NHLBI BioData Catalyst Data Generator Guidance

V1.0 - 20201229

Version 1.0

Document Owner

Primary Authors: Susan Eversole & Sweta Ladwa

Contact: NHLBI BioData Catalyst Coordinating Center (Tom Madden - bdc3@renci.org)

Revision History

Date (YYYYMMDD)	Version Number	Revision Reviewed/ Approved By	Brief Description of Change
20201229	V1.0	Tom Madden	Comms reviewed/approved
20201218	V1.0	Sweta Ladwa	Version shared with the Comms team for review
20201012	V0.2	Sweta Ladwa	Development version shared w/ DRMWG and as RFC Draft-7
20200511	V0.1	Sweta Ladwa	Development version

Table of Contents

1. Introduction

1.1 Overview

The National Heart, Lung, and Blood Institute (NHLBI) BioData Catalyst Ecosystem is a cyberinfrastructure, where Heart, Lung, Blood, and Sleep (HLBS) researchers can go to find, search, access, share, store, crosslink, and compute on large-scale datasets. The NHLBI BioData Catalyst Ecosystem serves as a novel, fully functioning resource in which users from a variety of disciplines and levels can perform complex operations and access newly available scientific data to make significant strides in research and beyond. For additional information, please visit the NHLBI BioData Catalyst website: https://biodatacatalyst.nhlbi.nih.gov/about

1.2 Purpose of the Data Generator Guidance

This Data Generator Guidance document is one component of a series of data management documents for Principal Investigators and study data owners (see Table 1). These documents aim to describe how data within the NHLBI BioData Catalyst Ecosystem is managed, ingested, and made available, including metadata generation, data preservation, data security, and ethics, in alignment with NIH policies and the FAIR (Findable, Accessible, Interoperable, Reusable) data principles.

Data management for the Ecosystem will continue to evolve to address new challenges as the Ecosystem matures. As such, this living document will serve as a reference/guide for any data generator looking to submit and share their data on NHLBI BioData Catalyst.

For overall governance structure, including the assignment of ownership, accountability, and roles for NHLBI BioData Catalyst, please reference the NHLBI BioData Catalyst Data Management Strategy document.

Table 1. NHLBI BioData Catalyst Data Management Documents

Document	Audience	Scope	Description
Data Management Strategy	External stakeholders & NHLBI BioData Catalyst Consortium members: Research community, Program Directors, NHLBI BioData Catalyst Consortium	Strategy document for data management	High-level introduction to purpose and key concepts; description of key principles; governance strategy

Data Release Management Process	Consortium members	Process documentation for NHLBI BioData Catalyst data ingestion;	Data release train phases; QA process; Goals, Roles & Responsibilities; Data release retrospective
Data Generator Guidance	External stakeholders: Future data generators, PIs, study data owners	Guidance document for prospective data ingestion	Principles and expectations of data before ingest into the NHLBI BioData Catalyst Ecosystem; Data flow and submission guidance

1.3 Objectives

The NHLBI BioData Catalyst Ecosystem serves as a custodian of NHLBI-relevant data. It is responsible for providing secure access to data within the Ecosystem for discoverability and reuse, and to maximize scientific utility by enabling users to upload additional datasets for private analysis. As such, the Ecosystem is aligned with the NIST model of custodian management of data, which provides additional information on the roles of data owner, steward, and custodianship https://nvd.nist.gov/800-53, inheriting privacy and consent, as well as other data access controls and requirements from source systems.

Supporting the Ecosystem's role as a data custodian, the NHLBI BioData Catalyst Consortium provides controls that may take the form of policy, process, or technology in four main areas:

- Data ingestion and indexing
- Data standards and quality metrics
- Data lifecycle management and process definition
- Community outreach and training

The details describing these controls as applied to each data release will be found in the Data Release Management Process (***in development***). A data release is defined as an identified set of data that is hosted on BioData Catalyst and available to users across the Ecosystem with the appropriate data access.

2. Data Flow and Submission Guidance

2.1 Data Management Core

A Data Management Core (DMC) will be established to support the overall research mission of the NHLBI. In coordination with NHLBI and the NHLBI BioData Catalyst DRMWG, the goals of the DMC will be to:

- Identify and prioritize data for onboarding into NHLBI BioData Catalyst
- Standardize and streamline the onboarding process (study registration, GUID assignment, data preparation, bucket management, etc.)
- Maximize the value of (meta)data through harmonization prior to NHLBI BioData Catalyst ingestion

The DMC will interact closely with stakeholders across the NHLBI research enterprise. The DMC will educate data generators about data standards and may subcontract to particular data generators to configure their data for ingestion into NHLBI BioData Catalyst. Prior to data ingestion, the DMC will work with data providers to assess the data with the intent of harmonizing data to required standards, and will assess data once ingested to determine the status of the data to reduce the cost of harmonization.

The DMC will work with NHLBI BioData Catalyst developers to understand the NHLBI BioData Catalyst architecture, ingest processes, and requirements for ingestion. The DMC will work with NHLBI BioData Catalyst developers to develop cloud-based workspaces for data preparation. The DMC will also seek guidance from NHLBI staff on data standards and identifying content experts, and will educate NHLBI staff on expectations for aligning datasets to agreed upon standards.

2.2 Data Flow Process

This section describes a generalized data flow process for data coming into the NHLBI BioData Catalyst Ecosystem from a data source or data generator.

The NHLBI BioData Catalyst Consortium works to streamline the data submission process and to reduce the burden of work required by data source/generator. To this end, it is expected that advancements will change the data flow and submission process as the Ecosystem matures.

