

Packaging research data with DataCrate - a cry for help!

Peter Sefton
Michael Lynch

Motivation



Package data with useful context

- Who ... made it? ... funded it?
- What ... format are these files? ... is the research about?
- Where / when ... was it collected? ... is it about?
- Why ... was it done? ... <link>
- How ... was the data created? ... can I repeat that process?



What is it?

- **Files** in a directory structure
- which may be in a **BagIt** bag
- described by a **JSON-LD** catalog
- which uses **Schema.org**
- and meets the **DataCrate spec**
- and can be rendered as a **set of HTML documents**




CATALOG.json

- Linked data in JSON-LD
- “Graph” of metadata
- Mandatory: minimal
- As much detail as needed
- Classes and relations from schema.org



Human-readable HTML metadata

isPartOf	Pictures
thumbnail?	
creator?	Peter Sefton
path?	2017-06-11 12.56.14.jpg
encodingFormat?	Exchangeable Image File Format (Compressed)
contentSize?	4.88 MB
exifData?	<ul style="list-style-type: none">▶ AELock : Off --to-- CompressionFactor : 4▼ Contrast : Normal --to-- ExposureMode : Manual<ul style="list-style-type: none">• Contrast : Normal• ContrastSetting : 0 (min -5, max 5)• CreateDate : 2017:06:11 12:56:14• CropHeight : 3024• CropLeft : 28 0• CropTop : 30 0• CropWidth : 4032• CustomRendered : Normal

Data provenance

@id	DataCapture_wcr03
@type	CreateAction?
agent?	Robert Zlot
instrument?	bentwing
object?	Victoria Arch
result?	<ul style="list-style-type: none">• wcr03_victoria_arch.laz• wcr03_victoria_arch_traj.txt

This file was created at 2018-10-04T01:39:50.283Z by [Calcyte](#) which implements the [Draft DataCrate Packaging format](#), version 1.0

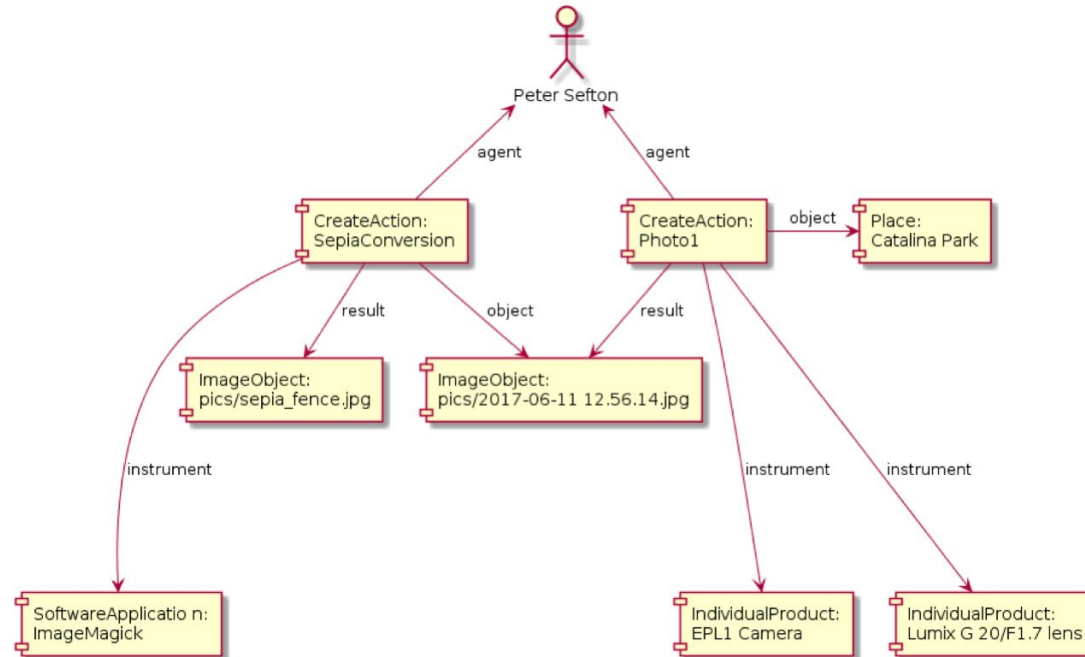
Software can be an instrument too

@id	SepiaConversion
name?	Converted dog picture to sepia
@type	CreateAction?
description?	convert -sepia-tone 80% test_data/sample/pics/2017-06-11\ 12.56.14.jpg test_data/sample/pics/sepia_fence.jpg
endTime?	2018:09:19T17:01:07+10:00
instrument?	ImageMagick
object?	pics/2017-06-11 12.56.14.jpg
result?	pics/sepia_fence.jpg

