Unlocking Insights:
Seamless Data Ingestion, Harmonization, and Curation for Multiomics Datasets with Integrated Ontologies

Yasodha Kannan S, Subuhi Yete, Hima Varghese, Manjushree G Sahoo, Suraj Kumar Sharma, Swaraj Basu

Strand Life Sciences, Bangalore, India

strand
WE SOLVE PROBLEMS IN GENOMICS.

Conference Name
Bio-IT World Conference & Expo 2024
Dates - April 15th to 17th
Place - Boston, MA, USA

Contact
Yasodha Kannan S
Bioinformatics Engineer
+91 95855 16367
yasodha@strandls.com

Introduction

In multiomics research, gaining valuable outcomes hinges on the **systematic management** of **diverse datasets** while adhering to **FAIR** data principles.

The process involves **data ingestion, harmonization**, and **curation**, crucial for data analysis.

Integrating ontologies is important for maintaining **consistent** and **controlled vocabularies** across datasets.

This study focuses on the **management of multiomics datasets**, from initial ingestion to harmonization and curation, while **preserving data integrity.**

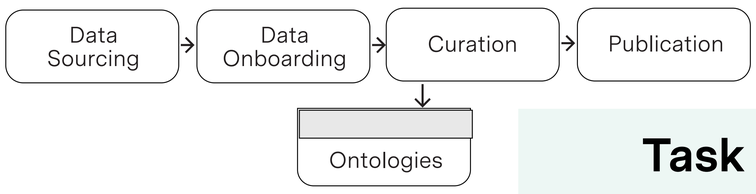


Figure 1: Overview of Data Harmonization Workflow

Highlights

Gene model improvements, achieved by transforming 1000+ datasets to suit a particular model and thus improve interoperability.

Assessment of **data availability** for model building involving: (i) sourcing **scRNA** and **scATAC** data from public databases (ii) extensive metadata curation, in order to enable dataset selection for pipeline development and modeling.

Strand's expertise in **curation** helps in evaluating tools based on **LLM models for omics meta-data mining.**

In-depth knowledge of input data characteristics, quality and relevance is essential for **effective model rebuilding** - e.g. Gene prediction model rebuilding for **Cell Painting datasets.**

Controlled Vocabulary

1. **Standardization:** Employ standardized terms to systematically organize and manage descriptive data, enhancing quality and interoperability.
2. **Metadata Schema:** Develop a metadata schema aligned with specific data requirements to ensure effective organization and retrieval.
3. **Naming Conventions:** Establish and enforce naming conventions for metadata elements to maintain consistency and facilitate data integration.
4. **Clear Definitions:** Provide clear definitions and guidance for each controlled vocabulary term to reduce ambiguity and enhance understanding.
5. **Training:** Offer training for metadata creators and users on the use of controlled vocabularies to optimize data management practices and promote effective utilization.

