

Pathways for Discovery of Free Software

Katherine Thornton, Morane Gruenpeter

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

25 March 2018



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

- 
- 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

We are...

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

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**

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 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?





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

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

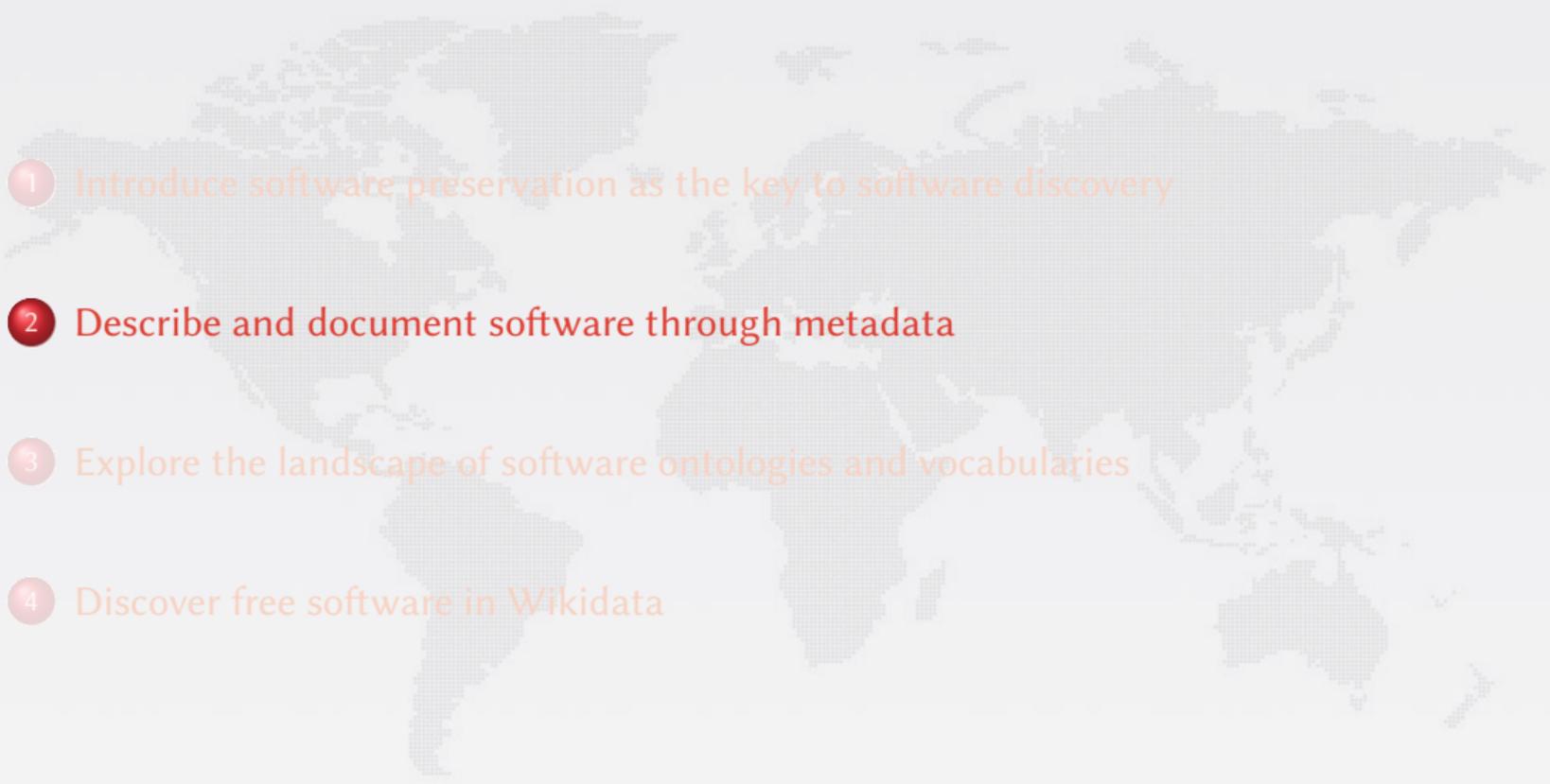
943,061,517



Projects

71,814,787



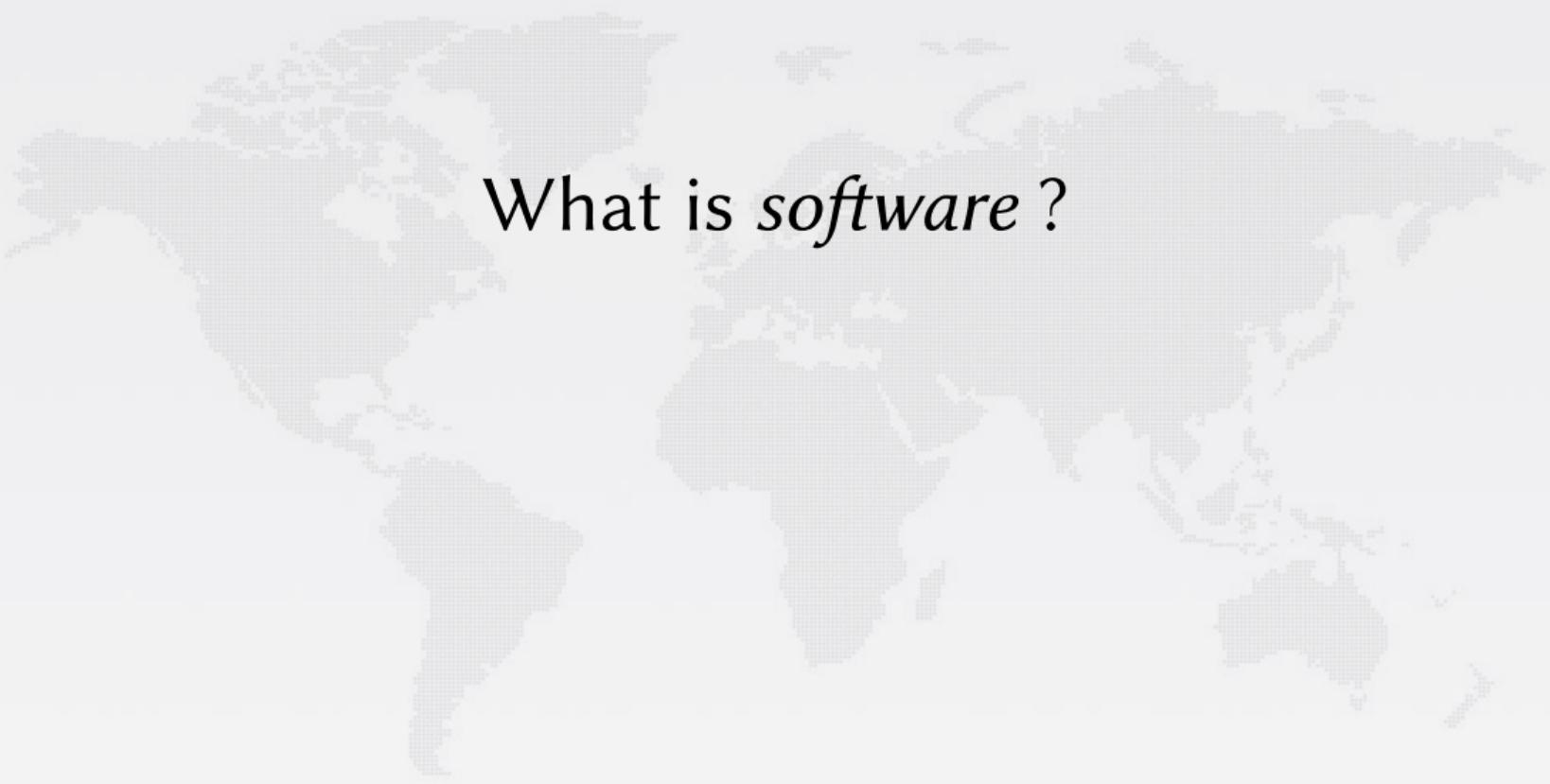
- 
- 1 Introduce software preservation as the key to software discovery
 - 2 Describe and document software through metadata
 - 3 Explore the landscape of software ontologies and vocabularies
 - 4 Discover free software in Wikidata

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



What is *software* ?

What is *software* ?

