataset
- `associatedOrganisms` can be used to name the second organism
- `environmentalMedium` can be used to describe the surrounding context (e.g., coral tissue)

⚠️ **Note:** You can only assign **one taxon** to the `materialSampleID`, meaning only the host or the symbiont can be formally described in the taxonomy fields.

2. Use two `materialSampleID`s (Recommended for independent analysis)

   Use this option when you want to treat the host and symbiont (or other biological entities) as distinct samples with their own metadata and taxonomy.

- Create separate `materialSampleID`s for each organism
- Assign individual taxonomic classifications, collection information, and sequence data
- Use the `associatedOrganisms` field to reference the linked sample by its materialSampleID

⚠️ **Note:** The second option allows for *independent metadata tracking*, taxonomic assignment, and clearer linkage between host and partner species.

Both approaches are supported, and the choice depends on how closely the entities are integrated in your study design and how you intend to analyze or share the data.

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

You can upload your sample metadata to GEOME at any stage of your project, even if your genetic data has not yet been submitted to NCBI/GenBank or is still in preparation. Including your metadata early helps validate your dataset and ensures it is ready for integration with genetic data later. For next-generation sequencing data, we recommend using GEOME's **FASTQ tool** to prepare your SRA (Sequence Read Archive) submission package. This guarantees metadata is complete and properly linked before submission to NCBI/GenBank.

When uploading your data to GEOME, you have the option to control access to your data as follows:

1. Keep your data private during manuscript preparation or peer review. GEOME allows data to remain private for up to **2 years**; after that, you can either make it public or remove it.
2. Use the "**discoverable**" setting under *Project Settings* to allow users to find your project metadata, even if the underlying data is not yet public.

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

You can still upload your sample metadata to GEOME even if no genetic data has been generated yet. This helps others discover the existence of your sample and promotes future collaboration. In this case, fill out the metadata template as fully as possible, especially the **Tissue entity** section. From there, enter the word **available** in the field `tissueRemarks`, to indicate that the tissue sample exists and may be shared upon request.

**What if I no longer have any sample/tissue corresponding to that genetic data?**

Even if the physical sample or tissue is no longer available, GEOME still encourages you to upload the metadata associated with that sample and its collection event. This ensures that your genetic data is properly contextualized and can be reused or referenced in future research. In this case, fill out the metadata template as thoroughly as possible, especially fields related to the collection location, date, and sampling protocol. From there, enter the word **unavailable** in the field `tissueRemarks`, to indicate that the tissue no longer exists or cannot be provided.

**How do I represent pooled RADSeq data?**

To represent pooled RADSeq data, where multiple individuals are sequenced together, you should:

- Enter the count of the number of pooled individuals in the `individualCount` field.

  > Any `materialSampleID` with `individualCount` > 1 will be interpreted as pooled data

- If the pooled individuals were collected across a spatial range, indicate this by setting an appropriate value in the `coordinateUncertaintyInMeters` field.