[Project scope 1](#_Toc253597074)

[Demographic data review 1](#_Toc2086590325)

[Recommendations on encoding data 2](#_Toc1566542099)

[An example from the archives 2](#_Toc581944189)

[Worked Databrary examples 3](#_Toc289803210)

[Recommendations on documentation 3](#_Toc1883200129)

# Project scope

Databrary (<https://databrary.org>) is a restricted-access data library specialized for storing, streaming, and sharing video and audio recordings collected as research data or documentation.[[1]](#footnote-9667) Databrary 2.0 is an NSF-funded project to build on the existing Databrary platform to ensure datasets stored on it comply with FAIR (findable, accessible, interoperable, reusable) standards through implementation of open metadata schema and vocabularies, documentation of these guidelines, and outreach to Databrary stakeholders.

To move toward FAIR compliance, demographic information about participants in research studies for which data is stored in Databrary should comply with one or more open schema. Databrary’s internal logic regarding user privacy and compliance with federal and depositor rules around health data dictate the terms by which users and user groups may view certain data properties and values. This permission logic, internal to Databrary, must also comply with open schema, to be built and maintained by the Databrary project team.

In furtherance of these goals the University Libraries reviewed the current demographic data stored about datasets and their constituent files in the current Databrary infrastructure, aligning each data property with one or more open schema in the domain of that property. This alignment will be reviewed by the Databrary team. The University Libraries also recommend the strategies below for documentation of the internal permission logic of Databrary, and on the documentation of data properties and values visible for each entry within that permission logic scheme.

# Demographic data review

The uses for which metadata in Databrary are designed are documented on its GitHub project page (<https://github.com/databrary/metadata>). The first priority of Databrary’s metadata design is to support open metadata schema and standards whenever possible. Among the schema supported by Databrary are the National Insitutes of Health (NIH) Common Data Elements; DataCite; and schema.org. Open serialization formats supported by Databrary include JSON-LD and the CSV Schema Language.

In addition to supporting open metadata initiatives, Databrary seeks to permit self-curation of data by allowing depositors to select the schema that best represent their data sets. These can be at cross-purposes. For example, a depositor may have collected data representing demographic information about project participants that is from a closed scheme, or one that is proprietary to their own project. To ensure interoperability, Databrary may require provisions to align such closed or proprietary schema with selected open schema endorsed by the project.

The *PLAY and Learning Across a Year* (PLAY) project (<https://play-project.org>) is a collaborative initiative by 65 researchers across 45 institutions in North America. Its primary purpose is to document play practices in early childhood development; a secondary purpose of the project is to model open, collaborative research and data management practices. To this end Databrary has documented the demographic data properties for which values are collected by PLAY.[[2]](#footnote-1316) These are derived from participant agreements, pre- and post-visit questionnaires, and data collected at the home visits themselves.

The PLAY project has two obstacles to ensuring FAIR data publication practices. The first is the collected data properties, which are not currently aligned thoroughly with open metadata schema (though work has been done to align key demographic properties with the NIH Common Data Elements). The second, and the more difficult to navigate, is the values assigned to those properties. There are open vocabularies in common use for assigning demographic values to properties, such as race and ethnicity, level of educational attainment, and geographical data. However, Databrary’s desire to allow self-curation of data extends to an individual researcher or project’s choice of demographic values; this may be at cross-purposes with a mandate to select a particular vocabulary for demographic information. Moreover, open schema with fixed values may not be sufficiently representative for demographic information such as gender or household structure. Open schema adoption within Databrary should be balanced against the needs of researchers to accurately reflect their project participants. ~~To this end the metadata strategies outlined here should be adopted with the consent of both researchers and, to the extent possible, project participants, through participant agreements.~~Participant agreements may need to be revised in order to allow participants to consent to the use of metadata about them for data reuse purposes.

# Recommendations on encoding data

JSON-LD is a generally accepted standard for performing this work. JSON-LD is a data exchange format using JavaScript Object Notation (JSON), a widely accepted format for encoding structured data on the World Wide Web. JSON-LD is easily recognized by search engines for indexing and retrieval, and can be used for both general Web searches using tools like Google or DuckDuckGo as well as specific searches within scholarly indexes such as Google Scholar.