Submission to the NHLBI BioData Catalyst Ecosystem is a two-step process in which the data source/ generator works first with the NHLBI Genomic Program Administrator (GPA) to complete the dbGaP Submission System (SS) and register the study data in dbGaP, which serves as the central registration authority for NHLBI BDCatalyst data.

Per the dbGaP Study Submission Guide, the GPA will determine the following:

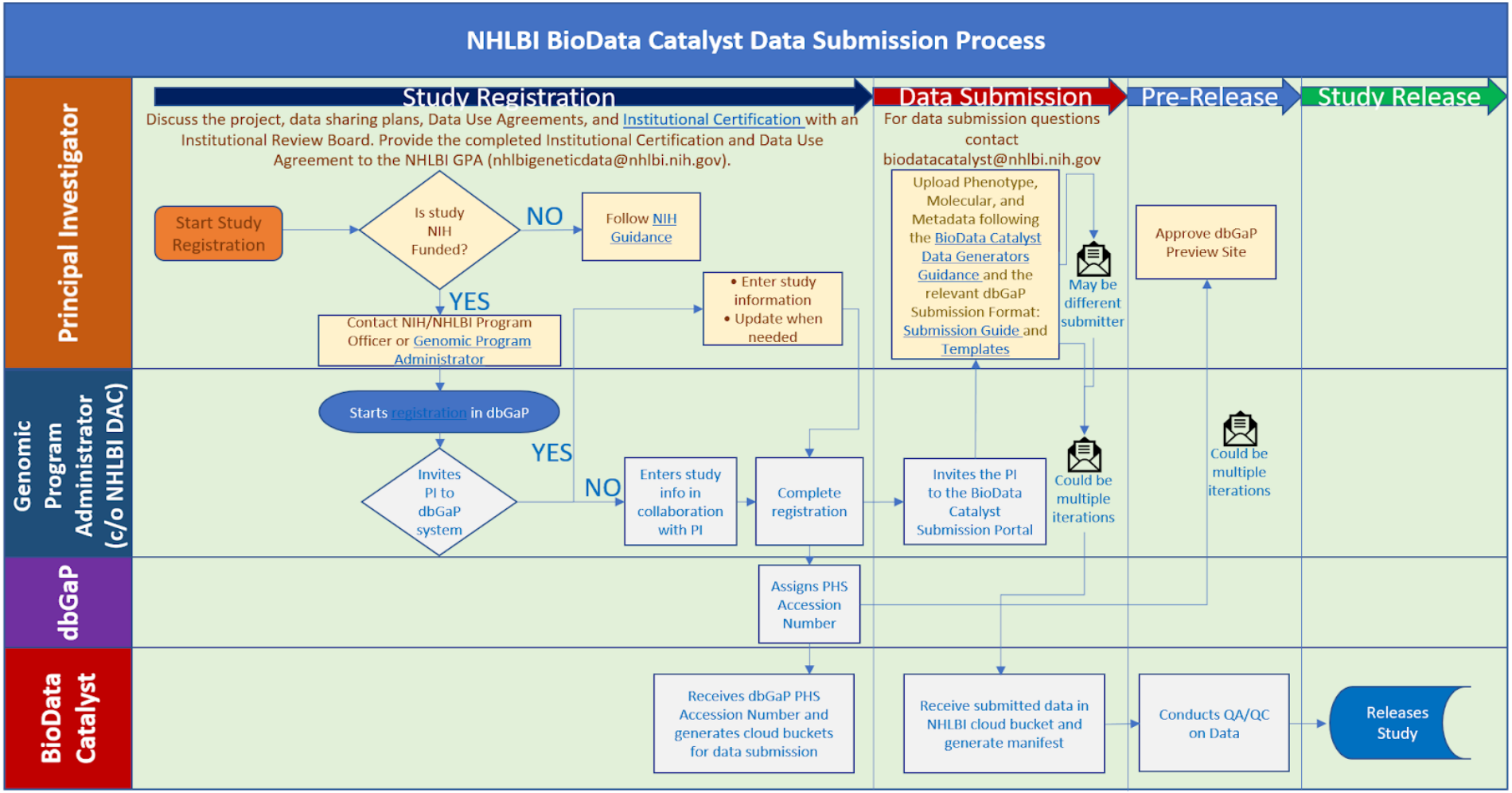
- Study Principal Investigator (PI)
- Study Project Officer (PO)
- NIH administration and funding
- Target data delivery date
- Target public release date
- Release type
- Types of data submission expected
- Inclusion in CADA (Compilation of Aggregate Genomic Data - a collection of analyses across many dbGaP studies that can be accessed with a single Data Access Request)
- Estimated study participants
- SRA submission expected
- PI assistant for study submissions

The GPA will upload the Submission Certification, Institutional Certifications, and Data Use Certification, which specifies the data use limitations (DUL). The DULs form the consent groups that will be used to parse the study data, and also determine which data access requests (DAR) can be approved through dbGaP authorized access.

BioProjects are created for each new study registered in the SS. The SS is only accessible by the GPA, PO, and PI. The SS is not the same as the dbGaP Submission Portal (SP). To make changes to the registration entry in the Submission System, contact your GPA (dbGaP Study Submission Guide).

In the second step of submission, the data source/generator will work with the GPA to submit their data to NHLBI BioData Catalyst, either directly or via the NHLBI Data Management Core. More specifically, a data generator may submit to NHLBI BioData Catalyst following the steps outlined below in Figure 1.

