Research Data Description From Structures Within and Around Metadata

Julianna Pakstis
Children's Hospital of Philadelphia, United States
pakstisj@chop.edu

Christiana Dobrzynski
Illumina, United States
cdobrzynski@illumina.com

Perry Evans
Spark Therapeutics, United States
James.Evans@sparktx.com

Stephanie Huang
Children's Hospital of Philadelphia, United States
huangs4@chop.edu

Ene Belleh
Children's Hospital of Philadelphia, United States
bellehe@chop.edu

Allison Olsen
Children's Hospital of Philadelphia, United States
olsenar@chop.edu

Hannah Calkins
Children's Hospital of Philadelphia, United States
calkinsh@chop.edu

Abstract

The Arcus Archives at the Children's Hospital of Philadelphia (CHOP) aims to collect and describe research data from across the Research Institute. To accomplish this, the Arcus Library Science team has established processes, structures, and descriptive metadata schemas informed by all phases of the research data preservation and reuse lifecycle at CHOP including archiving, cataloging, display, and usability. This paper introduces the Arcus Archives metadata schema and the key structures it relies upon.

Specifically, this paper explains how a small set of dataTypes subfields within a hierarchical, archivally informed metadata structure utilize a shared file directory organization structure to thoroughly and accurately process, catalog, and surface for discovery metadata about petabytes of archived pediatric research data from multiple and growing data modalities.

The metadata schema itself is flexible but consistent enough to apply to a myriad of data types produced during the conduct of pediatric research. Each metadata record reflects the unique archival arrangement done for each collection. It encompasses the framework of a shared recommended "project template" file directory structure which includes manifest and protocol files that allow for meaningful capture and organization of numerous complex data files and their relationships to one another.

Keywords: metadata; research data; file structure; pediatrics; implementation

1. Introduction to Arcus and the Library Science Team

Arcus is a multi-year strategic initiative of the Research Institute at the Children's Hospital of Philadelphia (CHOP). According to the Department of Biomedical and Health Informatics (2023),

> Arcus is a suite of tools and services developed to enhance research efforts by helping researchers to explore available data, see overlaps among datasets, build new cohorts, and determine if there are data or samples available for additional research projects. Incubated within the Department of Biomedical and Health Informatics (DBHi) at CHOP, Arcus connects CHOP's clinical and research data to enable biomedical researchers to conduct highly innovative, data-driven, reproducible research within a managed scalable framework. This framework includes 1) user access controls; 2) patient privacy and

https://doi.org/10.23106/dcmi.953336477

confidentiality protections through regulatory review; 3) electronic honest-brokered data de-identification and re-identification; and 4) data retention, management, sharing, and destruction services in an auditable computational environment.

The initiative was established as a practical pathway for promoting the ethical sharing of data among researchers. The Arcus environment provides a secure data Archives run by librarians and archivists who collaborate with software developers, platform engineers, privacy experts, educators, and research teams themselves in order to link research data contributions of any kind (such as genetic, biobanking, and survey data) with clinical data on more than two million CHOP patients, including information on procedures, diagnoses, and imaging. All CHOP researchers with valid credentials have access to the Arcus Archives.

Arcus' extensive internal scope was created to assist researchers in overcoming the siloing of data produced during clinical care or course of study. There are several reasons why scientists or doctors might not be willing to share data. Problems, including pressure to publish (Rawat, 2014) and fear of data scooping (Laine, 2017), occur both within institutions and out to the broader scientific research community. Even while the scientific community and individual researchers recognize the advantages of having open and shareable data (Lane, 2007; Mesirov, 2019; Peng, 2011; Stodden, 2016; Mendes, 2018), resistance remains. According to the Roundtable on Environmental Health Sciences, Research, and Medicine et al. (2016), reproducible research increases scientific rigor, scientific trust in the data, and awareness of related studies. At CHOP, Arcus advances the objective of repeatable and reusable research by enhancing data access across areas within the institution through its role as the informatics and Archives center. To make data available for reuse known, the Arcus Library Science team uses metadata and an upcoming data catalog.

Researchers do not always have the organizational, data management, or descriptive abilities to disseminate their work effectively, even if they desire to do so. By providing tools and strategies to aid researchers in better stewarding their own data from the project's inception through study conduct, publishing, and archiving, the Research Data Management services provided by the Arcus Library Science team try to close these gaps. A standard recommended file directory layout called the project template is crucial for better data management. Given the structure is the same for all submissions to the Archives, metadata automation and arrangement are possible even for very large collections.

Metadata is essential for an in-depth program like Arcus to document the institution's research data, where and how it was created, and how it has been used. Arcus' reach and its aspirational objectives of promoting data sharing and encouraging cultural shifts toward reproducibility have created numerous chances to develop a distinctive yet interoperable metadata ecosystem.