Software as a concept

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

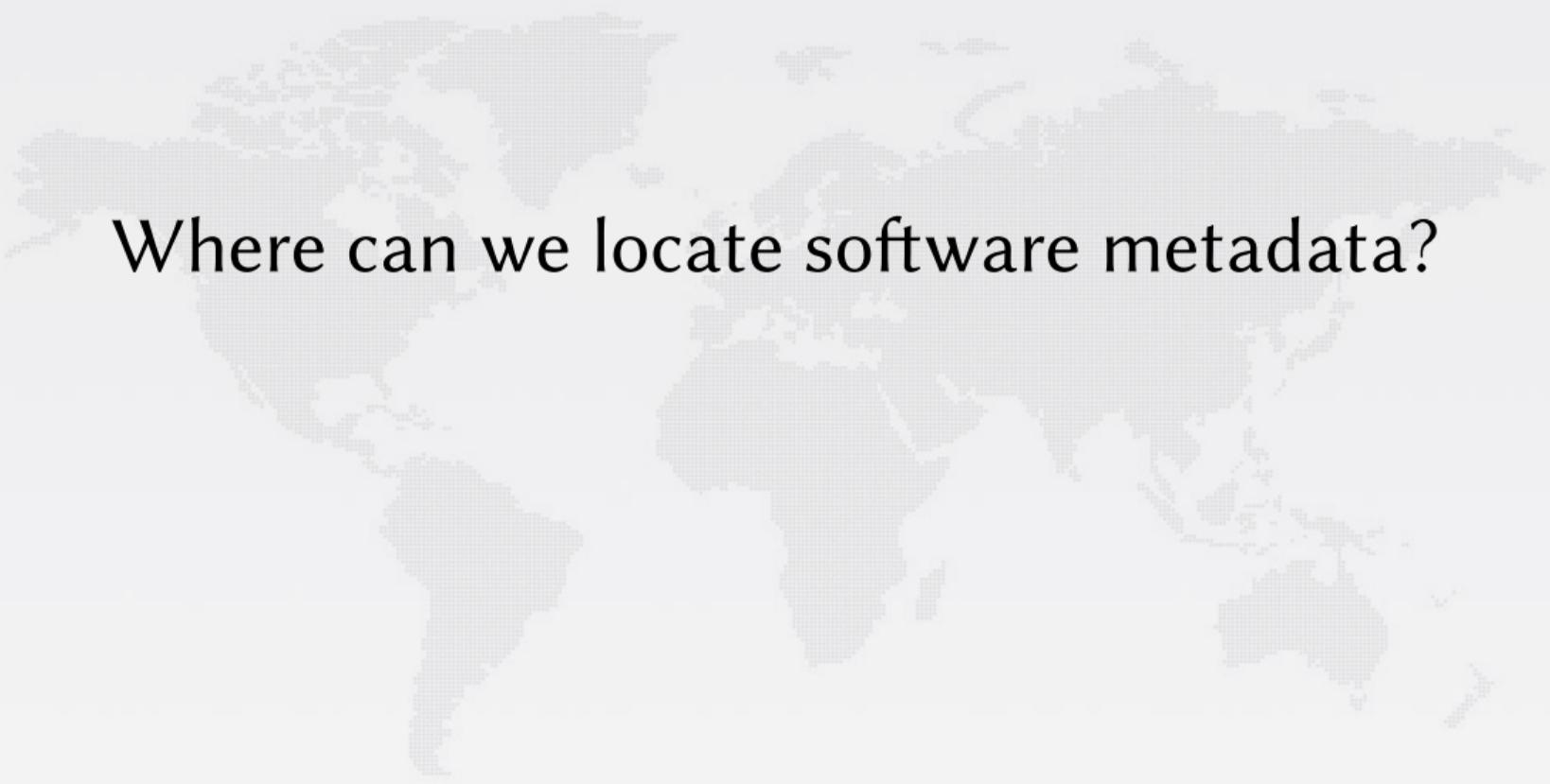
What is *software* ?

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



Where can we locate software metadata?

Where can we locate software metadata?

With the source code

- part of the software repository
- software deposits

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)

Where can we locate software metadata?