Figure 1: NHLBI BioData Catalyst Data Submission Process Flow



[V1.0] High-level step-by-step process, per Figure 1:

1. Register your study data with dbGaP (regardless of your data type).
2. Once you register your study with dbGaP, someone from the NHLBI Data Access Committee (DAC) will be in contact with you for next steps.
    a. Provide the Data Sharing Plan, Data Use Agreement, and Institutional Certification to NHLBI DAC.
    b. dbGaP will assign an accession number to the study
3. Complete the NHLBI BioData Catalyst Ingestion Form with guidance from the NHLBI DAC.
    a. Ingestion Form (***in development -- to be integrated with the website/help desk).***
4. Once you successfully register the study data with dbGaP and complete the NHLBI BioData Catalyst Ingestion Form, you will be invited to the NHLBI BioData Catalyst Submission Portal to submit your data.
    a. Upload your data to the portal using the prompts on the Data Submission page.
5. Once the study is ready for release on NHLBI BioData Catalyst, you will be invited to preview the data prior to its release to the appropriate audience(s) on NHLBI BioData Catalyst. All Data in NHLBI BioData Catalyst may be accessed only by those authorized to do so by submitting Data Access Requests (DARs) via dbGaP and receiving approval from the NHLBI Data Access Committee (DAC).
    a. If you have concerns, please convey them to a NHLBI BioData Catalyst representative within 5 business days of notification of the preview.
    b. If you have no concerns, the data will be live on NHLBI BioData Catalyst after 5 business days.
6. Data is released on NHLBI BioData Catalyst for access by authorized individuals.

[V1.0] NHLBI BioData Catalyst Detailed Data Workflow

1. Initiate a data submission request.
    a. Generator initiated:
        i. Contact NHLBI BioData Catalyst via email/website.
        ii. NHLBI BioData Catalyst team reroutes the request to the Program Team.
    b. NHLBI initiated:
        i. Contact the Program Team.
        ii. The Program Team sends the request to the GPA.
2. Begin registration (help video).
    a. Study PI contacts the NHLBI GPA to initiate the study registration.

b. Study PI obtains an Institutional Certification (an assurance that plans for the submission of large-scale human data to the NIH meets the expectations of the NIH Data Sharing Policy) with their institutional IRB and Signing Official.
    i. Address:
    GDS Program Administrator
    NHLBI, NIH, DHHS
    6701 Rockledge Drive
    Room 10120
    Bethesda, MD 20892-7936
    ii. Reference the NIH *Points to Consider in Developing Effective Data Use Limitation Statements* document, if necessary (e.g., if data does not already have a parent study in NHLBI BioData Catalyst that defines the DULs).
c. Study PI or designee completes the NHLBI BioData Catalyst Ingest Form [LS([1] on the NHLBI BioData Catalyst website and notifies the GPA of completion.
    i. Future – NHLBI BioData Catalyst website notifies GPA?
d. NHLBI Program Officer or GPA completes the Data Use Certification [LS([2], outlining terms and conditions for secondary use of the data.
e. Study PI prepares [LS([3] the dbGaP Study Config Template [LS([4].
f. GPA reviews information and approves the data submission to NHLBI BioData Catalyst.

3. Data preparation.
    a. Study prepares data files for submission as per the dbGaP Study Submission Guide.
        i. For questions, contact dbgap-help@ncbi.nlm.nih.gov.
    b. If study PIs require assistance, they may leverage the NHLBI Data Management Core to prepare data files, especially unharmonized data.
4. Data submission to NHLBI BioData Catalyst cloud data buckets.
    a. Study PI self-uploads data via the NHLBI BioData Catalyst auto-upload tool.
        i. Fill out the auto-upload tool access request form to gain access to the tool.
            1. Tool access request is reviewed.
            2. Tool access request is granted or rejected with feedback.
                a. If rejected with feedback, corrective action or clarification is required and the form must be resubmitted for approval.
                b. If approved, the form contents are sent to the NHLBI Cloud Bucket Team to configure the appropriate cloud bucket for data submission prior to tool use.
            3. Study PI receives the tool to auto-upload data to the NHLBI-assigned bucket (SOP).

4. NHLBI Cloud Bucket Team generates manifest [LS([5] or the tool-generated manifest is shared to provide to the Data Indexing Team (Gen3).
5. Study PI confirms the manifest contents.
6. NHLBI Cloud Bucket Team notifies the Gen3 Team that data is ready for indexing and ingestion [LS([6].

ii. Study PI provides data to the NHLBI Data Management Core.
1. Data Management Core processes, prepares, and submits data (workflow TBD).

5. NHLBI BioData Catalyst data ingest and indexing.
   a. See the Gen3 - Preparing Data for Ingestion into BioData Catalyst document
   b. See the TOPMed Freeze 5b Retrospective document
   c. Data Release Management Process document (***in development)**

Referenced Artifacts:

dbGaP Study Submission Guide

NIH Study Registration and Data Submission to an NIH-Designated Controlled-Access Data Repository

3. Data Policies

Data Protection Policy
The Ecosystem architecture is designed to implement a series of policies to protect human subjects and associated data in collaboration with NHLBI. Additional information can be found at: https://biodatacatalyst.nhlbi.nih.gov/data-protection/

Data Access Policy
NHLBI BioData Catalyst is neither the Master Data Manager for a specific data collection, nor the System of Record for a dataset. The data is provided for access with inheritance of privacy and consent from the originating system, with quality and metadata standards as provided by the source system or study. If study data does not have or provide governance or ownership designation, NHLBI provides de facto ownership and determines governance. This will continue to evolve as standards are defined and can be found at: https://biodatacatalyst.nhlbi.nih.gov/resources/data

Data managed within the system conforms to the NIST model of custodian management of data in a FISMA Moderate environment (https://nvd.nist.gov/800-53), inheriting privacy and consent from source systems, with the responsibility of providing tools for analysis for the managed data and the availability to upload additional datasets for analysis. Similar to dbGaP, the data within

NHLBI BioData Catalyst is managed by NHLBI staff and contractors. All staff and contractors with access to the data hold a Public Trust Clearance that is based on an extensive background check. All activities including data access are logged and monitored.