## An example from the archives

To draw on the example of another discipline, members of the University Libraries team implement ArchivesSpace as the collection management and discovery system for archival materials held at both University Park and the Commonwealth Campus Libraries at Penn State. ArchivesSpace supports JSON-LD natively through its built-in public user interface, a public catalog of archival metadata used for search, discovery, and item retrieval. The public user interface of ArchivesSpace selectively publishes data from the database, allowing for access to materials without unauthorized access to information that is of a sensitive nature (such as the contact or demographic information of donors), not suitable for public viewing (such as internal collection processing notes or accession records), or is of a particularly technical nature (such as digital file checksums).

JSON-LD is embedded within <script> tags in the HTML generated by ArchivesSpace when a user clicks through to a record. An example of ArchivesSpace’s JSON-LD, in this case to describe the [T.R. Johns papers](https://aspace.libraries.psu.edu/repositories/3/resources/215), is provided below.

<script type="application/ld+json">  
{  
 "@context": "<http://schema.org/>",  
 "@id": "<https://aspace.libraries.psu.edu/repositories/3/resources/215>",  
 "@type": [  
 "Collection",  
 "ArchiveComponent"  
 ],  
 "name": "T. R. Johns papers",  
 "identifier": "01616",  
 "description": "T.R. Johns was the General Manager and Vice President of the Bethlehem Mines Corporation/Industrial Collieries Corporation, the mining subsidiary of the Bethlehem Steel Corporation.",  
 "creator": [  
 {  
 "@id": "<https://aspace.libraries.psu.edu/agents/people/12997>",  
 "@type": "Person",  
 "name": "Johns, T. R. (Thomas Richards)"  
 }  
 ],  
 "dateCreated": [  
 "1853-1948"  
 ],  
 "materialExtent": [  
 {  
 "@type": "QuantitativeValue",  
 "unitText": "Linear Feet",  
 "value": "6"  
 }  
 ],  
 "about": [  
 {  
 "@type": "Intangible",  
 "name": "Bituminous coal industry -- History -- 20th century"  
 },  
 {  
 "@type": "Intangible",  
 "name": "Coal mines and mining -- Pennsylvania -- 20th century"  
 },  
 {  
 "@type": "Intangible",  
 "name": "Coal mines and mining -- West Virginia -- 20th century"  
 },  
 {  
 "@type": "Intangible",  
 "name": "Coal Strike, Heilwood, Pa., 1922"  
 },  
 {  
 "@type": "Intangible",  
 "name": "Industrial management -- History -- 20th century"  
 },  
 {  
 "@type": "Intangible",  
 "name": "Labor movement -- History -- 20th century"  
 }  
 ],  
 "genre": [  
 "Scrapbooks",  
 "Photographic",  
 "Artifacts"  
 ],  
 "inLanguage": [  
 {  
 "@type": "Language",  
 "name": "English"  
 }  
 ],  
 "holdingArchive": {  
 "@id": "<https://aspace.libraries.psu.edu/repositories/3>",  
 "@type": "ArchiveOrganization",  
 "name": "Eberly Family Special Collections Library",  
 "address": {  
 "@type": "PostalAddress",  
 "streetAddress": [  
 "104 Paterno Library",  
 "Penn State University"  
 ],  
 "addressLocality": "University Park",  
 "postalCode": "16802",  
 "addressCountry": "USA"  
 }  
 }  
}  
</script>

While this particular use case of JSON-LD lacks features desired for the Databrary 2.0 project – it only implements linked data URIs that are native to the software, and it makes little effort to specify semantic characteristics of linked subject headings – it points toward a potential solution for making Databrary data sets more publicly accessible, and for researchers without Databrary credentials to identify research data in which they may have an interest.

Schema.org has specific design notes for how it treats [data and datasets](https://schema.org/docs/data-and-datasets.html).

## Worked Databrary examples

The below is a partial worked example of how a Databrary volume may be expressed in schema.org JSON linked data. In cases where Schema properties do not exist, or only exist in partial forms (e.g. in the way that Schema handles race and ethnicity data for individuals), Databrary predicates have been provided.