With the source code

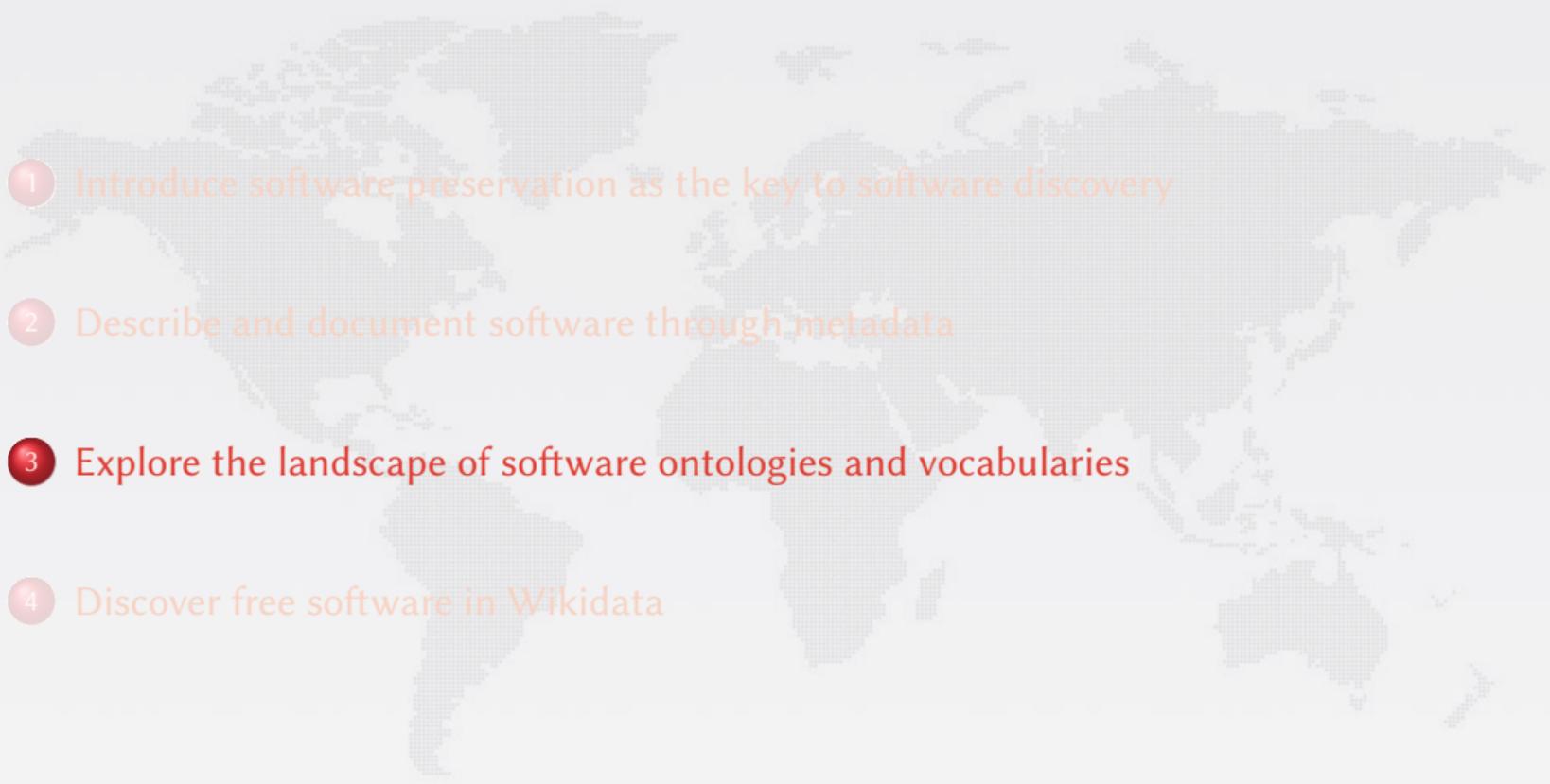
- part of the software repository
- software deposits