Access to studies registered in dbGaP is controlled by the NHLBI Data Access Committee (DAC) utilizing the database of Genotypes and Phenotypes (dbGaP) permissions infrastructure and the NHLBI Data Access Committee (DAC). To access controlled-access data in NHLBI BioData Catalyst, a user must have an approved Data Access Request (DAR) in dbGaP. The role of the NHLBI DAC is to review and approve (or deny) investigator submitted DARs and to ensure the requesting institution's investigators comply with the NIH Genomic Data Sharing Policy and any NHLBI BioData Catalyst policies.

Both the Amazon Web Services (AWS) and Google Cloud Platform (GCP) conform with industry-recognized certifications and security audit processes. Quality auditing processes will be necessary at onboarding, ingestion of data into the NHLBI BioData Catalyst Ecosystem, and acquisition of data within buckets for processing in the ETL pipelines for specific platforms, with additional points of validation to include data archiving, deleting, and publishing data within the system. This process is continuously evolving, and as such, these steps will be modified in the future to reflect processes as these are enacted.

Points at which quality auditing will be performed are:

1. At submission of data, the accession number assigned to your study will be verified.
2. Information submitted on the ingestion form will be used to verify information about the files submitted.
3. Ingestion of data processes will be verified for availability within the correct buckets according to the manifest created for the study.
4. Availability of metadata about the information submitted will be verified within the platforms.
5. Once the study is ready for release on NHLBI BioData Catalyst, data preview will be verified before you will be invited to review the data.
6. Access levels and platforms permitted for access will be verified.
7. Upon publishing and release to the data system, periodic verification of manifest information will be performed.

Research Authorization System (RAS)

Access to data will be managed by NIH Research Authentication Service (RAS). RAS is a service provided by NIH's Center for Information Technology to facilitate access to NIH's open and controlled data assets and repositories in a consistent and user-friendly manner. The RAS initiative is advancing data infrastructure and ecosystem goals defined in the NIH Strategic Plan for Data Science.

RAS will leverage a set of APIs to allow access to several data repositories. A researcher accessing NIH data resources can log in with any preferred credential (eRA Commons, ORCID, or Google) and access any integrated repository without having to log in again. Existing rules for authorization will be enforced so a user can only access data they have been authorized to view.

4. Types of Data

NHLBI BioData Catalyst is an Ecosystem capable of supporting all types of data and analyses. From sequencing data to GWAS and clinical data to image analysis, there is support for the full spectrum of scientific research. For data to be FAIR, it is advisable to contribute data conforming to de facto or de jure standards, where applicable. Additionally, for optimal use and reuse within the NHLBI BioData Catalyst Ecosystem and interoperable use across complementary NIH Institutes and Centers data commons ecosystems, adherence to metadata guidelines is strongly encouraged. ***Note: The NHLBI BioData Catalyst Consortium is currently working to develop file-level minimal metadata requirements and will include, once finalized.**

FAIR Principles

The goal of data management within NHLBI BioData Catalyst is to provide support for FAIR-TLC, an environment that ensures that the data and tools within the Ecosystem are Findable, Accessible, Interoperable, Reusable - Traceable, Licensed, Connected. This implies the creation of harmonized metadata, in a manner that allows the easy and repeatable discovery of information within an environment that supports the open access of tools and data. These principles are the basis of the Data Commons goals and must provide a coherent method for access across all data held and accessible within the ecosystem.

Metadata is, as its name implies, data about data. It describes the properties of a dataset and is critical to supporting discoverability. The datasets on the NHLBI BioData Catalyst Ecosystem are either added by a user, ingested from a controlled source such as dbGaP, or transferred from collaborative programs. In NHLBI BioData Catalyst, the definition of a set of metadata elements is necessary in order to allow identification of the vast amount of information resources managed for which metadata is created, its classification and identification of its geographic location and temporal reference, quality and validity, conformity with implementing rules on the interoperability of spatial datasets and services, constraints related to access and use, and organization responsible for the resource. Additionally, the NHLBI BioData Catalyst Ecosystem has a number of different search functionalities, and it is imperative for usability that these functionalities return consistent information about the same digital objects.

Types of data currently hosted within NHLBI BioData Catalyst include:

- Whole Genome Sequencing
- Whole Exome Sequencing
- Genotyping

- RNASeq
- Proteomic
- Clinical/Phenotypic
- Imaging
- Single-cell 'omics
- Other

5. File Formats

Study Description File
Metadata about the study files is currently provided prior to the Gen3 ingest; the full provenance information to be requested has not yet been defined. It will be necessary to track and catalog digital objects originating from studies to be used with derived files to establish provenance for these derived objects. These descriptive files should contain information about the study, duration, timing, and measurement information critical to the study design.

Manifest File
A manifest lists the contents and often location of items. Within NHLBI BioData Catalyst, the data manifest is a file containing data about the files ingested (md5, file size) as well as Access Control Lists, file name(s), and the URL for the deposition bucket (AWS, Google, or other cloud bucket). The metadata collected at the file level for NHLBI BioData Catalyst are as follows:

file_name	Name of the file
file_size	Size of file object in bytes
md5	MD5 checksum calculated for the file upon ingest
urls	Upon ingest, determination of the full path to the file in the cloud bucket (including bucket name)

Genetic Data - Genotypic/Phenotypic
SAM/BAM and related specifications: https://samtools.github.io/hts-specs/

VCF
http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

GWAS (Genome-wide Association Study)
https://osp.od.nih.gov/wp-content/uploads/What_are_Genomic_Summary_Results.pdf

Clinical Data – EHR and non-EHR
http://www.fhir.org/

6. Provenance

Current provenance information is provided by some studies in text files, providing information on source, collection, and methods. This information is necessary to track, and catalog both the original and derived files from studies. These descriptive files should contain information about the study, duration, timing, and measurement information critical to the study design.