2. Metadata in the Arcus Archives

Metadata in the Arcus Archives is subject to the same inherent tension that exists between preserving the relationship information of archival arrangement (Miller, 1990) while surfacing the item-level granularity present in descriptive metadata that bolsters discoverability. The difficulty in reconciling this tension had often meant that archival arrangement is described through an Encoded Archival Description (EAD) finding aid whereas item-level descriptive metadata has been implemented in something like a Machine-Readable Cataloging (MARC) record.

https://doi.org/10.23106/dcmi.953336477

Arcus descriptive metadata must be able to provide the full context of a file being described by presenting its relationship to other files and/or series in the collection. In traditional library cataloging techniques, this is achieved by linking title, identifier, or both (OCLC, 2018). This type of reference does not offer a method for capturing structured hierarchical relationships, like those between a collection and its series, that are essential in archival arrangement and prescribed in the Arcus Archives. Thus, Arcus metadata must reflect archival finding aid structure to describe its collections.

Archival finding aid structures alone are not sufficient for metadata in the Arcus Archives. Within the landscape of datasets and research files, users need a high level of differentiation between objects (Riley, 2017) and the people they contain information about. This is especially important in medical research data where tracking the files in which protected health information (PHI) or potential-PHI may appear directly impacts research users' ability to uphold the required privacy and confidentiality standards, such as components of Health Insurance Portability and Accountability Act (HIPAA). Finding aids do not represent enough granularity about individual items or their contents to meet this need. Thus, Arcus metadata also needs to draw on practices from bibliographic description to enhance the finding aid's structural sketch of item grouping relationships.

At the outset of the program, Arcus Metadata Librarians designed metadata in the Arcus Archives to handle this tension and hold both archival relationship details and granular discoverability information at once. Research objects in the Arcus Archives are grouped together into collections that contain diverse assets from across the research lifecycle. This includes datasets, certainly, but also archives tools and scripts, administrative files, project documentation, and contextual information. Comprehensive description of a collection of this scope, then, requires a metadata schema that blends both library and archival description.

Arcus descriptive metadata meshes bibliographic and archival description by using the JSON format. The structured nature of JSON coupled with its nesting abilities allows us to include relational information within a single collection's metadata. The prevalence of JSON structured data and JSON based tools in web and systems development also helps us more finely integrate with the software developers on our team. This integration allows for librarians and developers to work together to create tools for automation that make cataloging vast amounts of data possible.

At its core, the Arcus Archives metadata schema is a custom implementation of existing structures together with local fields and vocabularies, all written in JSON. Though unique in structure, the descriptive metadata schema incorporates shared fields and vocabularies from major data repositories. The schema adapts fields from version 2.2 of the Data Tag Suite (DATS) (Gonzalez-Beltran, 2017) and from the DataCite Metadata Working Group's 2017 4.1 version of the DataCite schema. The Arcus schema also draws from available controlled vocabularies, with terms from the Data Use Ontology (DUO) (Lawson, 2021) to describe reuse allowances and from Medical Subject Headings (MeSH) to apply topics. These adaptations ensure the schema is interoperable wherever possible with industry standards as well as those from constituent fields of biomedical research. It also allows contributing users to comply with data sharing mandates, such as that passed by the National Institutes of Health (NIH) in 2023.

2.1. Hierarchical Metadata Structure

The Arcus Archives metadata schema prescribes three levels of description as part of each record: collection, series, and file. The levels of description are modeled on archival finding aids and informed by the archival arrangements done for each contribution to Arcus. Each level contains fields designed to fully and specifically describe the objects or groupings at that level. When

https://doi.org/10.23106/dcmi.953336477

considered together as a single record, the three levels provide both specificity and context for a given object in Arcus. By grouping specific fields into hierarchical levels, we have adapted archival and bibliographic methods to fit the needs of Arcus materials and expand the current state of scientific research dataset description.

Collection is the broadest level of metadata. Metadata at the collection level should be representative of all items contained in lower levels. It provides an overview of the entire archival collection. Generally, each archival collection represents the data from one and only one research project.

Series is an intermediary level that exists below collection. All series must belong to a collection, and all collections have at least one series. Series allow for grouping of related materials within a collection. Series can also be broken down into multiple subseries if necessary. For example, a series may hold all the research data for a particular contribution. If the contribution is a simple one, the research data may not be further subdivided. However, if the contribution is a complex one that produced multiple sets of data from distinct efforts, it may be necessary to include multiple subseries underneath a parent research data series to represent these distinctions. Individual series and subseries may have child elements that are either subseries or files, but not both.