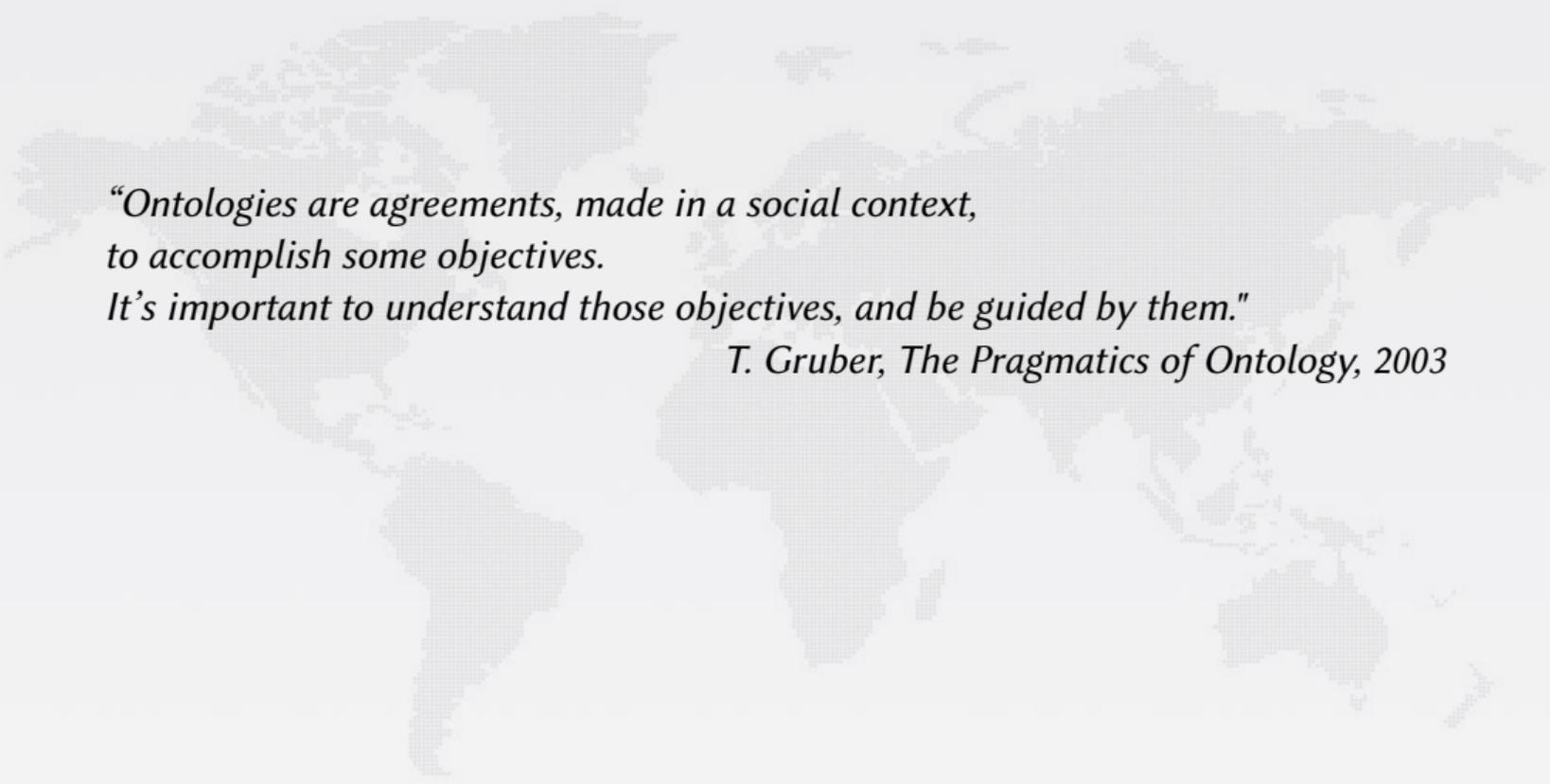
In the Source code

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

Registries/Catalogs

- Wikidata
- FSF-directory
- libraries.io

- 
- 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



*“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

*“Ontologies are agreements, made in a social context,
to accomplish some objectives.*

It’s important to understand those objectives, and be guided by them.”