The term "provenance" is borrowed from the confirmation of the authenticity of art in the form of documentation and the trail of ownership provided by this documentation. For our purposes, this documentation trail must provide the origination, ownership, stewardship, and transmission of data within NHLBI BioData Catalyst. At a minimum, the provenance documentation must identify the source study, creation, ownership, privacy, consent, complete inventory of related files, study design, duration, critical measurement information, and the metadata. It should also provide information on data transmission or ingestion into NHLBI BioData Catalyst.

Ingestion Form (***in development)**

The minimal metadata model we describe below is based on the ingestion form, linked above. We created this minimal metadata model (MMM) using the following considerations:

- The MMM will contain primarily provenance-related and study-level terms.
- We consider what may be common to all studies coming into NHLBI BioData Catalyst, whether they be dbGaP studies, COVID trials, etc.
- We assume all studies will be registered through dbGaP and therefore receive a dbGaP study accession ID.
- We include terms that, at a high level, will be relevant from the perspective of NHLBI.
- We support the inclusion of certain terms from the perspective of FAIR.

Table 2. Study Provenance Metadata

Section	Term	Reasoning
Data Submitter Contact Information	NHLBI Point of Contact - First Name	Point of contact for the "receiver" of data in NHLBI. CONNECTS Suggestion: NHLBI point of contact, such as a project scientist. These may be preloaded. Over the years, the PI and/or submitter may move or retire. By anchoring to an NIH

		person, you can have an idea of the history of oversight.
	NHLBI Point of Contact - Last Name	See above
	NHLBI Point of Contact - Email	See above
	Submitter First Name	Point of contact for the generator of data. Likely to be the study PI. Provenance: FAIR Principle R1.2
	Submitter Last Name	Point of contact for the generator of data. Provenance: FAIR Principle R1.2
	Submitter Email	Point of contact for the generator of data. Provenance: FAIR Principle R1.2
	Submitter Organization	What organization the data submitter is represented by, e.g., the Coordinating Center. Provenance: FAIR Principle R1.2
	Date of Submission	Provenance: FAIR Principle R1.2
Study Information	Submitted Study ID (Submitted GUID)	FAIR Principle F1
	dbGaP Accession Number (GUID)	FAIR Principle F1 - we assume all studies will be registered in dbGaP and thus assigned a dbGaP Accession Number
	Study Principal Investigator - First Name	Provenance: FAIR Principle R1.2
	Study Principal Investigator - Last Name	Provenance: FAIR Principle R1.2
	Study Principal Investigator - Email	Provenance: FAIR Principle R1.2

	Study Title	Relates to Findability; necessary to use for the dbGaP study configuration file
	NHLBI (Heart, Lung, Blood, Sleep) Domain	Relates to Findability; NHLBI may desire to perform faceting based on the NHLBI domain, including extending with MeSH and/or UMLS terminology
	Medical Condition(s)/Phenotype	Relates to Findability; must be aligned with the dbGaP list of phenotypes; uses the MeSH CV
	Study Type	Relates to Findability and is necessary for a faceted search. This will be generic enough to include dbGaP study types and clinical trials (and perhaps "other," which meets no designation)
	Sample Size	This is necessary for faceted searches and telling potential users what can be done with the study, given the size
	Data Collection Dates	A list of dates over which data was collected; likely important for provenance
Data Profile	GUID	The GUID created by the indexing entity, e.g., Gen3
	Submitted File GUID	The GUID created by the submitter; FAIR Principle F1
	File Checksum (md5/sha256)	What Gen3 uses for indexing; FAIR Principle F1
	Size in Bytes	Basic file-level metadata used for reporting
	Primary Data Location	URL of the file; likely in a cloud storage system; FAIR Principles F and A

	Consent Group(s)	Assuming different consent groups for different files under a single study, which may or may not be accurate. If not, this can be moved to Study Information.
	Data Type	Type of data, e.g., EHR Data, phenotype data, WGS, RNA-Seq, etc. Can possibly use EDAM ontology
	Data Format	Data format, e.g. CSV, TSV, BAM, CRAM, etc. Can possibly use EDAM ontology.
	File Name	Name of the file; string

Figure 3: Preliminary Diagram of the Metadata Model

The rest of the ingestion checklist is considered “nice to have” and is included in the appendix.

7. Additional Metadata: Data Documentation

Metadata is data that provides descriptions of the data, structure of the data, or contains administrative information about the data. Descriptive metadata is provided as unstructured files about the studies or source information. The structural and administrative data can be found as stated above, in the manifest files, created at the ingestion of the data into the Ecosystem.

If your dataset has any supporting metadata, such as case report form templates, data dictionaries, study protocols, and other documents, it is strongly encouraged to submit this along with your dataset.

8. Storage and Persistent Identifiers

The NHLBI BioData Catalyst systems and data are hosted in AWS & Google Cloud Platform's FedRAMP environments. To mitigate the risk of a breach or data leakage, multiple network firewalls and an intrusion-prevention and -detection system are in place to protect all communication in and out of the network, with a "default deny" rule that drops all traffic not explicitly permitted. All network traffic is monitored for malicious behavior and other specific attack signatures. In addition, within the cloud, data will be encrypted at rest (protecting data that's not moving through networks).