Data Sourcing

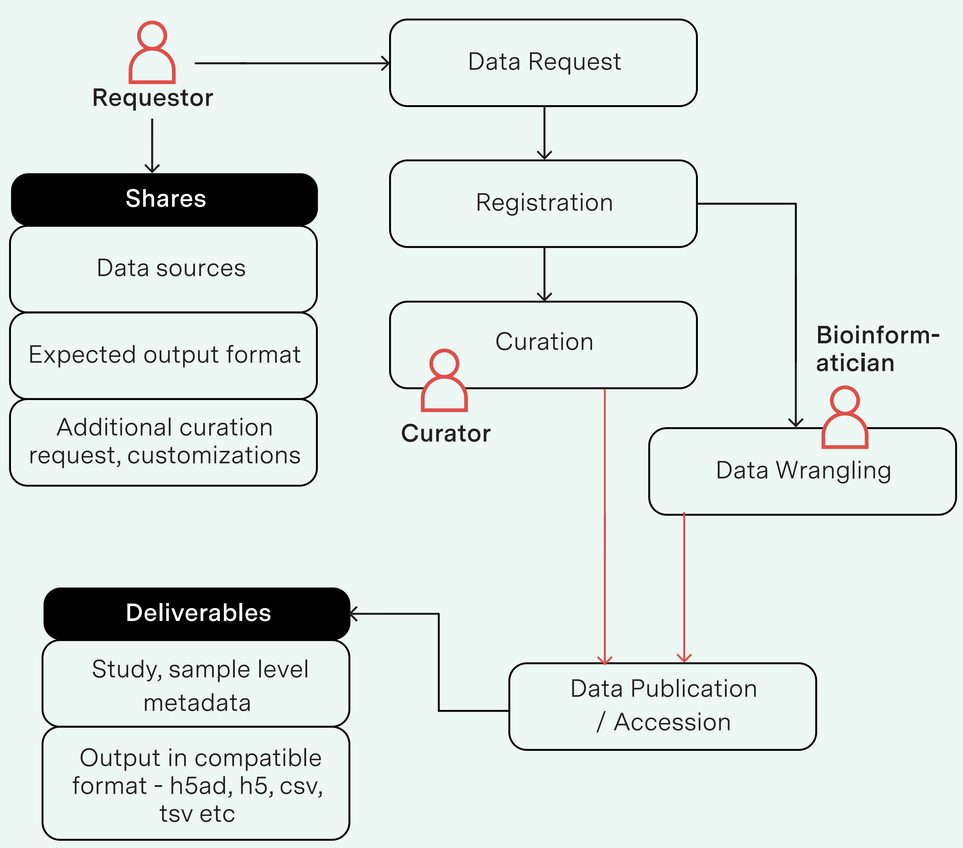
For multiomics datasets, the data to be processed fall into 5 major categories of sources

1. Big public databases - GEO, SRA, ENA
2. Published papers and their associated supplementaries
3. Data repositories on cloud, data transfer tools - AWS S3 buckets Gcloud buckets, Globus, imaging datasets
4. Custom databases - CellxGene, HCA, Archs4, refine.bio, recount3
5. Inhouse generated data by lab or collaborators, CROs

Task Management System

1. For prioritization of requests from multiple users.
2. To keep track of tasks, workload and deadlines.
3. Appropriate reviews and checks at multiple stages.

Figure 2: Harmonization Tasks Pipeline



Achievements

Metadata harmonization organized data, enabling downstream ML workflows.

We harmonized **35+ datasets** across various disease conditions. This approach helped in creating an **integrated atlas** and a **time-series model** for studying **cell-cell interactions** and understanding the cell types involved.

Ongoing curation reduces turnaround time for data ingestion into the data lake, enabling downstream ML.

Our typical TAT for 5-10 datasets is about **5 days**, while bulk curation involving **50+ datasets** is completed within **2-3 weeks.** The number of metadata fields involved in curation ranges between **10-60** depending on the available data and client specifications. Key time-consuming fields, in descending order, are **cell type, cell line, tissue, assay, disease, biological sex, and age**. Curating these fields requires extending beyond the dataset information and referring to **multiple related sources**, including publications, supplementary files, and processed data files, and mapping the fields to ontology Additionally, we're experimenting with **LLM models** to automate ontology mappings and are currently working on a proof of concept.

Tech Stack

Software

Public toolkits

Inhouse servers, Baremetal setup, VMs, AWS EC2

Experts

Curators, Bioinformaticians, Data Scientists, Software Engineers

Data Onboarding

1. Preparing and integrating data from various sources into a unified and accessible format for analysis or use in a specific system
2. Upload the final format to cloud or identifed location with open or restricted access according to team
3. Document the entire processing with Readmes and notes

Harmonized Datasets Publication

1. Fostering collaboration and innovation among data users across different domains.
2. Easily integrated with other datasets or systems.
3. Adhere to interoperability standards, such as those defined by international organizations or industry consortia.
4. Version control mechanisms to track changes to the dataset over time. This helps users understand the evolution of the data and ensures reproducibility.