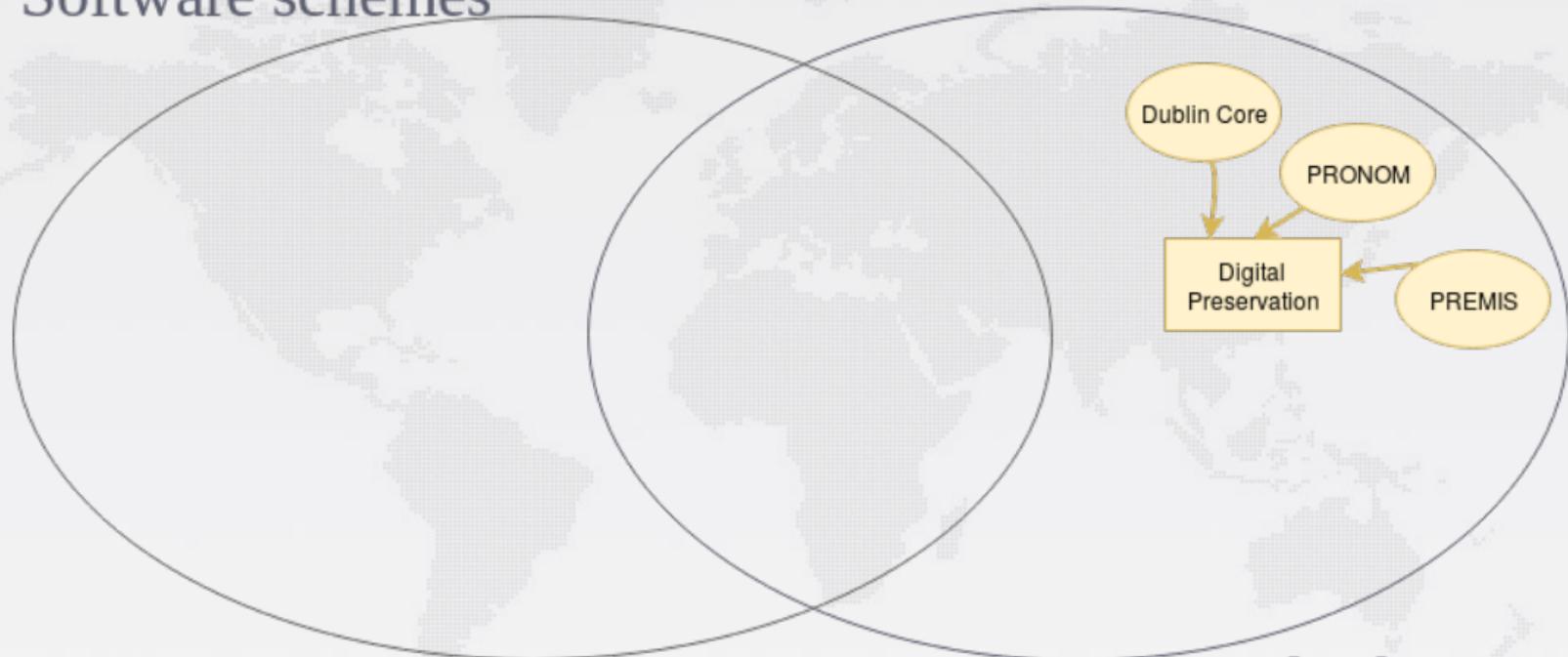
T. Gruber, The Pragmatics of Ontology, 2003

LOV- Linked open vocabularies

*“Vocabularies provide the **semantic glue** enabling data to become **meaningful data**. ”*