As part of the ingestion process, all files loaded to the AWS and Google cloud storage will have persistent identifiers created (GUIDs) as part of the indexing process. These identifiers will be available within the manifest files generated for study files.

For the NHLBI BioData Catalyst Ecosystem, the NHLBI Designated Authorizing Official has recognized the Authority to Operate (ATO) issued to the Broad Institute, University of Chicago, and Seven Bridges Genomics as presenting acceptable risk, and therefore the National Cancer Institute (NCI) ATO serves as an Interim Authority to Test (IATT) when used by designated TOPMed investigators and collaborators.

Quality assurance is a key element of NHLBI BioData Catalyst's backbone. As stated in the Data Management Strategy Document, Key Performance Indicators are in place to measure quality assurance with regards to data storage and other pertinent security and privacy features of NHLBI BioData Catalyst.

For reference, linked below are Genomic Data Security considerations for both AWS and GCP:

Architecting for Genomic Data Security and Compliance in AWS

Google Cloud Whitepaper: Handling genomic data in the cloud

9. Security and Privacy

NHLBI BioData Catalyst acts as a custodian of data produced or resulting from studies, cataloged and indexed as part of ingestion, and provided for use within one of many platforms for analysis. As such, the data managed within the system conforms to the NIST model of custodian management of data, inheriting privacy and consent from source systems, with the

responsibility of providing secure access to tools for analysis of the managed digital objects, and the availability to upload additional datasets for analysis.

NHLBI BioData Catalyst operates on Amazon Web Services (AWS) and Google Cloud Platform (GCP). Both AWS and GCP have received an Authority to Operate (ATO) from the General Services Administration FedRAMP (www.fedramp.gov) following a rigorous assessment process by a third-party assessor. The NHLBI Chief Information Officer has reviewed the System Security Plan from each system that comprises the NHLBI BioData Catalyst environment and has issued an ATO for their system to operate at the Moderate level. This type of authorization is consistent with the National Institute of Standards and Technology (NIST) guidance and complies with all requirements of the Federal Information Security Management Act.

The data within NHLBI BioData Catalyst is managed by NHLBI staff and contractors. All staff and contractors with access to the data hold a Public Trust Clearance that is based on an extensive background check. All activities, including data access, are logged and monitored.

10. Data Storage Considerations, Sharing and Archiving

10.1 Data Storage and Archiving

Due to the large data collection, archival of study information must be considered for efficient access of current information. Archival practices can either be manual or automatic. In the case of NHLBI BioData Catalyst study-specific information, a candidate for archive should be considered at two years past last access for any file, derived or original from the study in the collection, and be captured in offline storage as a complete set of manifest, data dictionary, clinical and non-clinical, raw, and analysis files. This collection should be available for search, and clearly designated as archived in offline storage for retrieval in the future.

10.2 File Types and Sizes Amounts/Size and Quantities Stored

NHLBI BioData Catalyst is able to ingest data types of all sizes, ranging from genomic and proteomic to clinical/phenotypic and imaging data. NHLBI BioData Catalyst currently has TOPMed dbGaP data measuring, in summation, in the single petabytes. The bulk of the size comes from genomic data in CRAM and VCF formats. This data comes from 70,000+ subjects and is complemented by over 240,000 phenotypic variables. NHLBI BioData Catalyst expects to accommodate data, phenotypic variables, and subjects from many more external studies in the near future.

Currently, genomic data is managed through the TOPMed IRC process in data freezes. Read alignments based on build GRCh38 and genotype call sets for Phase 1 studies from freeze 5 are

created and these files are in CRAM and VCF formats. All genotype data files from a given freeze contain the same set of variant sites. Site quality filtering information is contained in an associated “sites only” genotype file. While the size is large (####) the data is defined. The phenotypic, and imaging data of interest may come ad hoc and have an inherent size and provenance variability.

NHLBI BioData Catalyst has ingested primarily TOPMed dbGaP data that in summation measures in the single petabytes. The bulk of the size comes from genomic data in CRAM and VCF formats. This data comes from 70,000+ subjects and is complemented by over 240,000 phenotypic variables. NHLBI BioData Catalyst expects to accommodate data, phenotypic variables, and subjects from many more external studies in the near future.

10.3 Data Versioning

[To be developed by DRMWG]

10.4 Cloud Storage Costs Considerations

Currently, cloud storage costs for data ingested by and residing in NHLBI BioData Catalyst are covered by NHLBI. Information regarding cloud analysis credits for NHLBI BioData Catalyst users is available on the NHLBI BioData Catalyst website. Cloud analysis costs may also be subsidized by the NIH STRIDES program credits.

11. Citing Data and Licensing

11.1 Licensing

The web-based NHLBI BioData Catalyst platform conforms to an Open Source model with documentation conforming to the Creative Commons CC-BY-4.0 license for collaborative open science. Licensing of proprietary tools on platforms within the NHLBI BioData Catalyst is controlled by end user agreements according to a Bring Your Own License model: purchasing of licenses for specific proprietary tools must be performed by users to the system, as the application launching platforms do not check for end user licenses.

11.2 Publication Guidelines/Attribution

NHLBI BioData Catalyst Publications Guidelines v3

If you wish to cite the NHLBI BioData Catalyst ecosystem in your research, please use the following citation:

BioData Catalyst Consortium. (2020). The NHLBI BioData Catalyst. Zenodo. http://doi.org/10.5281/zenodo.3822858

12. Copyright and Data Protections

12.1 Copyright

NHLBI BioData Catalyst applications and the website conform to copyright applicable for U.S. Federal government web applications according to https://www.usa.gov/government-works.

