Pathways for Discovery of Free Software

Katherine Thornton, Morane Gruenpeter

Wikidata for Digital Preservation
katherine.thornton@yale.edu, morane@softwareheritage.org

25 March 2018

WIKIDATA
FOR DIGITAL PRESERVATION

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduce software preservation as the key to software discovery
2. Describe and document software through metadata
3. Explore the landscape of software ontologies and vocabularies
4. Discover free software in Wikidata

Wikidata for Digital preservation working group

We are...

- cultural heritage technologists with a mission
- metadata enthusiasts
- free software advocates

Wikidata for Digital preservation working group

We are...

- cultural heritage technologists with a mission
- metadata enthusiasts
- free software advocates

Goals

- document digital artifacts, software and software source code
- promote open standards and libre community vision
- contribute metadata for software preservation

Software Preservation

Many cultural heritage organizations have software in their collections

What do we want/need to preserve ?

- software binaries
- software source code
- hardware

Software Preservation

Many cultural heritage organizations have software in their collections

What do we want/need to preserve ?

- software binaries
- software source code
- hardware

is this enough ?

Software Preservation

Many cultural heritage organizations have software in their collections

What do we want/need to preserve ?

- software binaries
- software source code
- hardware

is this enough ?

What are the risks preserving software without the context?

- Sometimes different software resources have the same name
- Software description practices are variable-
  - without the compatible environment metadata
  - lack information necessary for reproducibility

if we preserve information about environments we can emulate or virtualize

Why now

Looking at the past

- a lot of old software misplaced, lost, or behind barriers, but…
- most legacy founders and the current maintainers are still here, and willing to share
- urgent to collect their knowledge

Only a few years left.

Why now

Looking at the past

- a lot of old software misplaced, lost, or behind barriers, but...
- most legacy founders and the current maintainers are still here, and willing to share
- urgent to collect their knowledge

Only a few years left.

Looking at the future

- software development skyrockets
- essential to preserve the software in its context for the future

Every year that goes by makes the problem worse.

What are the limitations today with software preservation and discovery?

Sourceforge
Debian
CPAN
Gitorious
Maven
Bitbucket
Inria
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

Software Heritage in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

Preserving our heritage, enabling better software for all

Software Heritage in a nutshell

Software Heritage

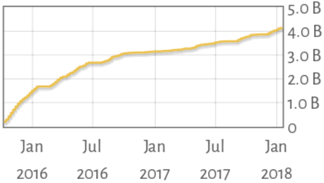
THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

Preserving our heritage, enabling better software for all

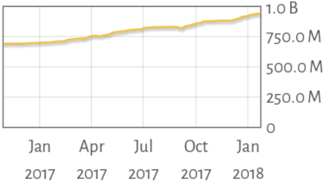
Source files

4,130,492,226



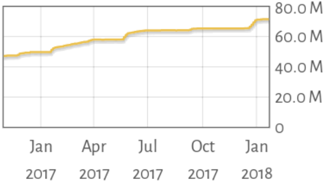
Commits

943,061,517



Projects

71,814,787



Outline

1. Introduce software preservation as the key to software discovery

2. Describe and document software through metadata

3. Explore the landscape of software ontologies and vocabularies

4. Discover free software in Wikidata

Describe and document software

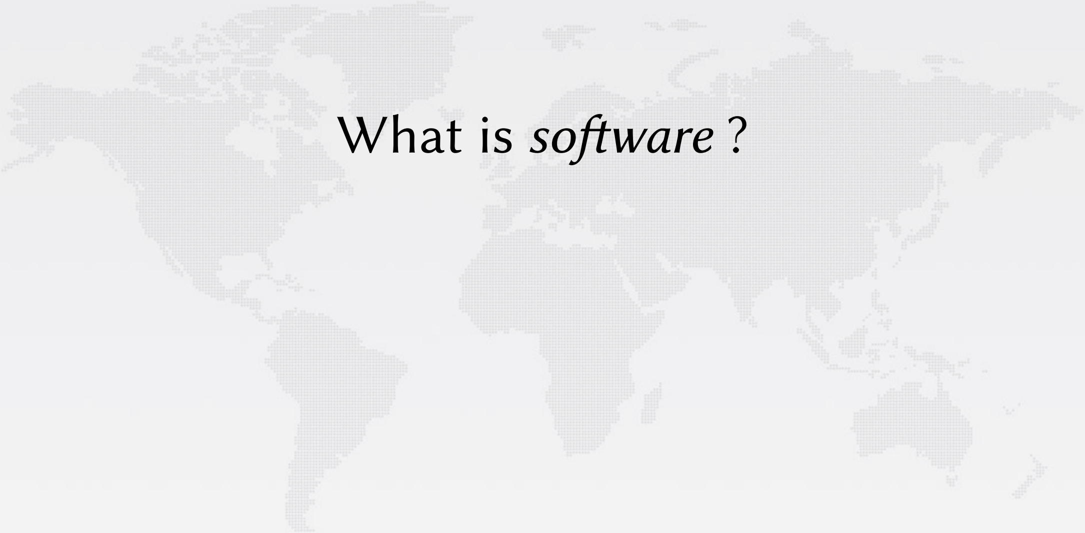
Why is it important?

without description and documentation these resources can't be located, reused, extended, etc

Use cases

- unique identification
- software reproducibility
- browse *source code* with context information
- software citation - cite and be cited
- semantic search: find software by author, version, keywords

Describe and document software



What is *software* ?

Describe and document software

What is *software* ?

Software as a concept

- project or entity
- the community around the project
- the software idea/algorithms/solutions

Describe and document software

What is *software* ?