Files are the most granular level of metadata. Each file will have its own distinct metadata, much of it technical (such as file type, size, and URL). Files are aggregated in a series or subseries.

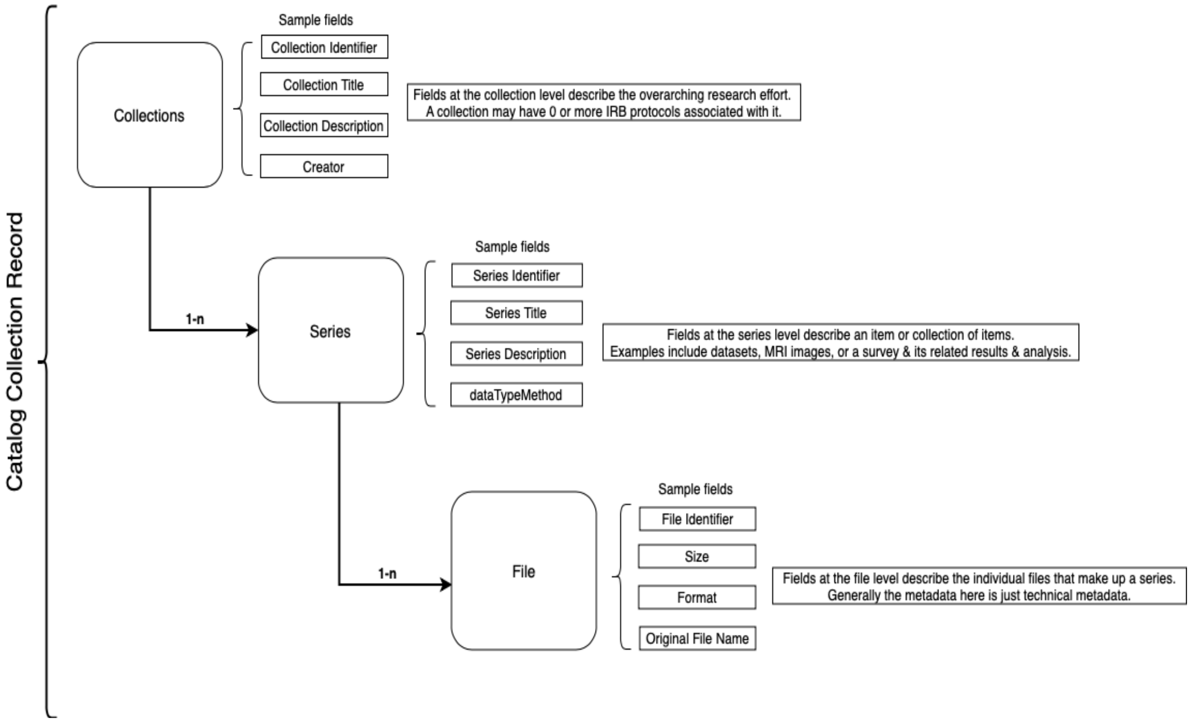

Catalog Collection Record
Collections
Sample fields
Collection Identifier
Collection Title
Collection Description
Creator
Fields at the collection level describe the overarching research effort.
A collection may have 0 or more IRB protocols associated with it.
1-n
Series
Sample fields
Series Identifier
Series Title
Series Description
dataTypeMethod
Fields at the series level describe an item or collection of items.
Examples include datasets, MRI images, or a survey & its related results & analysis.
1-n
File
Sample fields
File Identifier
Size
Format
Original File Name
Fields at the file level describe the individual files that make up a series.
Generally the metadata here is just technical metadata.

FIG. 1. Diagram of hierarchical structure of Arcus metadata.

2.2. Structured Yet Flexible Organization

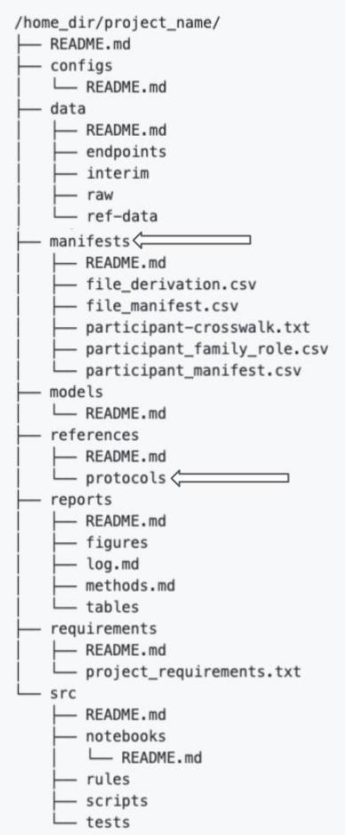
Project template, manifests

To facilitate the ingestion of large amounts of archival data and its cataloging, the Arcus Library Science team requires a high level of organization internally. Cooperative organization from our

https://doi.org/10.23106/dcmi.953336477

data contributors on the research side is essential as well. The cornerstone of archival organization is the project template (Dobrzynski et al., 2019), a basic recommended file directory structure. It was adapted from Cookiecutter Data Science (DrivenData, 2021) for use with CHOP research data by former CHOP Bioinformatics Scientist Perry Evans and former CHOP Digital Archivist Christiana Dobrzynski. The project template is designed to be broad enough to encompass the varied research endeavors across CHOP while remaining comprehensive and consistent enough to create a throughline of categorization among those diverse efforts.

