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
Collections as/are Dataset[s]

Outline

Datasets + Collection Worldviews

Collections are data

Benefits

Standards + Prototypes

Implications
Research

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

Always Already Computational

Thomas Padilla

@thomasgpadilla
A Dataset is a Worldview

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

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.
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
<table>
<thead>
<tr>
<th>Collections Worldview</th>
<th>Datasets Worldview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findability</td>
<td>Reproducibility</td>
</tr>
<tr>
<td>Discovery</td>
<td>Sharing + Reuse</td>
</tr>
<tr>
<td>Inventory</td>
<td>Preservation</td>
</tr>
<tr>
<td>Linkable</td>
<td>Citeable</td>
</tr>
</tbody>
</table>
Connecting digital collection worldviews to dataset worldviews helps us realize that collections are datasets.
Benefits?
Benefits

- Manifest for search engines
- Metadata
- Reuse
- Preservation
- Citable
Standards + Prototypes
## DataFeed

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

### Properties from `DataFeed`

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>dataFeedElement</code></td>
<td></td>
<td>An item within in a data feed. Data feeds may have many elements.</td>
</tr>
</tbody>
</table>

### Properties from `Dataset`

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>distribution</code></td>
<td><code>DataDownload</code></td>
<td>A downloadable form of this dataset, at a specific location, in a specific format.</td>
</tr>
<tr>
<td></td>
<td><code>DataCatalog</code></td>
<td>A data catalog which contains this dataset. Supersedes <code>catalog</code>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inverse property: <code>dataset</code>.</td>
</tr>
<tr>
<td><code>includedInDataCatalog</code></td>
<td></td>
<td>'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'</td>
</tr>
</tbody>
</table>
Standards - Research Object Crate (RO-Crate)

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 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, Sían Soiland-Reyes
- Authors: Peter Sefton, Eoghan Ó Carragáin, Sían Soiland-Reyes,
  Oscar Corcho, Daniel Garijo, Raul Palma, Frederik Coppens, Carole
  Goble, José María Fernández, Kyle Chard, Jose Manuel Gomez-Perez,
  Michael R Cruse, Ignacio Eguinoa, Nick Juty, Kristi Holmes, Jason A.
  Clark, Salvador Capella-Gutierrez, Alasdair J. G. Gray, Stuart Owen,
  Alan R Williams, Giacomo Tartari, Finn Bacal, 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)

```json
"@context": "http://schema.org/",
"@type": "DataFeed",
"name": "James Willard Schultz Collection",
"description": "API result for James Willard Schultz Collection",
"license": "https://creativecommons.org/licenses/by/4.0/",
"identifier": "https://arc.lib.montana.edu/schultz-0010/",
"dateModified": "2015-01-02",
"dataFeedElement": [
  {"@type": "DataFeedItem": [{
    "item": {
      "recordInfo_recordIdentifier": "64",
      "identifier_representation": "ark:/75778/m4sw9r",
      "identifier": "664.jpg",
      "proxyIdentifier": "",
      "titleInfo_title": "Photo of a young man in ceremonial dress playing a drum, Browning, Montana",
      "abstract": "Boy in full Indian dress standing on a rug outside of a building and playing an Indian drum. Nothing written on verso.",
      "extension": "clothing and dress, drums, blackfeet indians, indians of north america, montana",
      "name_namePart": "James Willard Schultz",
      "originalMetadata_name": "Unknown",
      "originalInfo_dateIssued": "1859-1947",
      "originalMetadata_dateCreated": "Unknown",
      "originalInfo_dateCreated": "2005-09-09",
      "location_physicalLocation": "Montana State University Library Collection 10 - James Willard Schultz Papers, 1867-1969",
      "typeOfResource": "1 Photographic Print, B&W, 15 x 9 cm",
      "originInfo_publisher": "Montana State University--Bozeman",
      "accessCondition": "https://creativecommons.org/licenses/by-nc-sa/3.0/us/",
      "subject_topic1": "Montana--History",
      "subject_topic2": "Indians of North America--Montana",
      "subject_topic3": "Schultz, James Willard, 1859-1947",
      "subject_topic4": "Native American Music",
      "note": "None",
      "genre": "Still Image",
      "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)
Prototype

James Willard Schultz Collection

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
Prototype

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)

```json
{
   "name": "James Willard Schultz Collection",
   "short_name": "SchultzPWA",
   "start_url": "./?utm_source=homescreen",
   "display": "standalone",
   "background_color": "#213c69",
   "description": "A digital library template app",
   "theme_color": "#213c69",
   "prefer_related_applications": false,
   "categories": [
      "books",
      "education"
   ],
   "icons": [
      {
         "src": "./img/icons/icon-72x72.png",
         "sizes": "72x72",
         "type": "image/png"
      },
      {
         "src": "./img/icons/icon-96x96.png",
         "sizes": "96x96",
         "type": "image/png"
      },
      {
         "src": "./img/icons/icon-128x128.png",
         "sizes": "128x128",
         "type": "image/png"
      },
      {
         "src": "./img/icons/icon-144x144.png",
         "sizes": "144x144",
         "type": "image/png"
      },
      {
         "src": "./img/icons/icon-152x152.png",
         "sizes": "152x152",
         "type": "image/png"
      }
   ]
}```
"@context": "http://schema.org/",
"@type": "DataFeed",
"name": "James Willard Schultz Collection",
"description": "API result for James Willard Schultz Collection",
"license": "https://creativecommons.org/licenses/by/4.0/",
"identifier": "https://arc.lib.montana.edu/schultz-0010/",
"dateModified": "2015-01-02",
"dataFeedElement": [
  {
    "@type": "DataFeedItem": [
      {
        "item": {
          "recordInfo_recordIdentifier": "64",
          "identifier_orc": "ark:/75788/m4sw9r",
          "identifier": "064.jpg",
          "proxyIdentifier": "",
          "titleInfo_title": "Photo of a young man in ceremonial dress playing a drum, Browning, Montana",
          "abstract": "Boy in full Indian dress standing on a rug outside of a building and playing an Indian drum. Nothing written on verso."
        }
      }
    }
  }
]
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
Research Implication

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