The landscape of software ontologies

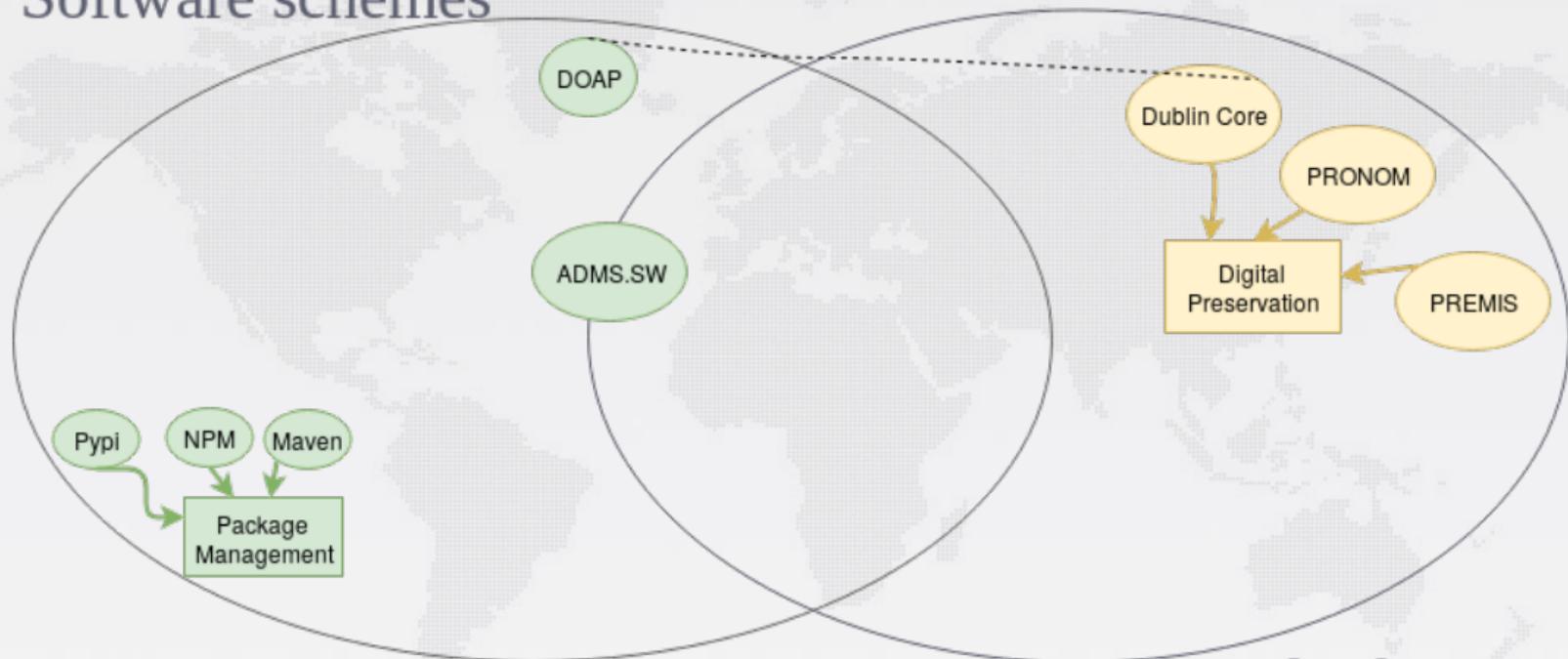
Software schemes



General schemes

The landscape of software ontologies