The project template structure is integrated throughout the Arcus Library Science team's initiatives. In research data consultations, the Arcus Library Science team trains researchers to use the project template for organizing their research files. Every archival collection Arcus intakes is arranged using the project template. The project template is furnished in each of the computational labs Arcus provides to researchers. All archival data is delivered to the labs using the structure. Broad usage of the project template enables us to connect datasets and funnel research efforts through the initiation, conduct, and archiving of their study with ease.



```
/home_dir/project_name/
├── README.md
├── configs
│   └── README.md
├── data
│   ├── README.md
│   ├── endpoints
│   ├── interim
│   ├── raw
│   └── ref-data
├── manifests
│   ├── README.md
│   ├── file_derivation.csv
│   ├── file_manifest.csv
│   ├── participant-crosswalk.txt
│   ├── participant_family_role.csv
│   └── participant_manifest.csv
├── models
│   └── README.md
├── references
│   ├── README.md
│   └── protocols
├── reports
│   ├── README.md
│   ├── figures
│   ├── log.md
│   ├── methods.md
│   └── tables
├── requirements
│   ├── README.md
│   └── project_requirements.txt
└── src
    ├── README.md
    ├── notebooks
    │   └── README.md
    ├── rules
    ├── scripts
    └── tests
```

FIG. 2. Project template file directory structure with manifests and protocol storage locations called out.

Our custom Metadata Management System (MMS) automates creation of the metadata record structure based on the archival arrangement informed by the project template framework. Within this framework, protocol files, standardized yml files that contain information about how the data was produced, are always stored in a consistent directory location (`/home_dir/project_name/references/protocols`). Manifests are always stored in a consistent directory location (`/home_dir/project_name/manifests`). The `file-manifest.csv` file is always stored in the manifests/ directory. This manifest lists out all the files, links them to participants in a deidentified fashion, and declares which protocol file holds information about the file in question. The explicit information from the protocol file and the

https://doi.org/10.23106/dcmi.953336477

implicit information about how the methods described in the protocol file are enacted come together in the dataTypes set of fields in the Arcus metadata schema. The file manifest and protocol files make describing numerous files (on the order of tens of thousands) feasible by, minimally, one Metadata Librarian with the help of one Data Liaison embedded in the research team providing the source information.

2.3. Metadata From Structure: dataTypes Fields

One of the strongest examples of the integration between the project template, its manifests, and the information captured in the Arcus Archives schema is the dataTypes set of fields (TABLE 1). This set of fields exists at the series level within the overall structure of the collection records. dataTypes is an array consisting of one or more objects made up of a number of subfields that capture input, process, and "is-ness" of the data. This set of fields was originally adapted from the Data Tag Suite (DATS) schema v.2.2 (Gonzalez-Beltran, 2017). The relationship between the dataTypes fields allows us to describe the whole of the research dataset alongside the methods of analyses. This is particularly important and novel for describing genomic data and methods of sequencing and analysis. Crucially, however, the dataTypes fields are extensible enough to describe not only genomic data but imaging, tabular data, contextual files, tools, and more.

TABLE 1: dataType subfields and their descriptions.

dataType Subfield Name	Description	Required	Repeat -able	Type
dataTypeName	The typology of the dataset, identifying the dataset type or nature of the data.	Y	Y	string
dataTypeInformation	The measurements or facts that the data is about.	N	Y	string in array
dataTypeRefinement	A qualifier to describe the processing level of the dataset and its distributions.	Y	Y	string in array
dataTypeMethod	The procedure or technology used to generate the information.	Y	Y	string in array
dataTypePlatform	The set of instruments, software, and reagents that are needed to generate the data.	N	Y	string in array
dataTypeSourceValue	Source of the data.	N	Y	string in array
dataTypeSourceNote	Additional free text information about the source of the data.	N	Y	string in array

dataTypes in Action: Genomics Example

Take the example of describing genomic research data in a study of the gut microbiome[1]. To fully describe this research data, we must describe both the raw and processed data. The raw data was generated directly from a biological sample using a genomic sequencing technique called the shotgun run on an ArcGene10 machine. The endpoint, or analyzed, data was processed using the Sunbeam pipeline to produce a metagenomic interpretation of "the microbial and genetic composition of samples" as described in Clarke et al., 2019.