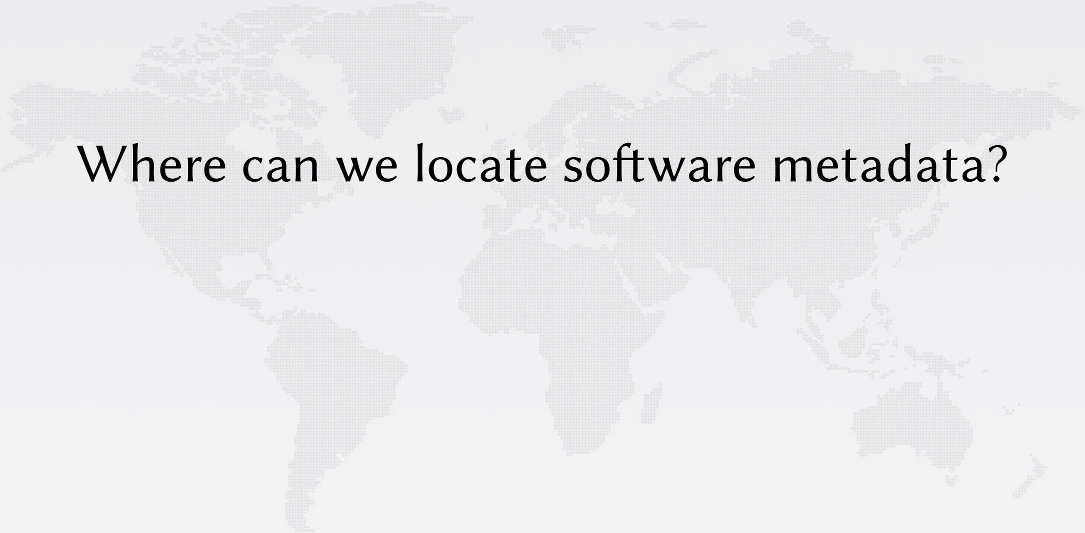
Software as a concept

- project or entity
- the community around the project
- the software idea/algorithms/solutions

Software artifact

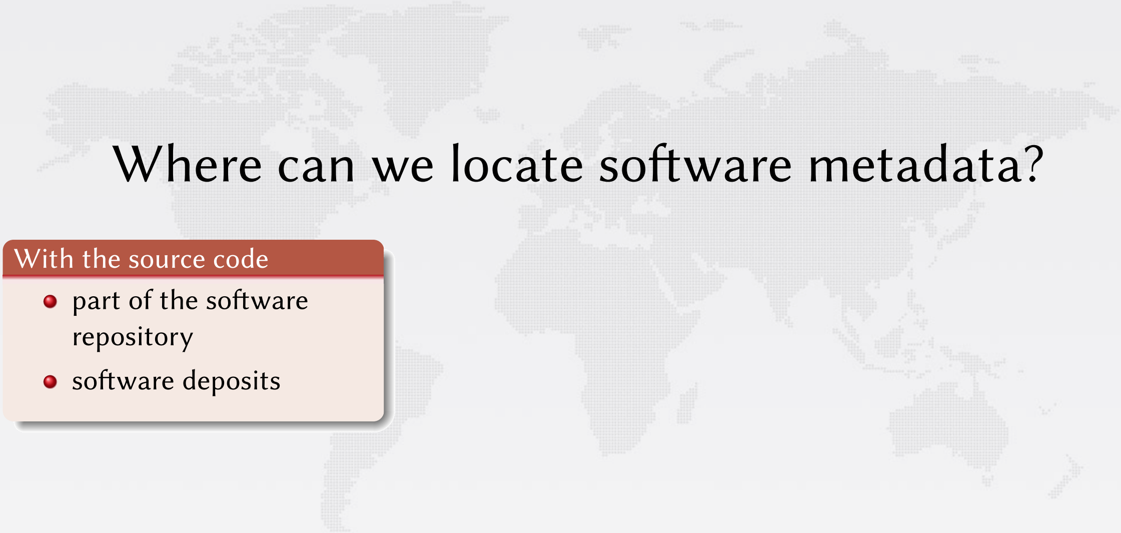
- each revision in source code form
- binaries produced for different environments

Describe and document software



Where can we locate software metadata?

Describe and document software



Where can we locate software metadata?

With the source code

- part of the software repository
- software deposits

Describe and document software

Where can we locate software metadata?

With the source code

- part of the software repository
- software deposits

In the Source code

- package management
- CodeMeta.json file (for citation)

Describe and document software

Where can we locate software metadata?

With the source code

- part of the software repository
- software deposits

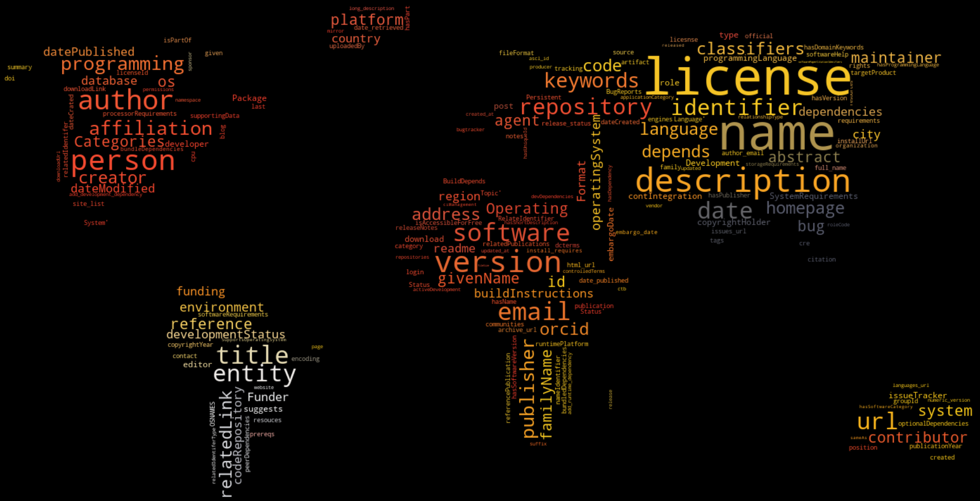
In the Source code

- package management
- CodeMeta.json file (for citation)

Registries/Catalogs

- Wikidata
- FSF-directory
- libraries.io

With what terms should we describe a *software artifact*?