{

"@context": {

"schema": "https://schema.org/",

"databrary": "https://databrary.org/"

},

"@id": "http://doi.org/10.17910/b7.11",

"@type": "schema:Dataset",

"schema:name": "The Ties That Bind: Cradling in Tajikistan",

"schema:description": "A traditional childrearing practice—“gahvora” cradling—in Tajikistan and other parts of Central Asia purportedly restricts movement of infants’ body and limbs. However, the practice has been documented only informally in anecdotal reports. Thus, this study had two research questions: (1) To what extent are infants’ movements restricted in the gahvora? (2) How is time in the gahvora distributed over a 24-hour day in infants from 1–24 months of age? To answer these questions, we video-recorded 146 mothers cradling their infants and interviewed them using 24-hour time diaries to determine the distribution of time infants spent in the gahvora within a day and across age. Infants’ movements were indeed severely restricted. Although mothers showed striking uniformity in how they restricted infants’ movements, they showed large individual differences in amount and distribution of daily use. Machine learning algorithms yielded three patterns of use: day and nighttime cradling, mostly nighttime cradling, and mostly daytime cradling, suggesting multiple functions of the cradling practice. Across age, time in the gahvora decreased, yet 20% of 12- to 24-month-olds spent more than 15 hours bound in the gahvora. We discuss the challenges and benefits of cultural research, and how the discovery of new phenomena may defy Western assumptions about childrearing and development. Future work will determine whether the extent and timing of restriction impacts infants’ physical and psychological development.",

"schema:funding": {

"@type": "schema:MonetaryGrant",

"schema:name": "Effects Of Traditional Cradling Practices On Infants' Physical, Motor, And Social Development",

"schema:url": "https://www.nsf.gov/awardsearch/showAward?AWD\_ID=1349044",

"funder": {

"@type": "schema:FundingAgency",

"schema:name": "National Science Foundation",

"schema:alternateName": "NSF",

"schema:sameAs": "https://doi.org/10.13039/100000001"

}

},

"databrary:Investigator": [

{

"@type": "schema:Person",

"name": "Karasik, Lana",

"institution": "City University of New York"

},

{

"@type": "schema:Person",

"name": "Tamis-LaMonda, Catherine",

"institution": "New York University"

},

{

"@type": "schema:Person",

"name": "Adolph, Karen",

"institution": "New York University"

}

],

"databrary:Collaborator": [

{

"@type": "schema:Person",

"name": "Kuchirko, Yana",

"institution": "Brooklyn College, City University of New York"

},

{

"@type": "schema:Person",

"name": "Schneider, Joshua",

"institution": "University of Pittsburgh"

},

{

"@type": "schema:Person",

"name": "Cole, Whitney Graham",

"institution": "New York University"

},

{

"@type": "schema:Person",

"name": "Rrjolli, Solida",

"institution": "College of Staten Island"

}

],

"schema:citation": {

"@context": "https://schema.org",

"@id": "http://hdl.handle.net/10.1371/journal.pone.0204428",

"@type": "schema:ScholarlyArticle",

"schema:author": [

{

"@type": "schema:Person",

"schema:name": "Karasik, L. B."

},

{

"@type": "schema:Person",

"schema:name": "Tamis-LeMonda, C. S."

},

{

"@type": "schema:Person",

"schema:name": "Ossmy, O."

},

{

"@type": "schema:Person",

"schema:name": "Adolph, K. E."

}

],

"schema:name": "The ties that bind: Cradling in Tajikistan",

"schema:publisher": "PLoS One"

},

"databrary:hasRecords": [

{

"@type": "databrary:Record",

"name": "NY017",

"type": "session",

"release": "private"

}

]

}

The referenced volume is “[The ties that bind: Cradling in Tajikistan](https://nyu.databrary.org/volume/11).”

# Recommendations on documentation

Visibility of data properties and values within Databrary is driven by an internal permission logic system. This system controls the extent to which users may view properties and values for a Databrary project, based on their status within that project. For example, end users unaffiliated with a project but who wish to reuse its data would see only descriptive metadata about the project; PIs or researchers who work on the project would see a fuller representation of item-level demographic data. The permission levels are unique to Databrary. These permissions will be published as an open schema within the Databrary metadata documentation, and the visibility of each demographic data property as a function of permission levels will be published as a property therein across all schema serialization formats.

Documentation of Databrary metadata practices should be done sufficiently to meet the needs of several stakeholder groups, including the Databrary project team, researchers or research teams interested in reusing data sets stored in Databrary or depositing their own, and metadata professionals seeking to reuse Databrary metadata products in different discovery systems. The documentation strategies outlined here should meet the needs of metadata specialists and Databrary project participants. Other documentation strategies, tailored to the needs of researchers without a technical background in metadata design, may be identified through future outreach work.

1. “About Databrary.” <https://databrary.org/about.html> (accessed 2021 March 23). [↑](#footnote-ref-9667)
2. https://github.com/databrary/metadata/blob/main/nih-cde.md [↑](#footnote-ref-1316)