Machine-readable metadata

```
{
  "@id": "SepiaConversion",
  "@type": "CreateAction",
  "description": "convert -sepia-tone 80% test_data/sample/pics/2017-06-11\\
12.56.14.jpg test_data/sample/pics/sepia_fence.jpg",
  "endTime": "2018:09:19T17:01:07+10:00",
  "identifier": "SepiaConversion",
  "instrument": {
    "@id": "https://www.imagemagick.org/"
  },
  "name": "Converted dog picture to sepia",
  "object": {
    "@id": "pics/2017-06-11 12.56.14.jpg"
  },
  "result": {
    "@id": "pics/sepia_fence.jpg"
  }
},
```

Other ways of viewing metadata



Terms link to their definitions

The screenshot shows the schema.org website interface. At the top, there is a dark red header with the 'schema.org' logo on the left, a 'Custom Search' input field with a magnifying glass icon on the right, and navigation links for 'Home', 'Schemas', and 'Documentation'. Below the header, the main content area is light gray. On the left, the term 'CreateAction' is displayed in bold red text, followed by its canonical URL: 'http://schema.org/CreateAction'. Below this, a breadcrumb trail reads 'Thing > Action > CreateAction'. The main definition text states: 'The act of deliberately creating/producing/generating/building a result out of the agent.' and 'Usage: Between 10 and 100 domains'. To the right of this text, a white box contains the text '@type CreateAction?'. At the bottom right of the main text area, there is a '[more...]' link. Below the main text is a table with three columns: 'Property', 'Expected Type', and 'Description'. The table lists properties from the 'Action' class: 'actionStatus' (type: 'ActionStatusType'), 'agent' (types: 'Organization' or 'Person'), and 'endTime' (type: 'DateTime').

schema.org Custom Search

Home Schemas Documentation

CreateAction
Canonical URL: <http://schema.org/CreateAction>

Thing > Action > CreateAction

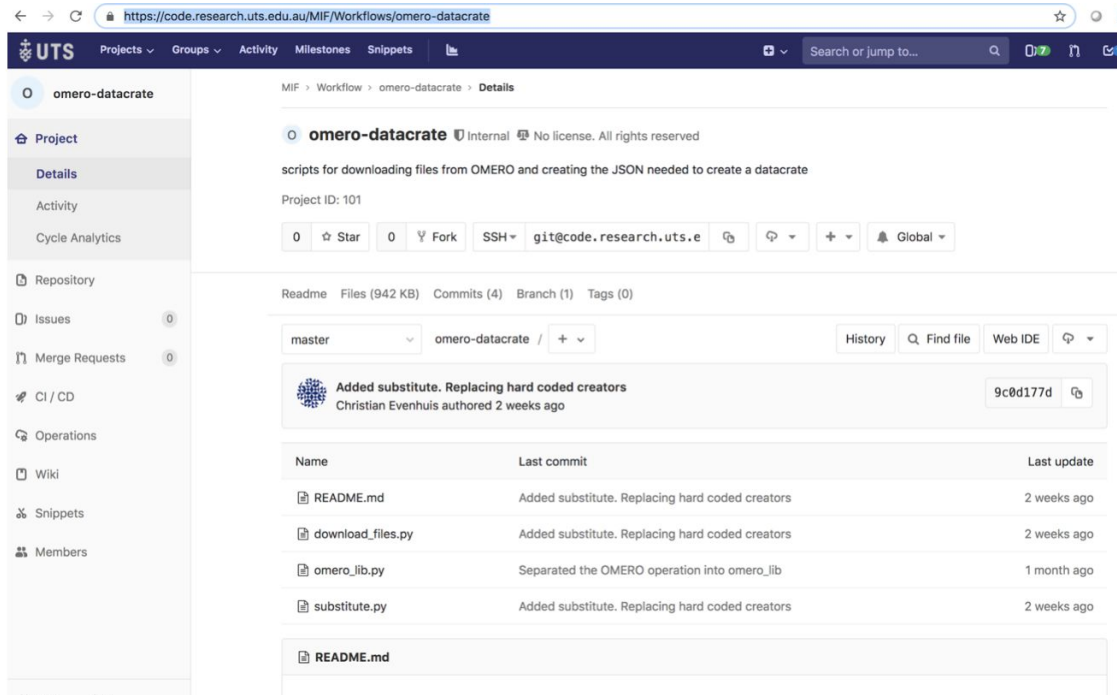
The act of deliberately creating/producing/generating/building a result out of the agent.