* Generated with wordcloud library in Python using CodeMeta's crosswalk tabe

Outline

1. Introduce software preservation as the key to software discovery
2. Describe and document software through metadata
3. Explore the landscape of software ontologies and vocabularies
4. Discover free software in Wikidata

Software ontologies and vocabularies

“Ontologies are agreements, made in a social context,
to accomplish some objectives.
It’s important to understand those objectives, and be guided by them."

T. Gruber, The Pragmatics of Ontology, 2003

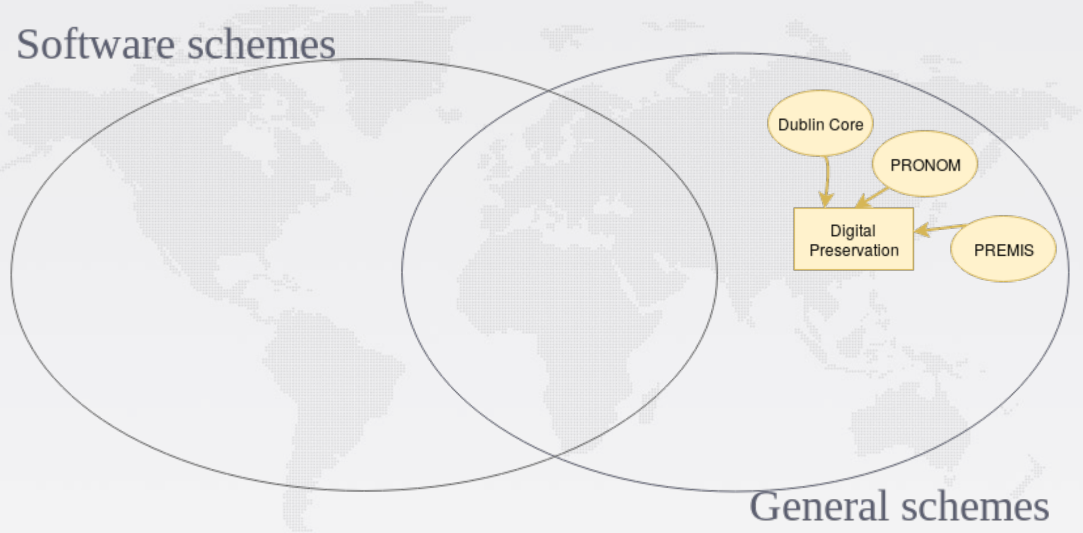
Software ontologies and vocabularies

"Ontologies are agreements, made in a social context,
to accomplish some objectives.
It's important to understand those objectives, and be guided by them."
T. Gruber, The Pragmatics of Ontology, 2003

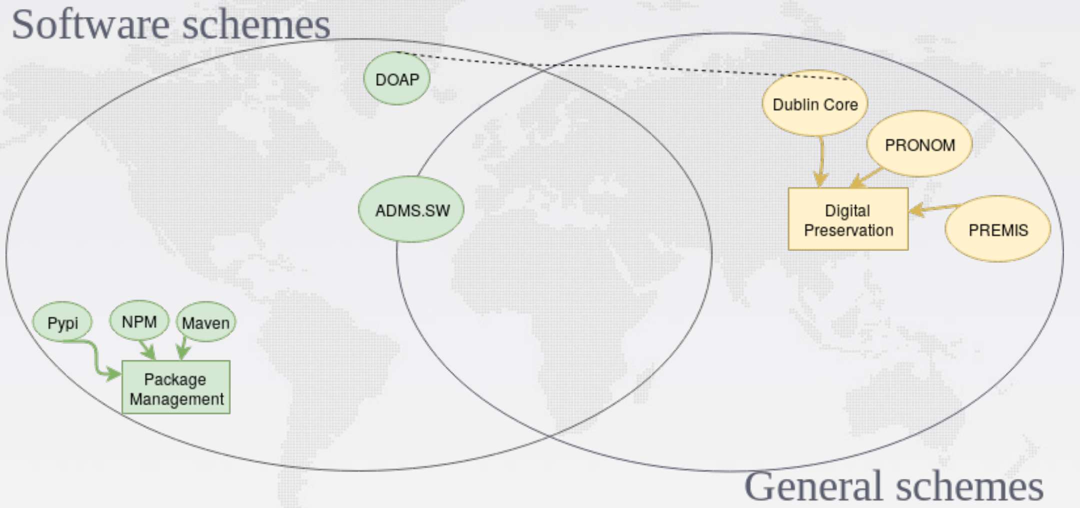
LOV- Linked open vocabularies

"Vocabularies provide the semantic glue enabling data to become meaningful data. "

