Modeling Software and File Formats

ShEx for data models

Kat Thornton

Yale University

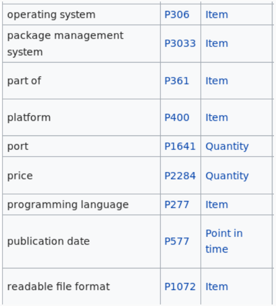
- We have talked about models for funded research projects
- We may also need to model the publications created in the context of research projects

- What software resources have been created in the context of ERC sponsored projects?
- What workflow pipelines have been created?
- What types of existing software are reused in ERC sponsored projects?
- Can I browse a list of available software pipelines and the licenses under which they are released?
- What databases has the ERC funded?

- What Journals do people who do research about openAIRE publish in?
- How many items have zenodo ids on Wikidata?

Status of software data in Wikidata

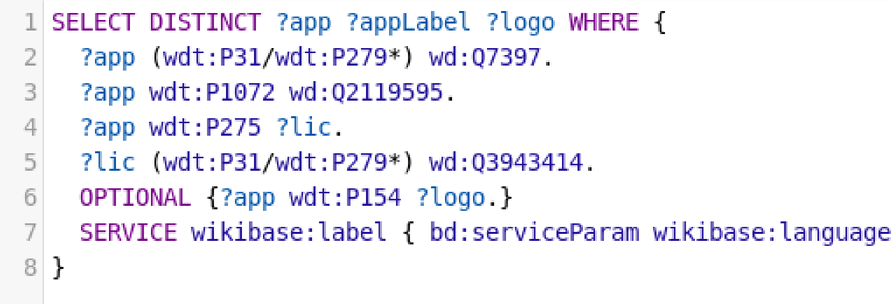
- **79,500** instances of software in Wikidata today
- FSF external ids for **1,428** software items (15,000+ resources total)
- Framalibre external ids for **336** software items



operating system	P306	Item
package management system	P3033	Item
part of	P361	Item
platform	P400	Item
port	P1641	Quantity
price	P2284	Quantity
programming language	P277	Item
publication date	P577	Point in time
readable file format	P1072	Item

Figure 1: Some of the properties used for software.

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

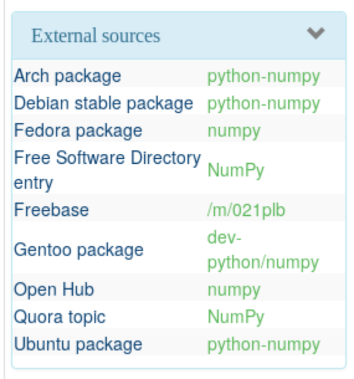


```
SELECT DISTINCT ?app ?appLabel ?logo WHERE {
  ?app (wdt:P31/wdt:P279*) wd:Q7397.
  ?app wdt:P1072 wd:Q2119595.
  ?app wdt:P275 ?lic.
  ?lic (wdt:P31/wdt:P279*) wd:Q3943414.
  OPTIONAL {?app wdt:P154 ?logo.}
  SERVICE wikibase:label { bd:serviceParam wikibase:language
}
```

Figure 2: 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



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 3: All external ids for NumPy

Status of file format data in Wikidata

- **2,852** instances of file format in Wikidata today
- PRONOM has 1,553 entries: of these we have **1,135** 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

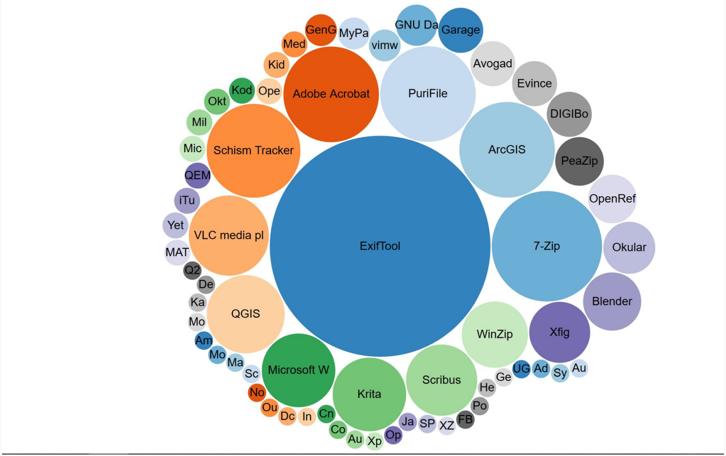


Figure 4: Try this query!

ShapeExpressions Uses

- **Legacy review** – develop punch lists for existing data issues that need fixing
- **Client pre-submission** – submitters test their data before submission to make sure they're saying what they want to say and that the receiving schema can accommodate all of their data
- **Server pre-ingestion** – submission process checks data as it comes in and either rejects or warns about non-conformant data
- **Generating** user interfaces
- **Transforming** RDF data into other data formats

Using ShEx to Validate Entity Data from Wikidata

- ShEx schema for file format
- Validating file format data

Learning More about ShEx

- ShEx homepage
- ShEx Primer
- Scholia profile for ShEx

Thank you!

katherine.thornton@yale.edu
@wikidigi