



# Wikidata and EaaSI

---

Katherine Thornton

6 November, 2018

- Wikidata and EaaS
- Wikidata
- WikiDP
- Resources

- Collaborative modeling of the domain of computing
- EaaS metadata will be interoperable with data models in Wikidata
- We are contributing to open knowledge

# Happy Birthday, Wikidata!

Wikidata went live on 29 October 2012.



# Wikidata Birthday Celebrations



**Figure 1:** Map of 6th birthday celebrations



# WIKIDATA

## This knowledge base of structured data is:

- Machine-readable linked open data
- Editable by anyone with Internet access
- Designed to support both human and algorithmic curation
- Fully-versioned wiki
- around 300 human languages

- over **19,000** active editors per month
- over **51 million** items have been created
- more than **260** bots



# Items, properties, unique identifiers

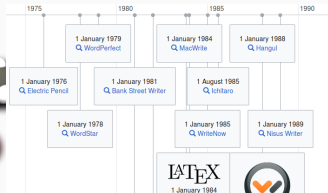
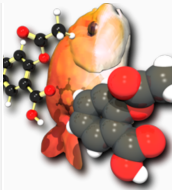
The screenshot shows the Wikidata page for 'Portable Document Format' (Q42332). The page is in English and displays various statements and their corresponding Wikidata properties. Annotations on the left side of the page identify specific properties and their values:

- Property: P3** (instance of) points to the statement 'is a' with the value 'file format family'.
- Property: P1813** (nickname) points to the statement 'nickname' with the value 'PDF'.
- Property: P144** (based on) points to the statement 'based on' with the value 'PostScript'.
- Property: P178** (developer) points to the statement 'developed by' with the value 'Adobe'.

On the right side of the page, a list of Wikidata items is shown, including 'Wikipedia' (Q42332), 'PDF' (Q42332), 'PostScript' (Q42332), and 'Adobe' (Q42332). The unique identifiers for these items are listed next to them: Q26085352, Q42332, Q218170, and Q11463.

# Status of software data in Wikidata

- **85,000** instances of software in Wikidata today
- commercial software
- research software
- free and open source software (FLOSS)



# Timeline of discontinued Google products

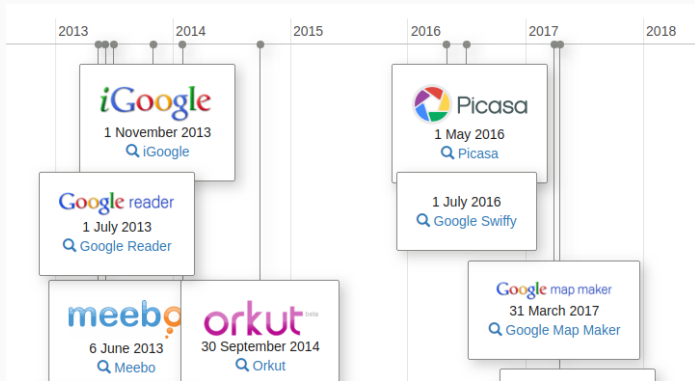


Figure 2: Try this query!

# What software available under a free software license can I use to open .obj files?

```
1 SELECT DISTINCT ?app ?appLabel ?logo WHERE {  
2   ?app (wdt:P31/wdt:P279*) wd:Q7397.  
3   ?app wdt:P1072 wd:Q2119595.  
4   ?app wdt:P275 ?lic.  
5   ?lic (wdt:P31/wdt:P279*) wd:Q3943414.  
6   OPTIONAL {?app wdt:P154 ?logo.}  
7   SERVICE wikibase:label { bd:serviceParam wikibase:language  
8 }
```

Figure 3: Try this query!

# Wikidata is a linking hub for external IDs

- External IDs have their own data type
- 58 percent of WD properties are external ids 2570/4439
- Joachim Neubert's paper on this topic

A screenshot of a Wikidata 'External sources' table for the NumPy entity. The table has a light blue header with the text 'External sources' and a downward arrow icon. The table contains ten rows, each with a source name and a corresponding external ID. The source names are in blue text, and the external IDs are in green text. The rows are: Arch package (python-numpy), Debian stable package (python-numpy), Fedora package (numpy), Free Software Directory entry (NumPy), Freebase (/m/021plb), Gentoo package (dev-python/numpy), Open Hub (numpy), Quora topic (NumPy), and Ubuntu package (python-numpy).