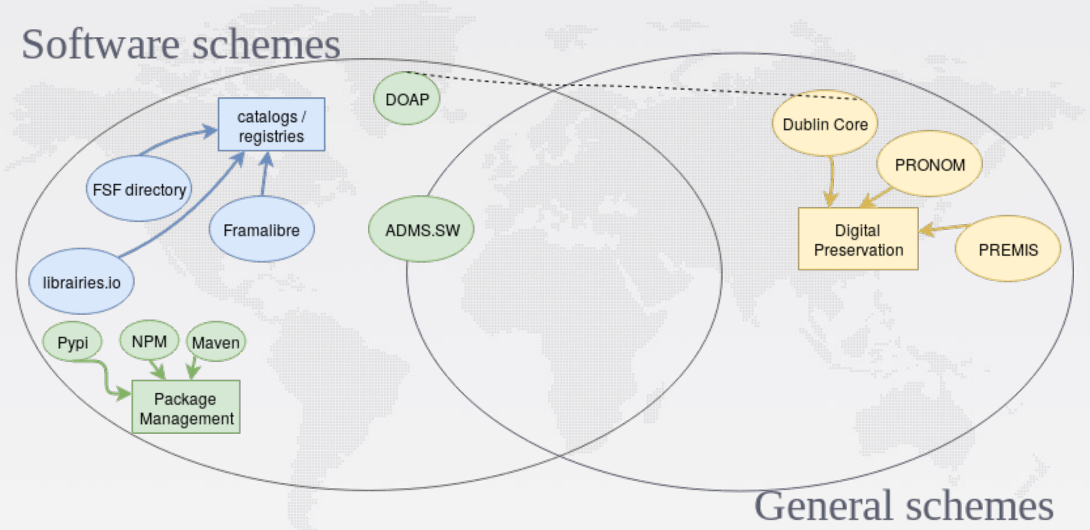
The landscape of software ontologies



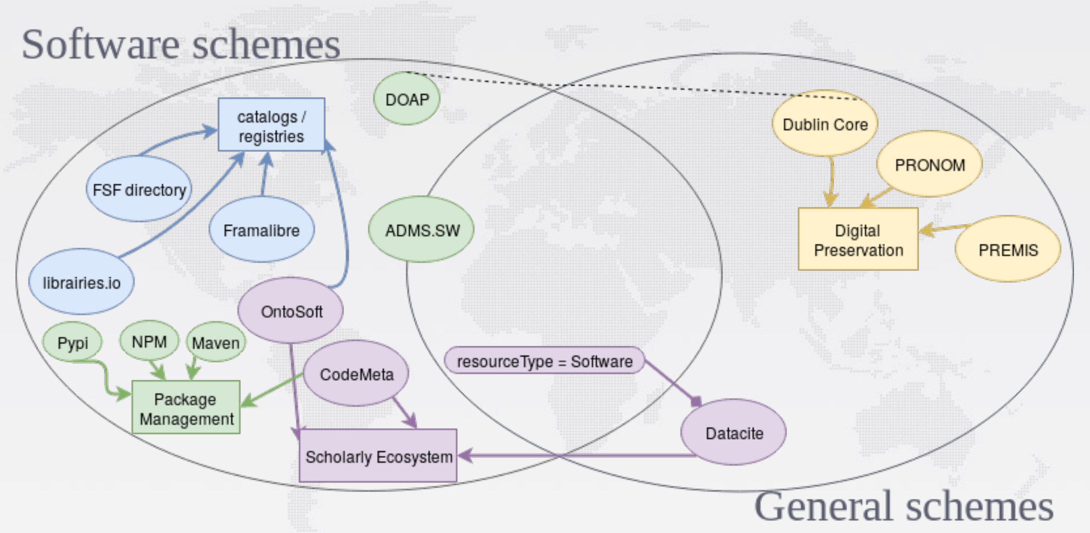
The landscape of software ontologies



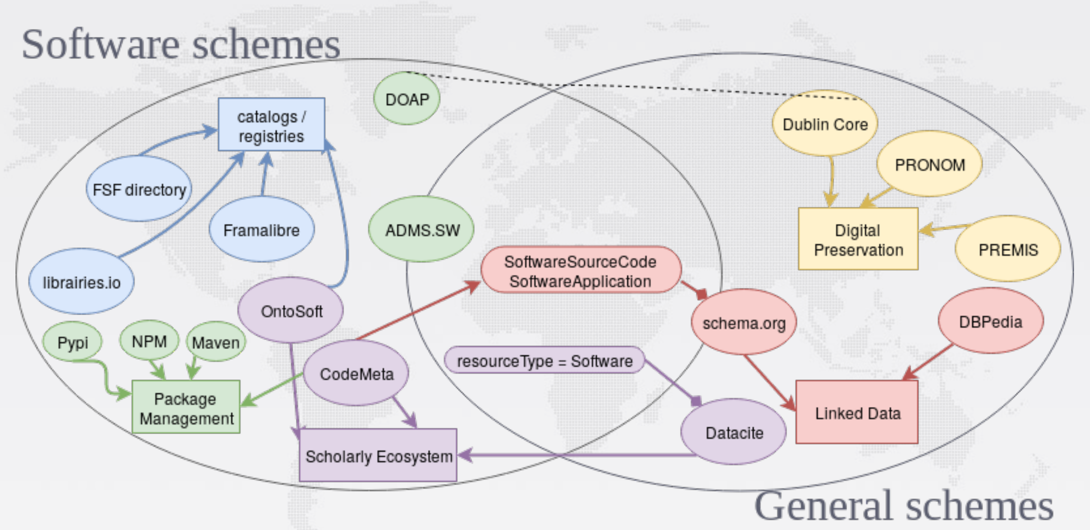
The landscape of software ontologies



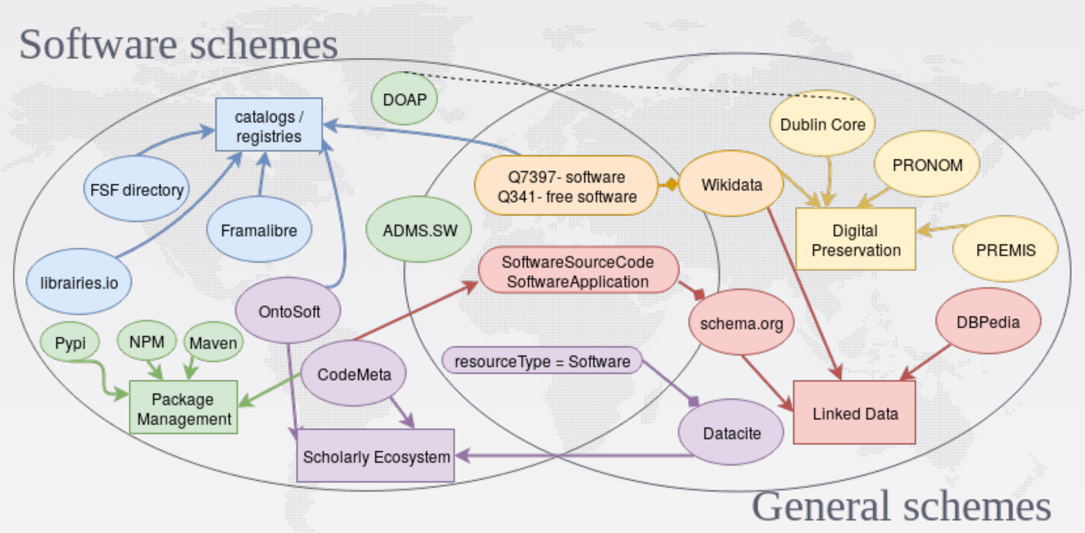
The landscape of software ontologies



The landscape of software ontologies



The landscape of software ontologies

