

# Collections as/are Data[sets]: Connecting Research Data Management to Digital Collections

Jason A. Clark @jaclark

Montana State University  
Research Data Access and Preservation Summit 2020

# Outline

Datasets + Collection Worldviews

Collections are data

Benefits

Standards + Prototypes

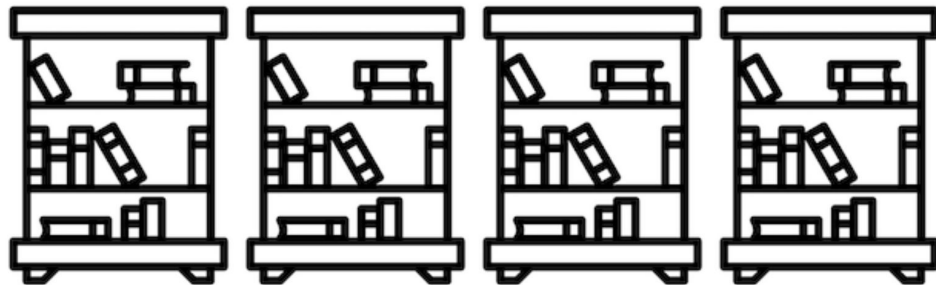
Implications

# Research Motivation

ALWAYS ALREADY COMPUTATIONAL - COLLECTIONS AS DATA



## Always Already Computational



Thomas Padilla

[@thomasgpadilla](https://twitter.com/thomasgpadilla)

## A Dataset is a Worldview

On subjective data, why datasets should expire, & data sabotage.



Hannah Davis [Follow](#)

Mar 5 · 6 min read



*This is a slightly expanded version of a talk given at the Library of Congress in September 2019.*

## A Dataset is a Worldview

My name is Hannah Davis and I'm a research artist, generative musician, and data scientist working with machine learning. Most of my work deals with ideas around emotional data, algorithmic composition, and dataset creation.

Hannah Davis

[@ahandvanish](#)

## Research Motivation - Collections as/are Dataset[s]

childbirth	anger	0	
childbirth	anticipation		0
childbirth	disgust	0	
childbirth	fear	0	
childbirth	joy	0	
childbirth	negative		0
childbirth	positive		0
childbirth	sadness	0	
childbirth	surprise		0
childbirth	trust	0	

The word "childbirth" had been tagged as being completely unemotional.

“This led to a perspective that has informed all of my work since: a dataset is a worldview. It **encompasses the worldview of the people who scrape and collect the data, whether they’re researchers, artists, or companies.** It **encompasses the worldview of the labelers**, whether they labeled the data manually, unknowingly, or through a third party service like Mechanical Turk, which comes with its own demographic biases.”

Hannah Davis (2019)

<https://towardsdatascience.com/a-dataset-is-a-worldview-5328216dd44d>

# Connecting Worldviews



## Collections Worldview

Findability

Discovery

Inventory

Linkable

## Datasets Worldview

Reproducibility

Sharing + Reuse

Preservation

Citeable

Connecting digital collection worldviews to  
dataset worldviews helps us realize that  
collections are datasets.

# Benefits?

# Benefits

Manifest for search engines

Metadata

Reuse

Preservation

Citeable

# Standards + Prototypes

**schema.org**

Custom Search

Q

HomeSchemasDocumentation

## DataFeed

[Thing](#) > [CreativeWork](#) > [Dataset](#) > [DataFeed](#)

A single feed providing structured information about one or more entities or topics.

[more...]

Property	Expected Type	Description
Properties from <a href="#">DataFeed</a>		
<a href="#">dataFeedElement</a>	<a href="#">DataFeedItem</a> or <a href="#">Text</a> or <a href="#">Thing</a>	An item within in a data feed. Data feeds may have many elements.
Properties from <a href="#">Dataset</a>		
<a href="#">distribution</a>	<a href="#">DataDownload</a>	A downloadable form of this dataset, at a specific location, in a specific format.
<a href="#">includedInDataCatalog</a>	<a href="#">DataCatalog</a>	A data catalog which contains this dataset. Supersedes <a href="#">catalog</a> , <a href="#">includedDataCatalog</a> . Inverse property: <a href="#">dataset</a> .
<a href="#">issn</a>	<a href="#">Text</a>	The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different formats of, or the linking ISSN (ISSN-L) for, this serial