External sources	
Arch package	python-numpy
Debian stable package	python-numpy
Fedora package	numpy
Free Software Directory entry	NumPy
Freebase	/m/021plb
Gentoo package	dev-python/numpy
Open Hub	numpy
Quora topic	NumPy
Ubuntu package	python-numpy

Figure 4: All external ids for NumPy

# Status of file format data in Wikidata

- **3,389** instances of file format in Wikidata today
- PRONOM has 1,553 entries: of these we have **1,208** file formats with PUID external ids
- **2,629** items connected to Just Solve the File Format Problem ids

# Links between descriptive and technical metadata

Bubble chart of software titles by number of readable file formats

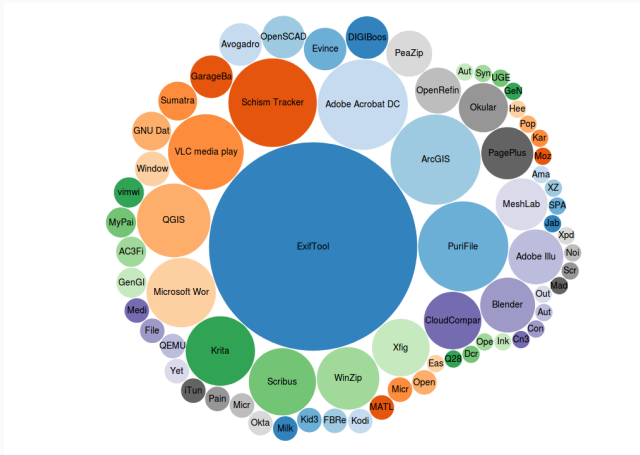


Figure 5: Try this query!

# What does LoC get from the creation of P2366 LoCFDD external id?

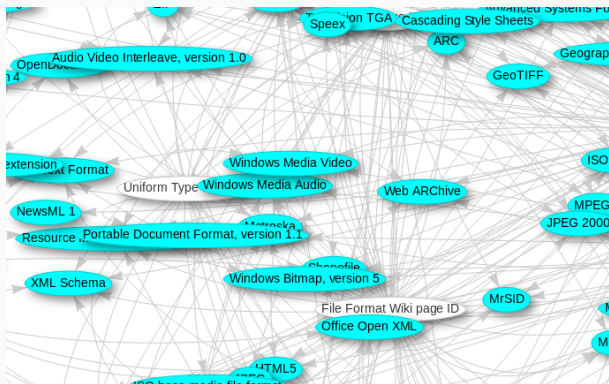


Figure 6: Try this query!



# Cross-domain knowledge base

- developers, organizations, events, places

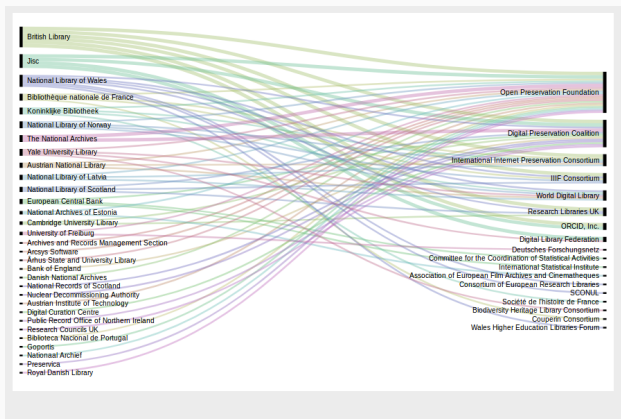


Figure 7: Orgs that are members of OPF and DPC plus other memberships

# What will we do with URIs for resources in the domain of computing?

- leverage structured data in digital preservation system
- software citation
- unambiguous identifiers for all parts of a pre-configured emulated computing environment

# Software titles for 3D graphics and file formats

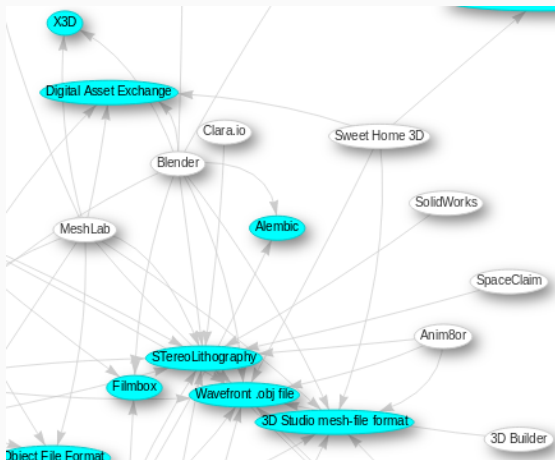


Figure 8: Try this query!