12.2 Data Protections

In addition to tools, applications, and workflows, NHLBI BioData Catalyst provides access to de-identified data according to NIH Data Sharing standards. Please reference the NHLBI BioData Catalyst Data Protection page for additional information: https://biodatacatalyst.nhlbi.nih.gov/data-protection/

13. Appendix

13.1 DRAFT Ingestion Form Provenance Metadata Model Tables

Included below is a draft model for what the NHLBI BioData Catalyst Ingestion Form *may* look like. An updated form will be communicated via the DRMWG for review and formal approval. The requirements for the planning and pre-ingestion phases of bringing data into the Ecosystem are under development and as such, will evolve over time.

Data Submitter Contact Information

Field	Format	Required	In MMM?	Question Type	Response Type
NHLBI Point of Contact	String	Y	Y	Text Box	Free Text
First Name	String	Y	Y	Text Box	Free Text
Last Name	String	Y	Y	Text Box	Free Text
Date (of submission)	Date	Y	Y	Date Box	Date

	(MM-DD-YYYY)				(MM-DD-YYYY)
Title	String	Y	N	Text Box	Free Text
Email	String	Y	Y	Text Box	Free Text
Organization	String	Y	Y	Text Box	Free Text
Address	String	Y	N	Text Box	Free Text
Phone	Numeric (1111111111)	Y	N	Text Box	Numeric (1111111111)
eRA Commons/NIH NED Username	String	Y	N	Text Box	Free Text

Study Information

Field	Format	Required	In MMM?	Question Type	Response Type
Lead Investigator	String	Y	Y	Text Box	Free Text
PI Email	String	Y	Y	Text Box	Free Text
Co-Investigator(s)	String	N	N	Text Box	Free Text
Co-Investigator(s) Email(s)	String	N	N	Text Box	Free Text
Co-Investigator(s) eRA Commons/NED Username (if available)	String	N	N	Text Box	Free Text
Study Title (Align with dbGaP Study Config File)	String	Y	Y	Text Box	Free Text
Study Title Short Name	String	N	N	Text Box	Free Text
Medical Condition(s)	String	Y	Y	Text Box	Free Text

NHLBI Domain	Drop-Down	Y	Y	Single Choice	Predefined Values (Heart, Lung, Blood, Sleep)
Disease Sub-Domain	String	N	N	Text Box	Free Text
NHLBI Intramural Study	Boolean	Y	N	Checkbox	Yes/No
NHLBI Extramural Study	Boolean	Y	N	Checkbox	Yes/No
Study Co-Sponsor	String	N	N	Text Box	Free Text
Study Type	Drop-Down	Y	Y	Multiple Choice	Pre-Defined Values (Study type + dbGaP values)
Sample Size	Numeric String	Y	Y	Text Box	Free Text - Numeric
Data Collection Dates	Date String	Y	Y	Text Box	Free Text - Date Range
Study Location(s) (Geographic)	String	Y	N	Text Box	Free Text
Multi-Center Study	Boolean	Y	N	Checkbox	Yes/No
Grant/Contract # (Extramural Studies Only)	String	N	N	Text Box	Free Text
Grant Program Officer	String	N	N	Text Box	Free Text
DIR Project/Collaboration Agreement # (Intramural Studies Only)	String	N	N	Text Box	Free Text
Administrative	String	N	N	Text Box	Free Text

Officer					
Related Studies	Boolean	Y	N		
dbGaP Action #(s)			N		
dbGaP Study Title(s)			N		
Publication Submission (If Y, attach documentation)	Boolean		N		
ClinicalTrails.gov Link	String	N	N	Text Box	Free Text
Study Website	String	N	N	Text Box	Free Text
Biospecimens available?	Boolean	N	N	Checkbox	Yes/No

Data Profile

Field	Format	Required	In MMM?	Question Type	Response Type
Data Type	Drop-Down	Y	Y	Single Choice	Predefined Values
File Format	Drop-Down	Y	Y	Single Choice	Predefined Values
Result Type	Radio Button	Y	N	Single Choice	Predefined Values (Preliminary, Interim, Final)
Anonymized Data	Boolean	Y	N	Checkbox	Yes/No
Controlled Access Data?	Boolean	Y	N	Checkbox	Yes/No
If Yes, Explain (provide consent code(s))			Y		

IRB for Secondary Use?	Boolean	Y	N	Checkbox	Yes/No
Continuous Review Required	Boolean	Y	N	Checkbox	Yes/No
If Yes, Provide Justification	String	N	N	Text Box	Free Text
Metadata Available?	Boolean	Y	N	Checkbox	Yes/No
Reference Ontologies (please list)	String	Y	N	Text Box	Free Text
Check all available and attach if 'Yes':			N		
Manifest	Boolean	Y	N	Checkbox	Yes/No
Data Dictionary	Boolean	Y	N	Checkbox	Yes/No
API	Boolean	Y	N	Checkbox	Yes/No
Study Protocol	Boolean	Y	N	Checkbox	Yes/No
Is data harmonized according to specified standards?	Boolean	Y	N	Checkbox	Yes/No
If Yes, are standards used?	Boolean	N	N	Checkbox	Yes/No
If Yes, describe standards below:	String	N	N	Text Box	Free Text
Target Data Delivery Date	Date (MM-DD-YYYY)	Y	N	Date Box	Date (MM-DD-YYYY)
Target Public Release Date	Date (MM-DD-YYYY)	Y	N	Date Box	Date (MM-DD-YYYY)
Primary Data Location	String	Y	Y	Text Box	Free Text