[1] This example is drawn from real data stored in the CHOP Arcus Archives. Specific values for the platform_name, instrument_model, sequencing_center, and dates in the example, however, have been changed to protect participant privacy and institutional information. The type of use case and cited software remain unchanged.

https://doi.org/10.23106/dcmi.953336477

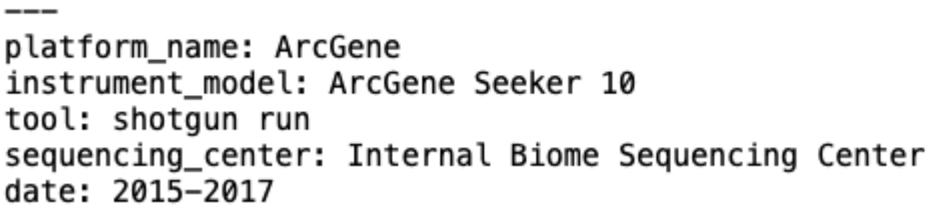
The `file-manifest.csv` (TABLE 2, column C) specifies which protocol file (examples in Figures 3 and 4) holds metadata about the genomic sequencing for a given biological sample. The first two rows show metadata about the raw multiplexed paired (meaning there were two reads from the sequencer: Read 1 and Read 2, denoted by the r1 and r2 appended to the file_type) fastq data produced from a biological sample. The last two rows show endpoint demultiplexed paired fastq data resulting from the raw data created from the biological sample.

TABLE 2: Excerpt of file-manifest.csv showing protocol file location for raw and analyzed sequencing data.

biosample_id	file_type	protocol	file_path	file_group
Case.StoolSample001	multiplexed-fastq_r1	references/protocols/fastq.yml	data/raw/multiplexed-fastq/Case.StoolSample001.fastq.gz	Case.StoolSample001_fqm
Case.StoolSample001	multiplexed-fastq_r2	references/protocols/fastq.yml	data/raw/multiplexed-fastq/Case.StoolSample001.fastq.gz	Case.StoolSample001_fqm
Case.StoolSample001	demultiplexed-fastq_r1	references/protocols/demultiplexing.yml	data/endpoints/demultiplexed-fastq/Case.StoolSample001.fastq.gz	Case.StoolSample001_fqd
Case.StoolSample001	demultiplexed-fastq_r2	references/protocols/demultiplexing.yml	data/endpoints/demultiplexed-fastq/Case.StoolSample001.fastq.gz	Case.StoolSample001_fqd

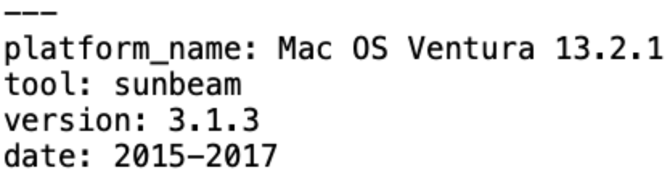
TABLE 3: file-manifest.csv data dictionary explaining column names for TABLE 2.

Column	Description
biosample_id	Unique identifier for the biological sample that produced the data being described in a given row of the manifest.
file_type	Free text description of the nature of the file.
protocol	File path to protocol yml file used to record and describe experiment metadata.
file_path	Path to the file being described in a given row of the manifest.
file_group	Formally related files. Examples include paired fastq files or a bam file and its index.



```
---
platform_name: ArcGene
instrument_model: ArcGene Seeker 10
tool: shotgun run
sequencing_center: Internal Biome Sequencing Center
date: 2015-2017
```

FIG. 3. `fastq.yml` protocol file showing raw sequencing metadata.



```
---
platform_name: Mac OS Ventura 13.2.1
tool: sunbeam
version: 3.1.3
date: 2015-2017
```

FIG. 4. `demultiplexing.yml` protocol file showing analyzed endpoints metadata.