# Machine-Readable “alternative to” website powered by Wikidata

- multiple serialization formats (.ttl, .rdf, .json)
- reverse look-up tool for file formats to software
- unique Wikidata URIs so we can discuss software without confusion

## Data Curation Stats as of Oct 1, 2018

Items	Jan 18	Oct 18	net change
Software Items	64,925	85,548	20,623
File Format Items	2,834	3,405	571
File Format Items with PUIDs	777	1194	417
File Format Signatures	167	183	16
Emulators	106	115	9
File Systems	146	153	7
Device Drivers	17	27	10
Plugins	155	176	21

# Wikidata for Digital Preservation

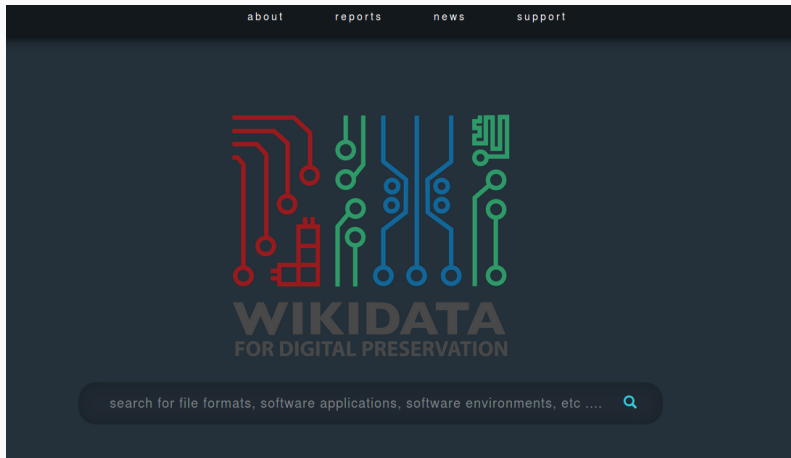


Figure 9: wikidp.org

- **Kenneth Seals-Nutt:** software engineer
- **Katherine Thornton:** data curation, data models, SPARQL queries
- **Carl Wilson:** technical mentor
- **Euan Cochrane:** digital preservation program of work

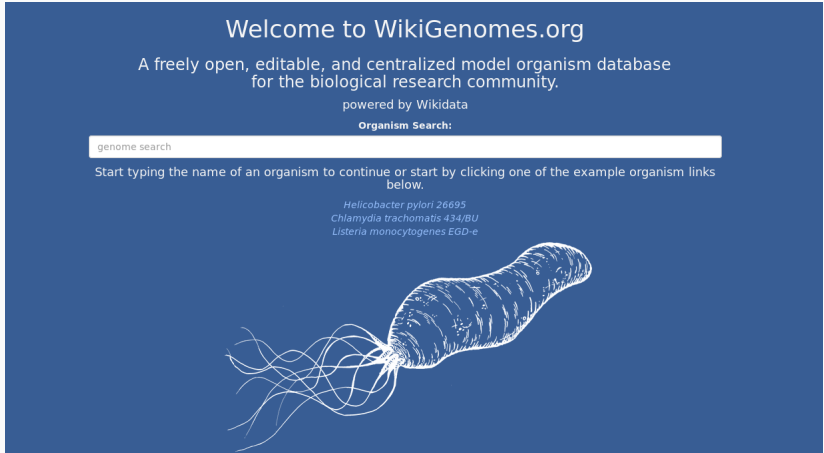


Figure 10: wikigenomes.org



- About **5,000** properties in Wikidata
- Data models are not pre-defined
- Portal has a domain-specific property checklist

- Python
- Flask
- SPARQL
- Wikidata Integrator
- Wikimedia API

The screenshot shows the WikiDP portal interface. At the top, there's a navigation bar with links for 'about', 'reports', 'news', and 'support'. The main content area is titled 'Ogg (Q188199)' and 'digital container format'. Below this, there's a section 'Current Data About Ogg' which lists several properties: 'file format (Q235557)', 'Instance Of (P31)' (digital container format (Q167772)), and 'multimedia container (Q28379876)'. Each property has a 'Reference Uri (P854)' and a 'Retrieved (P613)' date. To the left of the main content is a sidebar with a 'selected item' dropdown set to 'Ogg (Q188199)' and a 'property checklist' with various checkboxes. To the right is a section titled 'other details' which includes 'Aliases' (.ogg), 'Description' (digital container format), 'External Links' (File Format Wiki Page Id (P3381) pointing to Ogg), 'Freebase Id (P646)' (/m/05mxd), and 'Locfdd Id (P3266)' (idd000026). At the bottom, there's a large 'Logo Image (P154)' showing the Ogg logo.

Figure 11: Screenshot of search results in the WikiDP portal

- Custom search
- View data from Wikidata
- Contribute data from Wikidata
- Property checklists

- Wikidata
- WikiDP
- reuse metadata from within EaaS

- Getting started with Wikidata video
- Wikidata Editing Cheatsheet
- Tours of the Wikidata site
- WikiProject Informatics

- OCLC Hanging Together
- ZBW
- ODNB

- Github repository for source code
- SPARQL queries to reuse Wikidata data
- Kat's tweets full of SPARQL-y data from Wikidata @wikidigi



Thank you!

katherine.thornton@yale.edu



# Acknowledgements

- Andrew W. Mellon Foundation
- Alfred P. Sloan Foundation
- Software Preservation Network
- Open Preservation Foundation
- Wikidata community