<https://schema.org/Dataset>

<https://schema.org/DataFeed>

# Standards - Research Object Crate (RO-Crate)

ro-crate



Research Object Crate

View the Project on GitHub  
ResearchObject/ro-crate

This project is maintained by  
ResearchObject

Hosted on GitHub Pages — Theme by orderedlist

## RO-Crate Metadata Specification 1.0

- Permalink: <https://w3id.org/ro/crate/1.0>
- Status: Recommendation
- JSON-LD context: <https://w3id.org/ro/crate/1.0/context>
- This version: <https://w3id.org/ro/crate/1.0>
- Previous version: <https://w3id.org/ro/crate/0.2>
- Published: 2019-11-15
- Publisher: [researchobject.org](https://researchobject.org) community
- Cite as: <https://doi.org/10.5281/zenodo.3541888> (this version)  
<https://doi.org/10.5281/zenodo.3406497> (any version)
- Editors: Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes
- Authors: Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes, Oscar Corcho, Daniel Garijo, Raul Palma, Frederik Coppens, Carole Goble, José María Fernández, Kyle Chard, Jose Manuel Gomez-Perez, Michael R Crusoe, Ignacio Eguinoa, Nick Juty, Kristi Holmes, Jason A. Clark, Salvador Capella-Gutierrez, Alasdair J. G. Gray, Stuart Owen, Alan R Williams, Giacomo Tartari, Finn Bacall, Thomas Thelen

### 1. Introduction & definition of an RO-Crate

1. Terminology
2. Linked Data conventions

### 2. RO-Crate Structure

1. RO-Crate Metadata File ([ro-crate-metadata.jsonld](#))
2. RO-Crate Website ([ro-crate-preview.html](#) and [ro-crate-preview\\_files/](#))

a lightweight approach to  
packaging research data with their  
metadata

<https://w3id.org/ro/crate/1.0>

[#digital-library-and-repository  
-content](#)

# Collections API as DataFeed

## Application Programming Interface



# API Payload as a DataFeed (.jsonld)

```
1 {
2   "@context": "http://schema.org/",
3   "@type": "DataFeed",
4   "name": "James Willard Schultz Collection",
5   "description": "API result for James Willard Schultz Collection",
6   "license": "https://creativecommons.org/licenses/by/4.0/",
7   "identifier": "https://arc.lib.montana.edu/schultz-0010/",
8   "dateModified": "2015-01-02",
9   "dataFeedElement": [
10     {
11       "@type": "DataFeedItem": [
12         {
13           "item": {
14             "recordInfo_recordIdentifier": "64",
15             "identifier_ark": "ark:/75788/m4sw9n",
16             "identifier": "064.jpg",
17             "proxyIdentifier": "",
18             "titleInfo_title": "Photo of a young man in ceremonial dress playing a drum, Browning, Montana",
19             "abstract": "Boy in full Indian dress standing on a rug outside of a building and playing an Indian drum. Nothing written on verso.",
20             "extension": "clothing and dress, drums, blackfeet indians, indians of north america, montana",
21             "name_namePart": "James Willard Schultz",
22             "originalMetadata_name": "Unknown",
23             "originInfo_dateIssued": "1859-1947",
24             "originalMetadata_dateIssued": "Unknown",
25             "originInfo_dateCreated": "2005-09-09",
26             "location_physicalLocation": "Montana State University Library Collection 10 - James Willard Schultz Papers, 1867-1969",
27             "typeOfResource": "1 Photographic Print, B&W, 15 x 9 cm",
28             "originInfo_publisher": "Montana State University--Bozeman",
29             "accessCondition": "https://creativecommons.org/licenses/by-nc-sa/3.0/us/",
30             "subject_topic1": "Montana--History",
31             "subject_topic2": "Indians of North America--Montana",
32             "subject_topic3": "Schultz, James Willard, 1859-1947",
33             "subject_topic4": "Native American Music",
34             "note": "None",
35             "genre": "Still Image",
36             "physicalDescription_internetMediaType": "image/jpeg",
```

<https://arc.lib.montana.edu/schultz-0010/api.php?v=1&type=search&q=boy&limit=10&format=json>

# James Willard Schultz

Progressive Web App (PWA) built using DataFeed and  
RO-Crate standards as datastore (.jsonld)

## James Willard Schultz Collection

[Search](#) [Browse](#) [About](#)

James Willard Schultz Photos & Personal Papers Collection includes photos & documents of Blackfeet, Blood, Kootenai, Shoshone and Arapaho Native Americans, Glacier and Waterton Lakes National Parks, as well as his professional writing career and personal papers.