https://doi.org/10.23106/dcmi.953336477

The Sunbeam analysis pipeline takes as its input multiplexed raw sequence data. Thus, we need to record metadata about these inputs. We also need to be able to link these input files and their metadata to the next processing steps. Because the entire set of dataTypes subfields is repeatable as an object within the dataTypes array, we can represent the raw data as one object and then the analyzed data as another object, linked together conceptually.

```
"dataTypes": [
 {
  "dataTypeInformation": ["genomics"],
  "dataTypeMethod": ["shotgun run"],
  "dataTypeName": "genomics",
  "dataTypePlatform": ["ArcGene Seeker 10"],
  "dataTypeRefinement": ["raw data"],
  "dataTypeSource": [
   {
    "dataTypeSourceValue": "biological sample"
   }
  ]
 },
 {
  "dataTypeInformation": ["metagenomics"],
  "dataTypeMethod": ["sunbeam 3.1.1"],
  "dataTypeName": "pipeline",
  "dataTypePlatform": ["Mac OS Ventura 13.2.1"],
  "dataTypeRefinement": ["analyzed data”],
  "dataTypeSource": [
   {
    "dataTypeSourceValue": "file storage"
   }
  ]
 }
]
```

FIG. 5. JSON dataTypes metadata for a genomics example containing both raw and analyzed data.

Translating this JSON back into human readable language, we return to the original description of the endpoint data and its constituent raw data: The raw data (dataTypeRefinement) was generated directly from a biological sample (dataTypeSourceValue) using the genomic (dataTypeInformation) sequencing technique called the shotgun run (dataTypeMethod) on an ArcGene Seeker 10 machine (dataTypePlatform). The analyzed (dataTypeRefinement) endpoint data consists of the raw dataset pulled from file storage (dataTypeSourceValue) processed using the Sunbeam 3.1.1 pipeline (dataTypeMethod) on a computer running Mac OS 13.2.1 (dataTypePlatform) to produce a metagenomic (dataTypeInformation) interpretation of “the microbial and genetic composition of samples.”

2.4. Controlled Vocabularies for dataTypes Values

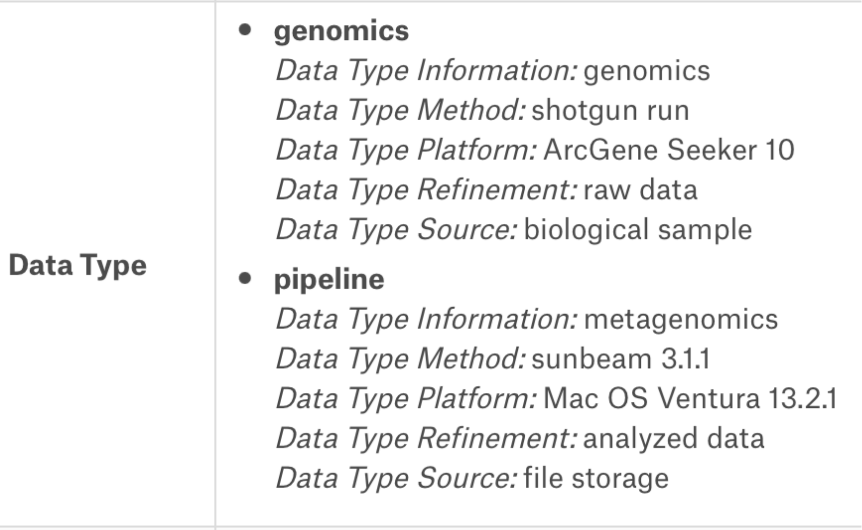
Because bioinformaticians, either within Arcus or on the study team, use non-standardized genomic files in which other scientists often supply various forms to refer to the same technologies, there is a proliferation of minimally different terms referring to the same concepts (Masseroli, 2019). This makes processing pipelines difficult to impossible to scale and makes description imprecise. Bioinformaticians need controlled vocabularies for the exact concepts the Arcus Metadata Librarian needs to complete the dataTypes fields. This presented an opportunity to move the controlled vocabulary standardization upstream and solve the problem for both the study and Library Science teams.

For the gut microbiome sequencing example, the platform_name options are standardized by a controlled vocabulary term to ensure all ArcGene products can be linked together. This prevents misspellings (e.g. ArcGen) or capitalization inconsistencies (e.g. Arcgene). The MMS tool validates records against the controlled vocabulary lists. It can also integrate and pull terms from external authoritative sources.

Though the Arcus metadata schema uses extensive controlled vocabularies for the dataTypes fields, we were able to pull out only the relevant snippets of the vocabularies into separate files reassembled using a JSON $ref key. Those files are then surfaced to bioinformaticians in a separate Github repository that also explains protocol file usage. The bioinformaticians can review these vocabulary files and select the relevant standardized terms. They can also request a new term is added if the concept they need to describe has not been included before. With this process, the Metadata Librarian can still create and control a comprehensive list of dataTypes terms as part of the full JSON schema but can do so without adding unnecessary complication to the work of the study team while still leveraging their expertise.