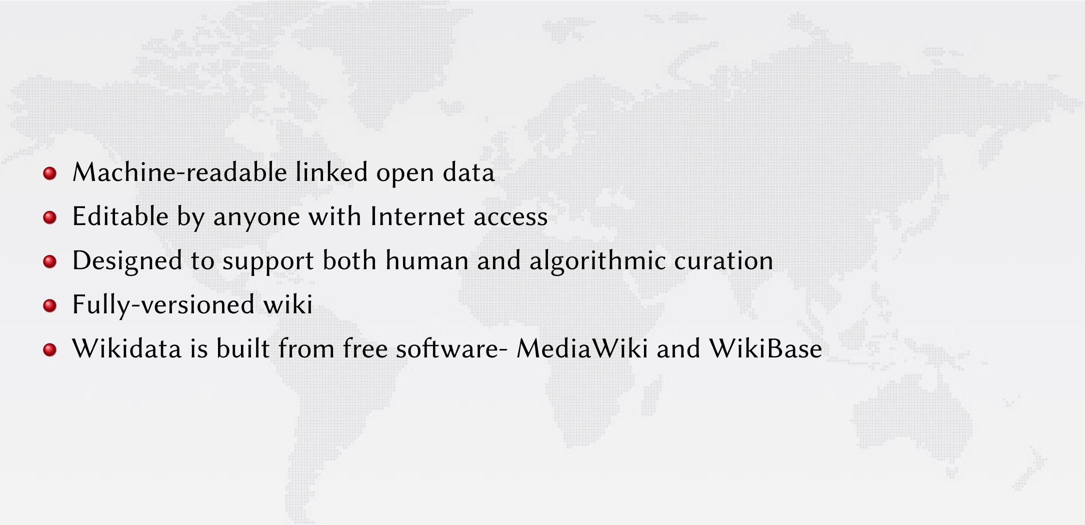


Outline

1. Introduce software preservation as the key to software discovery
2. Describe and document software through metadata
3. Explore the landscape of software ontologies and vocabularies
4. Discover free software in Wikidata

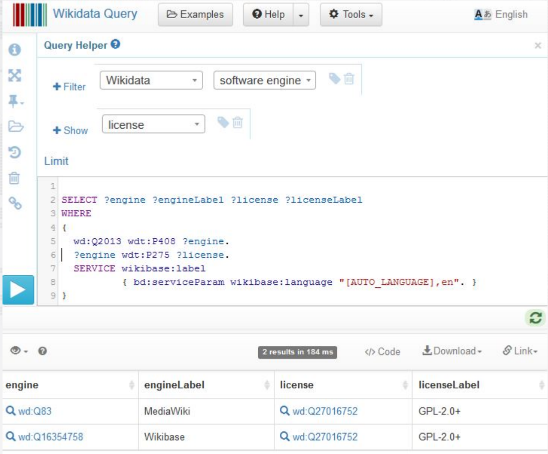
Wikidata

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
- Wikidata is built from free software- MediaWiki and WikiBase

