# Background

Databrary ([https://databrary.org](https://databrary.org/)) is a restricted-access data library for storing, streaming, and sharing video and audio recordings collected as research data or documentation. Its primary focus is the collection of data in behavioral, social, educational, and developmental science. The Databrary 2.0 project, supported by the National Science Foundation, builds on the existing Databrary platform to ensure its datasets, and the metadata that describes them, complies with FAIR (findable, accessible, interoperable, reusable) standards through implementation of open metadata schema and vocabularies, documentation of these guidelines, and outreach to Databrary stakeholders.

As part of the NSF study, the University Libraries consulted with the Databrary project team on metadata curation and publishing standards for research data. Particular attention was paid to the demographic data describing 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 report presents the results of this consultation. It also introduces further practices that the Databrary project may adopt in order to comply with the FAIR Principles.

# Databrary and the FAIR Principles

## Findability

The FAIR principles for findability are:

* Metadata are assigned a globally unique and persistent identifier
* Data are described with rich metadata
* Metadata clearly and explicitly include the identifier of the data they describe
* Metadata are registered or indexed in a searchable resource

Datasets in Databrary are assigned persistent identifiers; the aggregations of metadata that describe these datasets should be provided distinct persistent identifiers, so that they may be referenced separately. The metadata describing resources in Databrary is already quite rich, but more work may be done to ensure its fitness for the purposes for which it is used (e.g. self-curation and management of deposits, discovery of data by outside researchers). The identification and definition of core and extended sets of data properties that best facilitate finding and reusing Databrary datasets should be the subject of future user research studies. Finally, metadata describing Databrary datasets should be aggregated in such a way that they may be made available for discovery in searchable resources. This includes Databrary itself, as well as any third-party data repositories for which Databrary makes itself available for aggregation. [The FAIR Data Point project](https://www.fairdatapoint.org/) points a way forward for this.

## Accessibility

The FAIR principles for accessibility are:

* Metadata are retrievable by their identifier using a standardized communications protocol
  + The protocol is open, free, and universally implementable
  + The protocol allows for an authentication and authorization procedure, where necessary
* Metadata are accessible even when the data are no longer available

Databrary is in the process of making its metadata available through [an R package](https://github.com/databrary/databraryr) and through an open API. Determining the data structures for making metadata available through the API is one of the primary deliverables of the Libraries’ ongoing collaborations with the Databrary project. Databrary is well on its way to aligning with the FAIR accessibility principles, although in our work we identified two issues to address in future work. First, the categories of access restriction that may be placed on data vary even among individual users or within a single dataset. This has implications for how accessible data may be for a particular user, or for the degree to which a metadata aggregation may be visible to a user who requests it using an API. Databrary will need to keep an internal record of these permissions; this record will itself need to be published in accordance with the FAIR Principles, with appropriate identifiers and version control that can be referenced in a metadata aggregation. Second, datasets that are no longer available in Databrary for any reason (withdrawn from the repository; made unavailable to a user due to changes in circumstances) will still need to have metadata for them made available. Rationale for the unavailability of this data will need to be published in a standardized way.

## Interoperability

The FAIR principles for interoperability are:

* Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
* Metadata use vocabularies that follow FAIR principles.
* Metadata include qualified references to other metadata.

Databrary is in the process of establishing metadata encoding standards for the various places where metadata aggregations describing its datasets may be found, including the Databrary website, the API, and the Databrary R package. Wherever possible, Databrary should refer to property value sets that are published as open data with persistent, referenceable URIs to ensure data interoperability. For publishing data through a publicly-accessible Web frontend, JSON-LD (or JSON for Linking Data) is a widely-used standard. JSON-LD uses broadly applicable open vocabularies such as Schema.org to publish metadata property/value pairs in a way that search engines can recognize and integrate into their indexes. Examples of metadata aggregations encoded as JSON-LD for the Web may be found in this report and in the related Metadata Recommendations document.

In some cases this may come into conflict with the ways in which researchers and PIs have chosen to collect and document their own data. For these situations Databrary should have a statement or memorandum of understanding with depositors allowing for some post-deposit curation of metadata that prserves the provenance and original order of the data while allowing for interoperability and shared understanding of data properties, according to the FAIR standards (see the “Reusability” section below).

## Reusability

The FAIR principles for reusability are:

* Metadata are richly described with a plurality of accurate and relevant attributes.
  + Metadata are released with a clear and accessible data usage license.
  + Metadata are associated with detailed provenance.
  + Metadata meet domain-relevant community standards.

The principles regarding rich description using clearly-defined and relevant data properties are echoed in many of the other FAIR principles. The application of reusability principles, more so than the other three categories, follow from researched and documented use cases for metadata and data collected from the Databrary community, and are difficult to glean from metadata specialization alone. Data usage licenses should be implemented in Databrary that meet the needs for both data sharing among researchers and the protection of privacy for research subjects. Data provenance, both in the deposit of metadata and in its ongoing curation (e.g. the enhancement of metadata to conform with open vocabularies for attributes such as demographic data) should be documented and made available to aid in the understanding of datasets. Finally, direct observation of and outreach to researchers, in particular their needs for sharing data and re-using the shared data of others, should be considered in developing description practices.

# Appendix: JSON-LD examples

## From Databrary

The below is a partial worked example of how a Databrary volume may be expressed with JSON-LD. 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. The referenced volume is “[The ties that bind: Cradling in Tajikistan](https://nyu.databrary.org/volume/11).”

{

"@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"

}

]

}

## From University Libraries Archival Collections

To draw on the example of another discipline: [The University Libraries Archival Collections public frontend](https://aspace.libraries.psu.edu) for searching and retrieving archival records uses JSON-LD. 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>