2.5. dataTypes: Usability and Discovery

Exactly how to display dataTypes to be immediately understandable and searchable is a question separate from how to structure the JSON metadata. In the first release of our internal data catalog, an enterprise-wide instance of an Alation product called "Gene," we will surface each of the dataTypes fields on their own line and with the JSON formatting removed. Descriptions of each of the dataTypes fields will be searchable in a linked metadata glossary similar to TABLE 1 here. The dataTypeName value serves as a header for grouping the constituent dataTypes values related to that name.

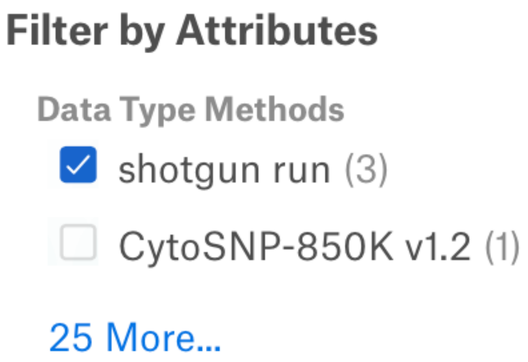


Data Type	**genomics** *Data Type Information:* genomics *Data Type Method:* shotgun run *Data Type Platform:* ArcGene Seeker 10 *Data Type Refinement:* raw data *Data Type Source:* biological sample **pipeline** *Data Type Information:* metagenomics *Data Type Method:* sunbeam 3.1.1 *Data Type Platform:* Mac OS Ventura 13.2.1 *Data Type Refinement:* analyzed data *Data Type Source:* file storage

FIG. 6. dataTypes metadata presented in the Gene catalog.

Each of the dataTypes fields will also be selectable values in a sidebar filter. For example, users will be able to select "shotgun run" from dataTypeMethod and view all datasets that have utilized this genomic sequencing type.

https://doi.org/10.23106/dcmi.953336477



Filter by Attributes

Data Type Methods

☑ shotgun run (3)

□ CytoSNP-850K v1.2 (1)

25 More...

FIG. 7. dataTypesMethod values as filters in the Gene catalog.

It remains to be seen whether glossing the dataTypes keys and increased socialization of the field name terms will be sufficient for user understanding. We remain open to the possibility of renaming these fields on display or showing them in a different order or visual presentation, for example, depending on the results of catalog user testing. Arcus has a User Experience (UX) Specialist embedded within the team who can perform such user testing to evaluate the efficacy of the presentation of the dataTypes fields as well as the efficacy of Arcus Archives descriptive metadata as a whole. Feedback and changes from user testing round out Arcus metadata ensuring it is informed by all phases of the research data preservation and reuse lifecycle at CHOP: archiving, cataloging, display, and usability.

3. Conclusion

An initial challenge in drafting the metadata schema was determining the proper balance between adherence to intellectual archival arrangements and flexibility in item description (Pakstis et al., 2019). Since the launch of Arcus, the Metadata Librarians and Digital Archivists have quelled the need for constant evaluation of this tension by broadly implementing the project template, protocol, and manifest framework. Although as the Archives grows there will undoubtedly be smaller refinements to the templates and schemas, the largest opportunity for change in this next phase of Arcus operationalization is to dynamically address display and delivery of archived data. This is made possible through continued collaboration with study teams and by working closely with the Arcus Data Catalog and Discovery Librarian and the Arcus UX Specialist.

Finally, training for bioinformaticians and research study team data liaisons supplied by the Arcus Library Science team make implementation of the project template, manifests, and protocol files even more efficient for both the team processing the data contribution as well as for the study team. The research team can use the project template, manifest, and protocol structure in their data management to keep track of what and who are represented in their research outputs and where to find those files. They can then smoothly deposit files into the Arcus Archives where they will be securely stored, described, and metadata about those files displayed for reuse possibilities - and the cycle continues.

Acknowledgements

This team is indebted to Christiana Dobrzynski and Perry Evans for their outstanding work on the project template and its implementation during their tenures at CHOP. We would also like to acknowledge Hannah Calkins for her work disseminating the project template at CHOP and for her

immeasurable contributions as a co-author of the heretofore unpublished Arcus Archives Metadata Schema.

References

Clarke, Eric L., Louis J. Taylor, Chunyu Zhao et al. (2019). Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome. Volume 7, Article 46. https://doi.org/10.1186/s40168-019-0658-x. Retrieved June 2, 2022 from https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0658-x.

DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. http://doi.org/10.5438/0014

Department of Biomedical and Health Informatics. Arcus Internal Data Sharing and Attribution Policy. (2023). Children's Hospital of Philadelphia.

Dobrzynski, Christiana, Stephanie Huang, and Perry Evans. (2019). Shared Agency: Data Science, Programming, and Archiving Collaborate for Digital Preservation of Big Data. NDSA Digital Presentation DLF Forum 2019, from https://osf.io/3m58w/#.