Secondary Data Location(s)	String	N	N	Text Box	Free Text
Keywords	String	Y	N	Text Box	Free Text

Genotype Platform Information

Field	Format	Required	In MMM?	Question Type	Response Type
Name	Drop-Down	Y	N	Single Choice	Predefined Values
Vendor	Drop-Down	Y	N	Single Choice	Predefined Values
Number of Probes	Numeric String	Y	N	Text Box	Numeric
URL	String	Y	N	Text Box	Free Text
Description	String	Y	N	Text Box	Free Text
Submitting SRA?	Boolean	Y	N	Checkbox	Yes/No
If Yes, follow dbGaP instructions	String	N	N	Text Box	Free Text

13.2 Key Terms and Concepts

The NHLBI BioData Catalyst Glossary is located at the following: https://bdcatalyst.gitbook.io/biodata-catalyst-consortium-guidance/glossary/biodata-catalyst-consortium-glossary

The following includes specific terms referenced within this document:

Audit: Process of systematic examination of the system determined by the quality assurance process.

AWS: Amazon **W**eb **S**ervices

CCB: Change **C**ontrol **B**oard within the NHLBI BioData Catalyst Consortium

Cloud: Storing and accessing data and programs in remote servers hosted on the internet instead of on local computing systems. NHLBI BioData Catalyst duplicates data across the Amazon Web Services and Google Cloud Platform for data storage.

Data: Includes all digitized information, data resources, derived data products, and results of digital extract from one store, transformation, and/or load of this digital information within the NLBI BioData Catalyst Ecosystem or a NHLBI BioData Catalyst platform.

Data Index: A data index is a unique identifier, created to allow future discovery of the information and/or metadata. For NHLBI BioData Catalyst, the Global Unique Identifiers (GUIDs) created at ingest for each object file serves as this unique identifier, added to the manifest file for all data files ingested.

Data Ingestion: The process of obtaining and importing data for availability across the NHLBI BioData Catalyst Ecosystem. Data can be streamed in real time or ingested in batches. For data contributions for individual use, see *Data Upload*.

Data Curation: The organization and integration of data from various sources including the processes defined for the receipt, transfer, accounting, safeguarding, and destruction of material within the purview of the Ecosystem.

Data Custodian: All personnel who have operational responsibility for the data, especially NHLBI BioData Catalyst stakeholders. A collection of data may have multiple data custodians.

Data Governance: An organizational strategy to support business goals.

Data Management: Describes how data as an asset is operationalized and used to support an organizational strategy.

Data Management Core: Ensures data is appropriately represented within the Ecosystem.

Data Manifest: A manifest lists the contents and often location of items. Within NHLBI BioData Catalyst, the data manifest is a file containing data about the files ingested (md5, file size) as well as Access Control Lists, file name(s), and the URL for the deposition bucket.

Data Owner: An individual who is accountable for the data in a legal or business sense. The data owner is the executive or senior staff member who (1) answers for the proper care of the data by all within the organization who have access to or control of the data; and (2) makes decisions about the dataset, system, or resource.

Data Steward: An individual or group who is responsible for the contents or values of the data, especially quality control and assurance. Data stewards may define business rules that apply to the data under their supervision.

Data Upload: Moving data from outside the NHLBI BioData Catalyst security boundary into a user accessible working location within the security boundary, i.e., Bring Your Own Data (BYOD).

DAWG: **D**ata **A**ccess **W**orking **G**roup within the NHLBI BioData Catalyst Consortium

DCF: Data **C**ommons **F**ramework

DCFS: Data **C**ommons **F**ramework **S**ervice

DevOps: Set of practices that combines software development (Dev) and information-technology operations (Ops), which aims to shorten the systems development life cycle and provide continuous delivery with high software quality.

DHWG: Data **H**armonization **W**orking **G**roup within the NHLBI BioData Catalyst Consortium

DRMWG: Data **R**elease **M**anagement **W**orking **G**roup within the NHLBI BioData Catalyst Consortium

Ecosystem: A software ecosystem is a collection of processes that execute on a shared platform(s) or across shared protocols to provide flexible services. For example, the NHLBI BioData Catalyst Ecosystem is inclusive of all platforms, tools, applications, data, and workflows enabling research investigators to find, access, share, store, and compute on large scale datasets in a secure workspace.

Element Team DevOps: Individual Other Transaction Awards (OTAs) led by a Principal Investigator (PI), or PIs, who will complete milestones and produce deliverables.

FAIR Principles: A set of guiding principles, including the creation of metadata, to make data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (Wilkinson et al. 2016).

Key Performance Indicator: A quantifiable measure used to evaluate the success of in meeting objectives for performance.

GCP: **G**oogle **C**loud **P**latform

Metadata: Data about data that describes the properties of a dataset and is critical to supporting discoverability (the "F" in FAIR - Findable).

NHLBI: **N**ational **H**eart, **L**ung, and **B**lood **I**nstitute

OTA: Other **T**ransaction **A**wards

PL: Program Leadership

Platform: A Data Commons serving as user-accessible applications and application programming interface comprising the NHLBI BioData Catalyst Ecosystem. Examples: Terra, Gen3, Seven Bridges Genomics, etc.

Quality Assurance: The maintenance of a desired level of quality in a service or product, especially by means of attention to every stage of the process of delivery or production.

Quality Control: A system of maintaining standards by testing a sample of the output against the specification.

STRIDES Initiative: NIH **S**cience and **T**echnology **R**esearch **I**nfrastructure for **D**iscovery, **E**xperimentation, and **S**ustainability (STRIDES) Initiative allows NIH to explore the use of cloud environments to streamline NIH data use by partnering with commercial providers (https://datascience.nih.gov/strides).