SPARQL query for the software licenses of the software that powers Wikidata



Filter: Wikidata, software engine

Show: license

Limit

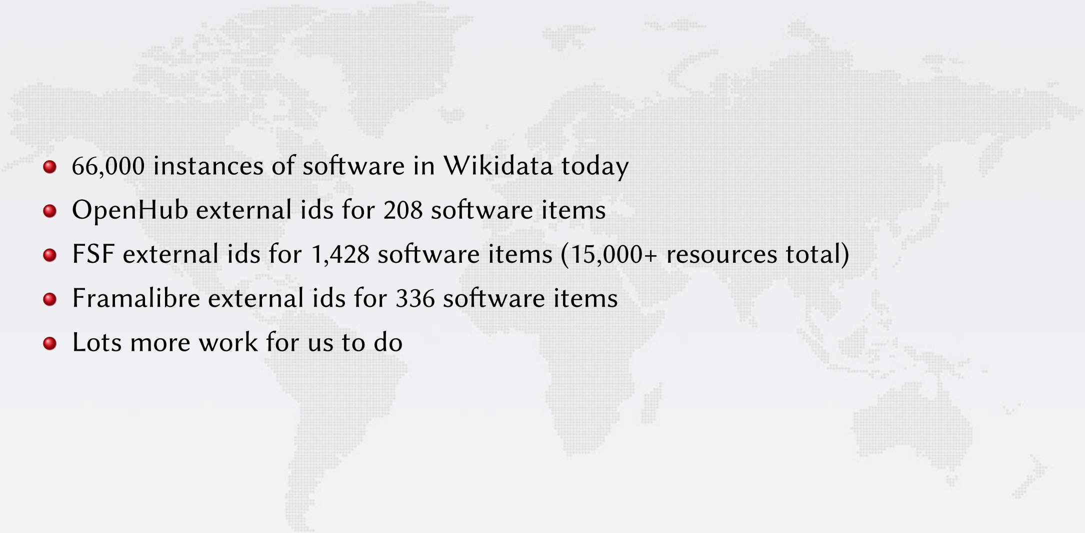
```
SELECT ?engine ?engineLabel ?license ?licenseLabel
WHERE
{
  wd:Q2013 wdt:P408 ?engine.
  ?engine wdt:P275 ?license.
  SERVICE wikibase:label
          { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

2 results in 184 ms

engine	engineLabel	license	licenseLabel
wd:Q83	MediaWiki	wd:Q27016752	GPL-2.0+
wd:Q16354758	Wikibase	wd:Q27016752	GPL-2.0+

Figure: Try this query!

Status of software data in Wikidata



- 66,000 instances of software in Wikidata today
- OpenHub external ids for 208 software items
- FSF external ids for 1,428 software items (15,000+ resources total)
- Framalibre external ids for 336 software items
- Lots more work for us to do

Free software in Wikidata

A bubble chart of licenses by number of software titles

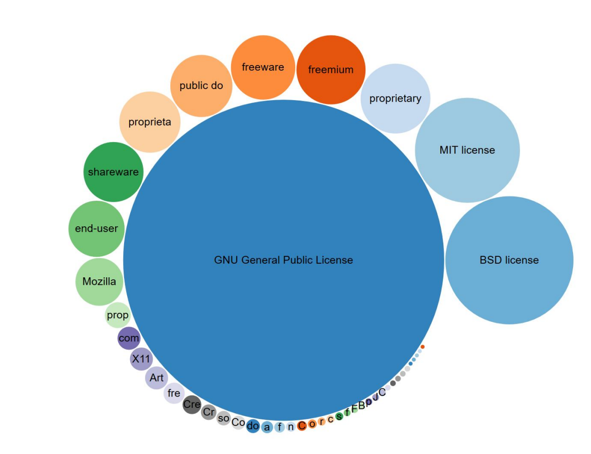
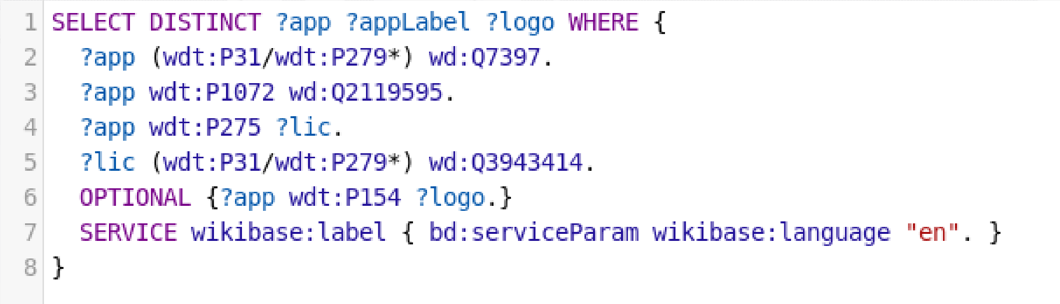


Figure: Try this query!

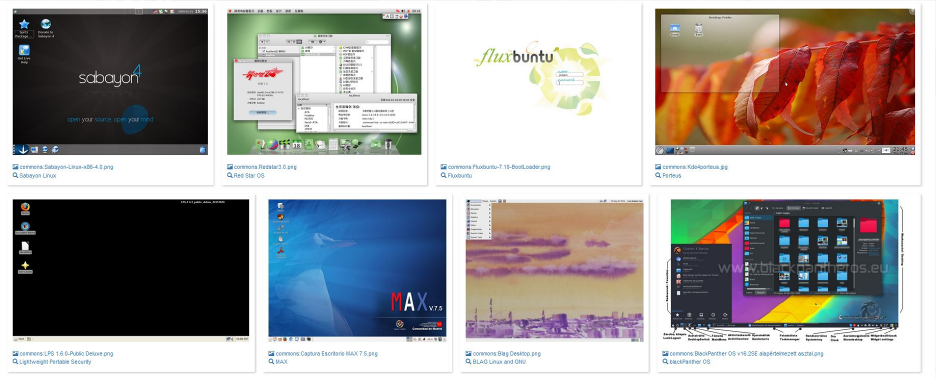
What software available under a free software license can I use to open .obj files?