DrivenData. (2021). Cookiecutter Data Science - Version 1 (Legacy). Retrieved June 1, 2022, from https://github.com/drivendata/cookiecutter-data-science/releases/tag/v1.

Gonzalez-Beltran, Alejandra, Philippe Rocca-Serra. (2017). biocaddie/WG3-MetadataSpecifications: DataMed DATS specification v2.2 - NIH BD2K bioCADDIE (v2.2). Zenodo. https://doi.org/10.5281/zenodo.438337

Laine, Christine, Steven N. Goodman, Michael E. Griswold, Harold C. Sox. 2007. Reproducible research: moving toward research the public can really trust. Annals of Internal Medicine. 146, 6 (March 2007), 450-453. DOI: https://doi.org/10.7326/0003-4819-146-6-200703200-00154

Laine, Heidi. (2017). Afraid of Scooping – Case Study on Researcher Strategies against Fear of Scooping in the Context of Open Science. Data Science Journal, 16, 29. DOI: http://doi.org/10.5334/dsj-2017-029

Lawson, Jonathan, Moran N. Cabili, Giselle Kerry, Tiffany Boughtwood, Adrian Thorogood, Pinar Alper, Sarion R. Bowers, Rebecca R. Boyles, Anthony J. Brookes, Matthew Brush, Tony Burdett, Hayley Clissold, Stacey Donnelly, Stephanie O.M. Dyke, Mallory A. Freeberg, Melissa A. Haendel, Chihiro Hata, Petr Holub, Francis Jeanson, Aina Jene, Minae Kawashima, Shuichi Kawashima, Melissa Konopko, Irene Kyomugisha, Haoyuan Li, Mikael Linden et al. 2021. The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genomics. Volume 1, Issue 2. https://doi.org/10.1016/j.xgen.2021.100028. Retrieved June 2, 2022, from https://www.sciencedirect.com/science/article/pii/S2666979X21000355.

Marco Masseroli. Biological and Medical Ontologies: Introduction. Editor(s): Shoba Ranganathan, Michael Gribskov, Kenta Nakai, Christian Schönbach, Encyclopedia of Bioinformatics and Computational Biology, Academic Press, Pages 813-822, https://doi.org/10.1016/B978-0-12-809633-8.20395-6. Retrieved April 19, 2023 from https://www.sciencedirect.com/science/article/pii/B9780128096338203956.

Mendes, Pedro. 2018. Reproducible research using biomodels. Bulletin of Mathematical Biology. 80, 3081-3087. DOI: https://doi.org/10.1007/s11538- 018-0498-z

Mesirov, Jill P.. 2010. Accessible reproducible research. Science. 327, 5964, 415- 416. DOI: https://doi.org/10.1126/science/1179653

Office of The Director, National Institutes of Health. NIH Policy for Data Management and Sharing. 2020. NOT-OD-21-013. Retrieved April 18, 2023 from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

Online Computer Library Center (OCLC). Bibliographic Formats and Standards. 2018. Retrieved April 18, 2023 from https://www.oclc.org/bibformats/en/4xx/490.html

Pakstis, Julianna, Hannah Calkins, Christiana Dobrzynski, Spencer Lamm and Laura McNamara, "Advancing Reproducibility Through Shared Data: Bridging Archival and Library Practice," 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 49-52, doi: 10.1109/JCDL.2019.00017.

Peng, Roger D.. 2011. Reproducible research in computational science. Science. 334, 6060 (Dec 2011), 1226-1227. DOI: https://doi.org/10.1126/science.12138847

Rawat, Seema and Sanjay Meena.. (2014). Publish or perish: Where are we heading?. Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences, 19(2), 87–89.

Riley, Jenn. 2017. Understanding Metadata: What is Metadata, and What is it For?. National Information Standards Organization, Baltimore, MD.

Roundtable on Environmental Health Sciences, Research, and Medicine; Board on Population Health and Public Health Practice; Health and Medicine Division; National Academies of Sciences, Engineering, and Medicine. 2016.The

https://doi.org/10.23106/dcmi.953336477

Benefits of Data Sharing. In Principles and Obstacles for Sharing Data from Environmental Health Research: Workshop Summary. National Academies Press (US), Washington DC.

Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, and Michela Taufer. 2016. Enhancing reproducibility for computational methods. Science. 354, 6317 (Dec 2016), 1240-1241. DOI: https://doi.org/10.1126/science.aah6168

U.S. Department of Health and Human Services (HHS). 2004. "Clinical Research and the HIPAA Privacy Rule."

U.S. Department of Health and Human Services (HHS). 2013. "Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule." Federal Register 78(17):5566-702.

https://doi.org/10.23106/dcmi.953336477