### FEATURED



1939



1939



1939



1939



1939



1939



1939



1932

[Search](#) [Browse](#) [About](#)

## BROWSE

a timeline view of the collection

### 1939 James Willard Schultz and Friends: in a Garden, 1939

2 older women and 1 older man sitting in garden in front of house. Nothing written on verso.

### 1939 Bernadotte: Blackfeet Indians: White Man and Woman Dressed in Indian Garb Pose with Three Indians, 1939

White man and woman dressed in Indian garb posing with 3 Indians in full dress outside in meadow. Nothing written on verso.

### 1939 Bernadotte: Blackfeet Indians: Man in Full Indian Dress in Front of Tipi, 1939

# Manifest (.json)

```
1 {
2   "name": "James Willard Schultz Collection",
3   "short_name": "SchultzPWA",
4   "start_url": "./?utm_source=homescreen",
5   "display": "standalone",
6   "background_color": "#213c69",
7   "description": "A digital library template app",
8   "theme_color": "#213c69",
9   "prefer_related_applications": false,
10  "categories": [
11    "books",
12    "education"
13  ],
14  "icons": [
15    {
16      "src": "./img/icons/icon-72x72.png",
17      "sizes": "72x72",
18      "type": "image/png"
19    },
20    {
21      "src": "./img/icons/icon-96x96.png",
22      "sizes": "96x96",
23      "type": "image/png"
24    },
25    {
26      "src": "./img/icons/icon-128x128.png",
27      "sizes": "128x128",
28      "type": "image/png"
29    },
30    {
31      "src": "./img/icons/icon-144x144.png",
32      "sizes": "144x144",
33      "type": "image/png"
34    },
35    {
36      "src": "./img/icons/icon-152x152.png",
37      "sizes": "152x152",
```

# Items (.jsonld)

```
1  {
2    "@context": "http://schema.org/",
3    "@type": "DataFeed",
4    "name": "James Willard Schultz Collection",
5    "description": "API result for James Willard Schultz Collection",
6    "license": "https://creativecommons.org/licenses/by/4.0/",
7    "identifier": "https://arc.lib.montana.edu/schultz-0010/",
8    "dateModified": "2015-01-02",
9    "dataFeedElement": [
10     {
11       "@type": "DataFeedItem": [
12         {
13           "item": {
14             "recordInfo_recordIdentifier": "64",
15             "identifier_ark": "ark:/75788/m4sw9n",
16             "identifier": "064.jpg",
17             "proxyIdentifier": "",
18             "titleInfo_title": "Photo of a young man in ceremonial dress playing a drum, Browning, Montana",
19             "abstract": "Boy in full Indian dress standing on a rug outside of a building and playing an Indian drum. Nothing written on verso.",
20             "extension": "clothing and dress, drums, blackfeet indians, indians of north america, montana",
21             "name_namePart": "James Willard Schultz",
22             "originalMetadata_name": "Unknown",
23             "originInfo_dateIssued": "1859-1947",
24             "originalMetadata_dateIssued": "Unknown",
25             "originInfo_dateCreated": "2005-09-09",
26             "location_physicalLocation": "Montana State University Library Collection 10 - James Willard Schultz Papers, 1867-1969",
27             "typeOfResource": "1 Photographic Print, B&W, 15 x 9 cm",
28             "originInfo_publisher": "Montana State University--Bozeman",
29             "accessCondition": "https://creativecommons.org/licenses/by-nc-sa/3.0/us/",
30             "subject_topic1": "Montana--History",
31             "subject_topic2": "Indians of North America--Montana",
32             "subject_topic3": "Schultz, James Willard, 1859-1947",
33             "subject_topic4": "Native American Music",
34             "note": "None",
35             "genre": "Still Image",
36             "physicalDescription_internetMediaType": "image/jpeg",
```

# Implications for Research Data Practice

# Implications

Identifying collections as datasets

Encoding collections as datasets

Prioritizing reuse

Prioritizing reproducibility

Connecting research data metadata  
practices to collections metadata



# Collections are datasets.

- emulatable and preservable
- set up for indexing and metadata (manifest file)
- static “databases” that can be reused/redeployed

# How can we continue to connect our worldviews?

- Cross-departmental appointments
- Recognition of emerging dataset standards
- Release our collections as datasets

**\*\*Thank you\*\***

Jason A. Clark  
@jaclark