Usage: Between 10 and 100 domains

[more...]

Property	Expected Type	Description
Properties from Action		
actionStatus	ActionStatusType	Indicates the current disposition of the Action.
agent	Organization or Person	The direct performer or driver of the action (animate or inanimate). e.g. <i>John</i> wrote a book.
	DateTime	The endTime of something. For a reserved event or service (e.g. FoodEstablishmentReservation), the time that it is expected to end.

OMERO to DataCrate



The screenshot shows a GitHub repository page for 'omero-datacrate' on the UTS domain. The page is viewed in a browser with the URL <https://code.research.uts.edu.au/MIF/Workflows/omero-datacrate>. The repository is owned by 'omero-datacrate' and is internal. The description states: 'scripts for downloading files from OMERO and creating the JSON needed to create a datacrate'. The project ID is 101. The repository has 0 stars, 0 forks, and 0 tags. The current branch is 'master'. The commit history shows a recent commit by Christian Evenhuis titled 'Added substitute. Replacing hard coded creators' with the hash 9c0d177d. The commit message is 'Added substitute. Replacing hard coded creators'. The commit was authored 2 weeks ago. The repository contains 942 KB of files, 4 commits, 1 branch, and 0 tags. The file list includes README.md, download_files.py, omero_lib.py, substitute.py, and README.md.

UTS Projects Groups Activity Milestones Snippets

Search or jump to...

omero-datacrate

MIF > Workflow > omero-datacrate > Details

omero-datacrate Internal No license. All rights reserved

scripts for downloading files from OMERO and creating the JSON needed to create a datacrate

Project ID: 101

0 Star 0 Fork SSH git@code.research.uts.e

Readme Files (942 KB) Commits (4) Branch (1) Tags (0)

master omero-datacrate / + History Find file Web IDE

Added substitute. Replacing hard coded creators
Christian Evenhuis authored 2 weeks ago 9c0d177d

Name	Last commit	Last update
README.md	Added substitute. Replacing hard coded creators	2 weeks ago
download_files.py	Added substitute. Replacing hard coded creators	2 weeks ago
omero_lib.py	Separated the OMERO operation into omero_lib	1 month ago
substitute.py	Added substitute. Replacing hard coded creators	2 weeks ago

README.md

Tooling

- JSON-LD tooling is not quite there
- Calcyte – allowed data description with spreadsheets, and also generated HTML
- Split this into Describo and the node.js datacrate package
- Gerard Devine is developing Python tooling at Western Sydney University



Aligning with other efforts

- OCFL for repositories of DataCrates
- Portland Common Data Model for modelling collections
- Research Object Bundle also uses JSON-LD but is more complicated and lacks the human-readable HTML
- dataspice is targeting the same problems with schema.org and JSON



Help wanted!

- Critique the standard
- Generate more sample datasets, as fixtures for people who will ...
- ... write packaging tools
 - Export from more data management systems like MyTardis
 - Write a GUI or web tool for people to create Data Crates



Links

- <https://github.com/UTS-eResearch/datacrate>
- <https://schema.org>
- <https://ocfl.io/>
- <https://en.wikipedia.org/wiki/BagIt>
- <https://pcdm.org/>



Thank
you