```
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 "en". }
}
```

Figure: Try this query!

Create an image grid of Gnu/Linux distributions



commons:Sabayon-Linux-x86-4.0.png
Sabayon Linux

commons:Redstar3.0.png
Red Star OS

commons:Fluxbuntu-7.10-BootLoader.png
Fluxbuntu

commons:Kde4porteus.jpg
Porteus

commons:LPS 1.6.0-Public Deluxe.png
Lightweight Portable Security

commons:Captura Escritorio MAX 7.5.png
MAX

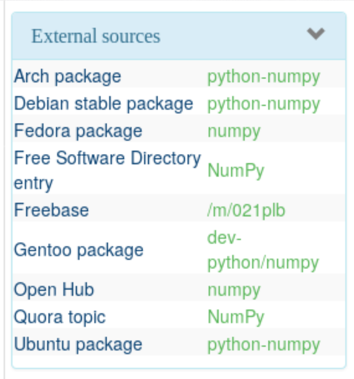
commons:Blag Desktop.png
BLAG Linux and GNU

commons:BlackPanther OS v16.2SE alapértelmezett asztal.png
blackPanther OS

Figure: 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



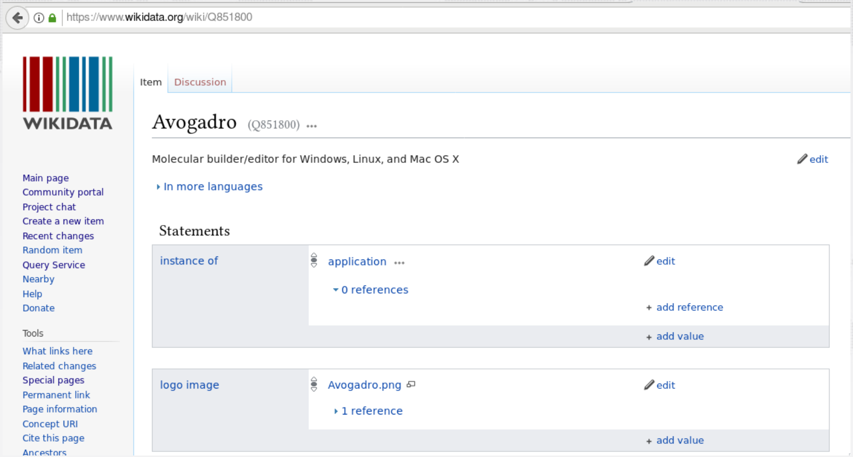
External sources

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: All external ids for NumPy

FSF Resource Directory external ID in Wikidata

- If a person or software agent visits the Wikidata item for a piece of software that is also in the FSF Resource Directory, they will find a URL to the page on the FSF wiki.



https://www.wikidata.org/wiki/Q851800

Avogadro (Q851800)

Molecular builder/editor for Windows, Linux, and Mac OS X

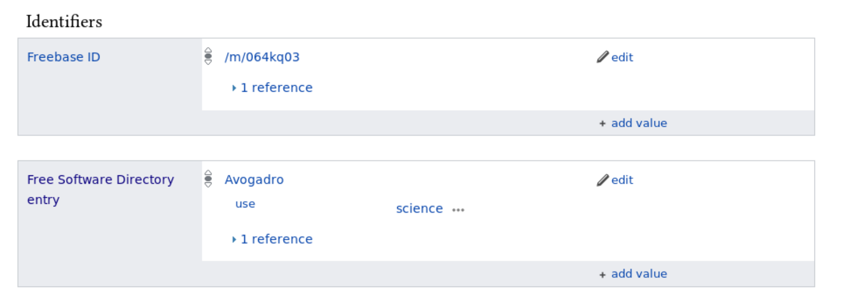
Statements

instance of: application

logo image: Avogadro.png

Scroll down to the bottom of the page to see the identifiers

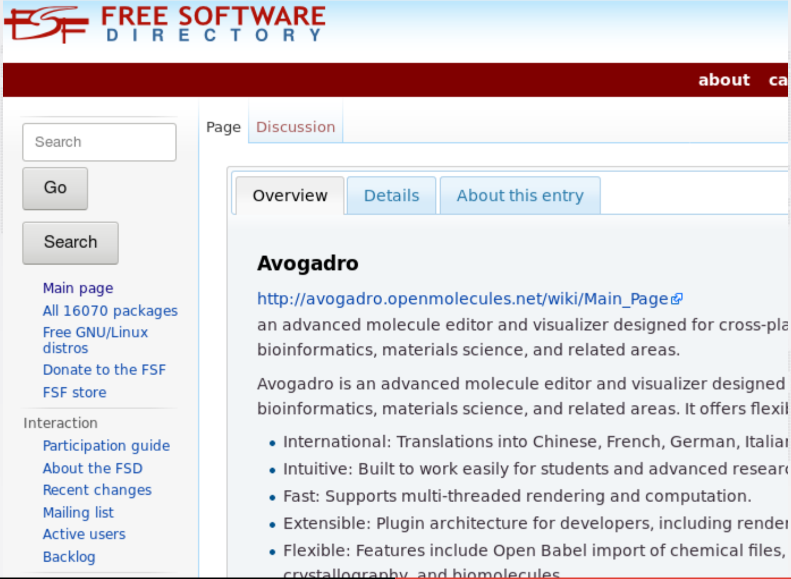
click on the link next to Free Software Directory ID



Identifiers

Freebase ID	/m/064kq03 1 reference
Free Software Directory entry	Avogadro use science 1 reference

Here is the wiki page for Avogadro in the FSF Resource Directory



Avogadro

http://avogadro.openmolecules.net/wiki/Main_Page

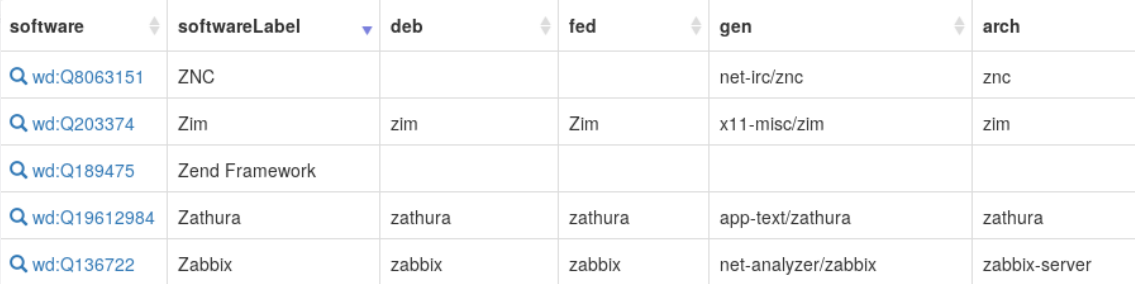
an advanced molecule editor and visualizer designed for cross-pla
bioinformatics, materials science, and related areas.

Avogadro is an advanced molecule editor and visualizer designed
bioinformatics, materials science, and related areas. It offers flexil

- International: Translations into Chinese, French, German, Italia
- Intuitive: Built to work easily for students and advanced researc
- Fast: Supports multi-threaded rendering and computation.
- Extensible: Plugin architecture for developers, including rende
- Flexible: Features include Open Babel import of chemical files, crystallography, and biomolecules

Wikidata crosswalks multiple external IDs

- We can write queries to return lots of different identifiers for software.



software	softwareLabel	deb	fed	gen	arch
wd:Q8063151	ZNC			net-irc/znc	znc
wd:Q203374	Zim	zim	Zim	x11-misc/zim	zim
wd:Q189475	Zend Framework				
wd:Q19612984	Zathura	zathura	zathura	app-text/zathura	zathura
wd:Q136722	Zabbix	zabbix	zabbix	net-analyzer/zabbix	zabbix-server

Figure: Try this query!

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 1,023 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

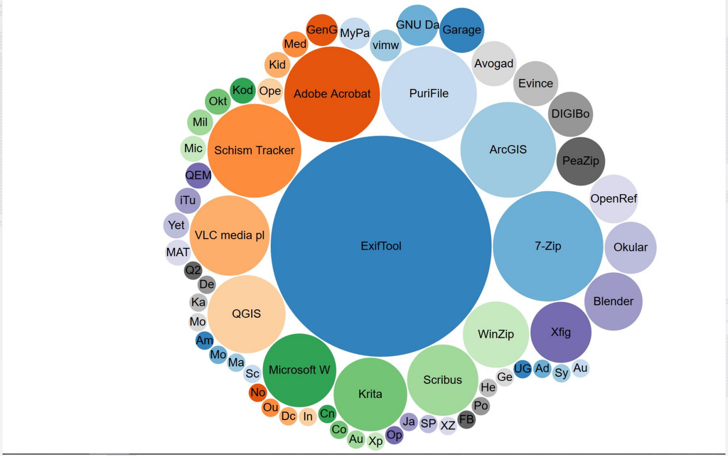
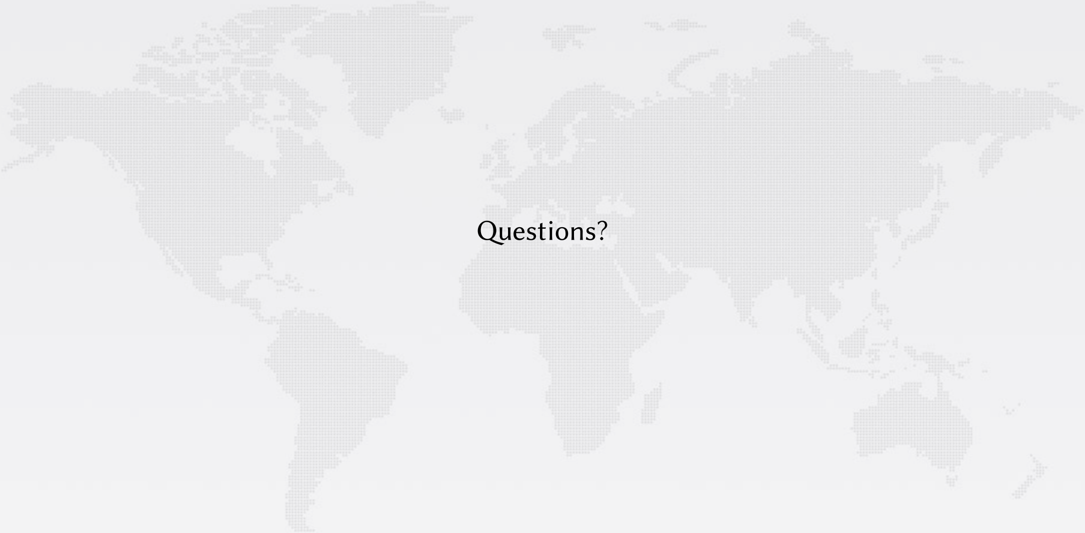


Figure: Try this query!

Machine-Readable "alternative to" website powered by Wikidata

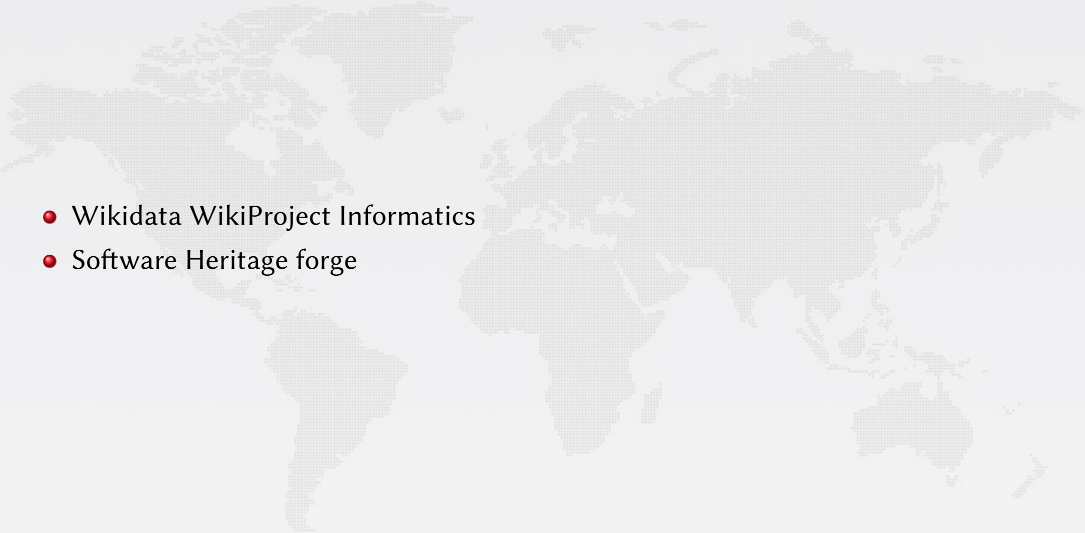
- multiple serialization formats (.ttl, .rdf, .json)
- reverse look-up tool for file formats to software
- unique Wikidata URIs so we can discuss software without confusion

Thank You!



Questions?

Want to contribute?



- Wikidata WikiProject Informatics
- Software Heritage forge

Acknowledgements

Wikidata for Digital Preservation working group

- Kat
- Morane
- Carl Wilson, Open Preservation Foundation
- Thomas Ledoux, National Library of France
- Bertrand Caron, National Library of France
- Ross Spencer, Artefactual Systems
- John Samuel, École Supérieure de Chimie Physique Électronique de Lyon
- David Russo, British Library

Acknowledgements

Communities

- Wikidata community
- Software Heritage
- Wikicite
- Yale University Library
- Council on Library and Information Resources
- Crossminer