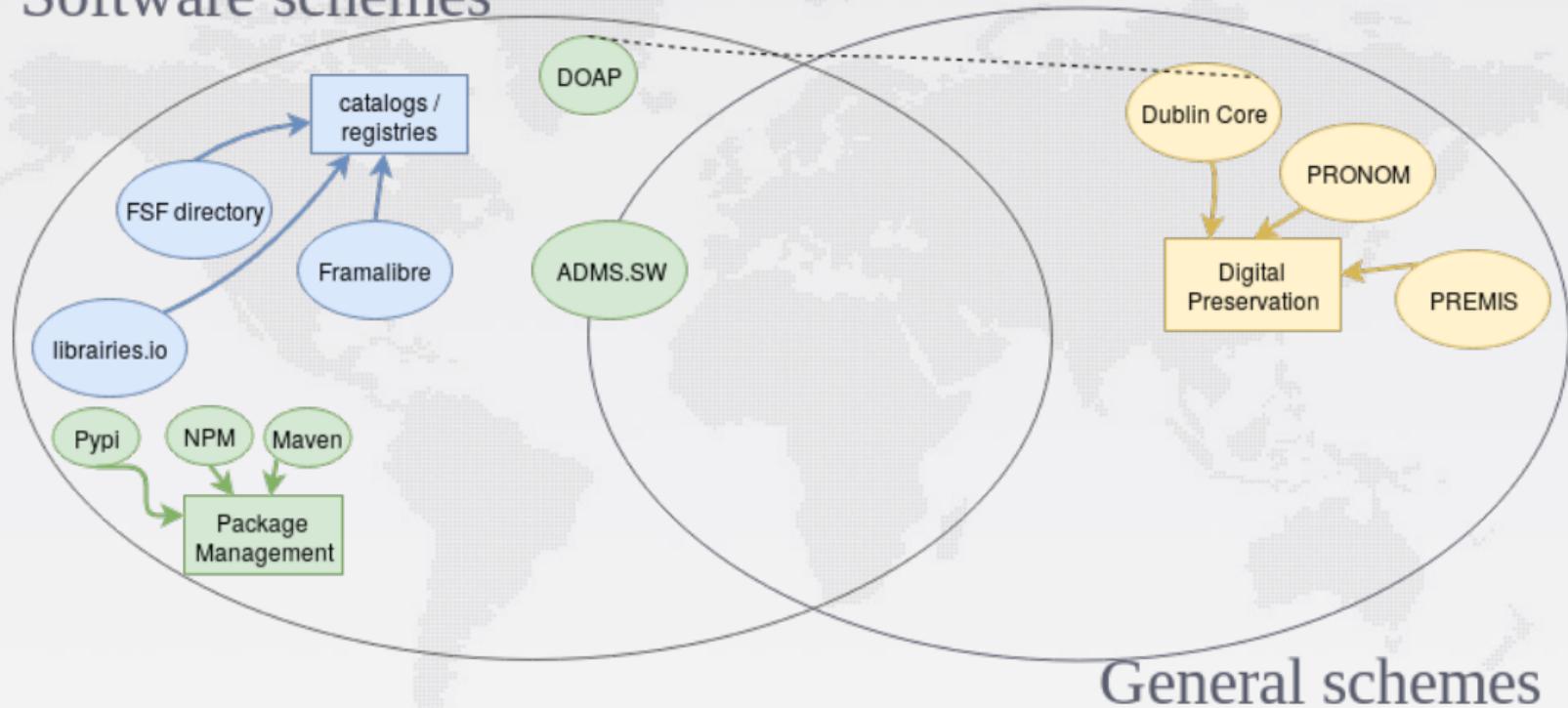
Software schemes



General schemes

The landscape of software ontologies

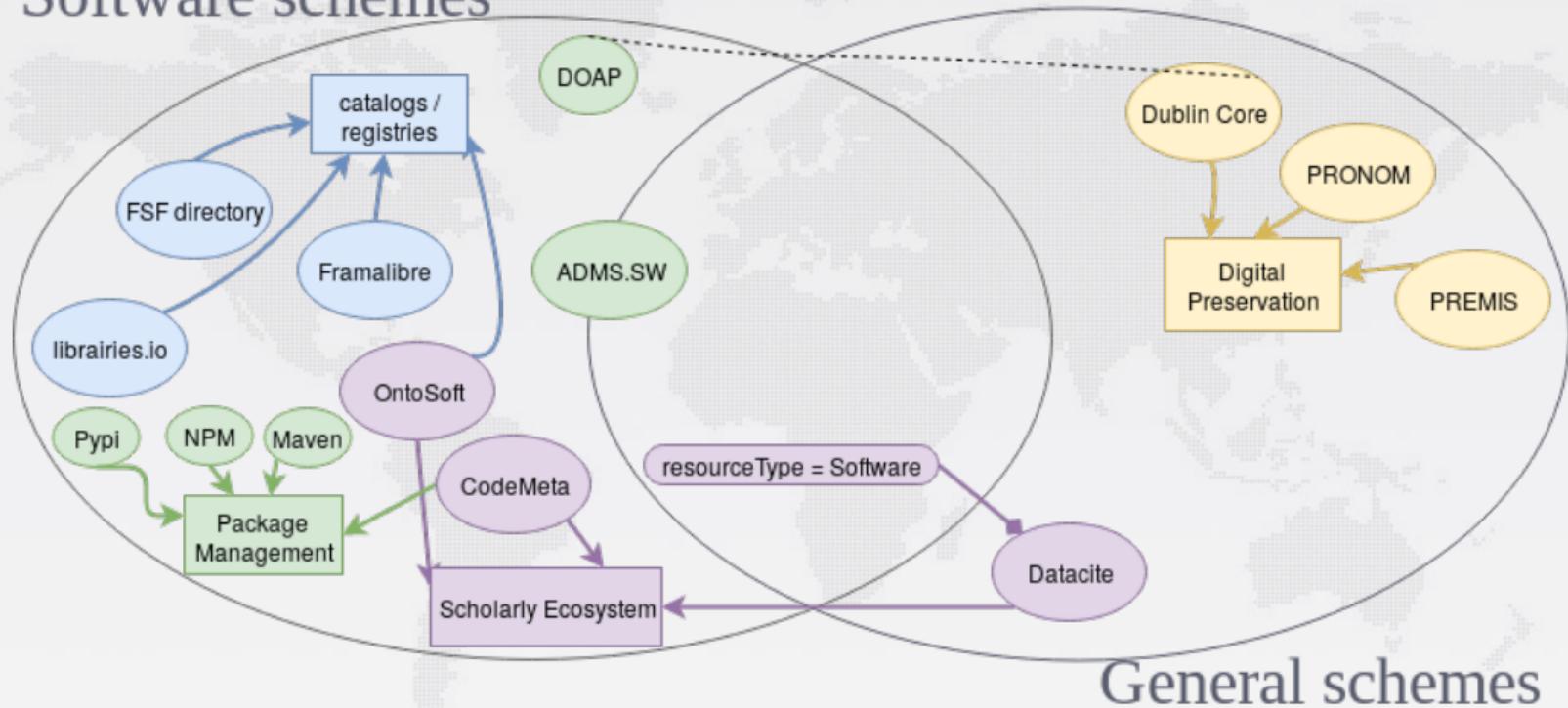
Software schemes



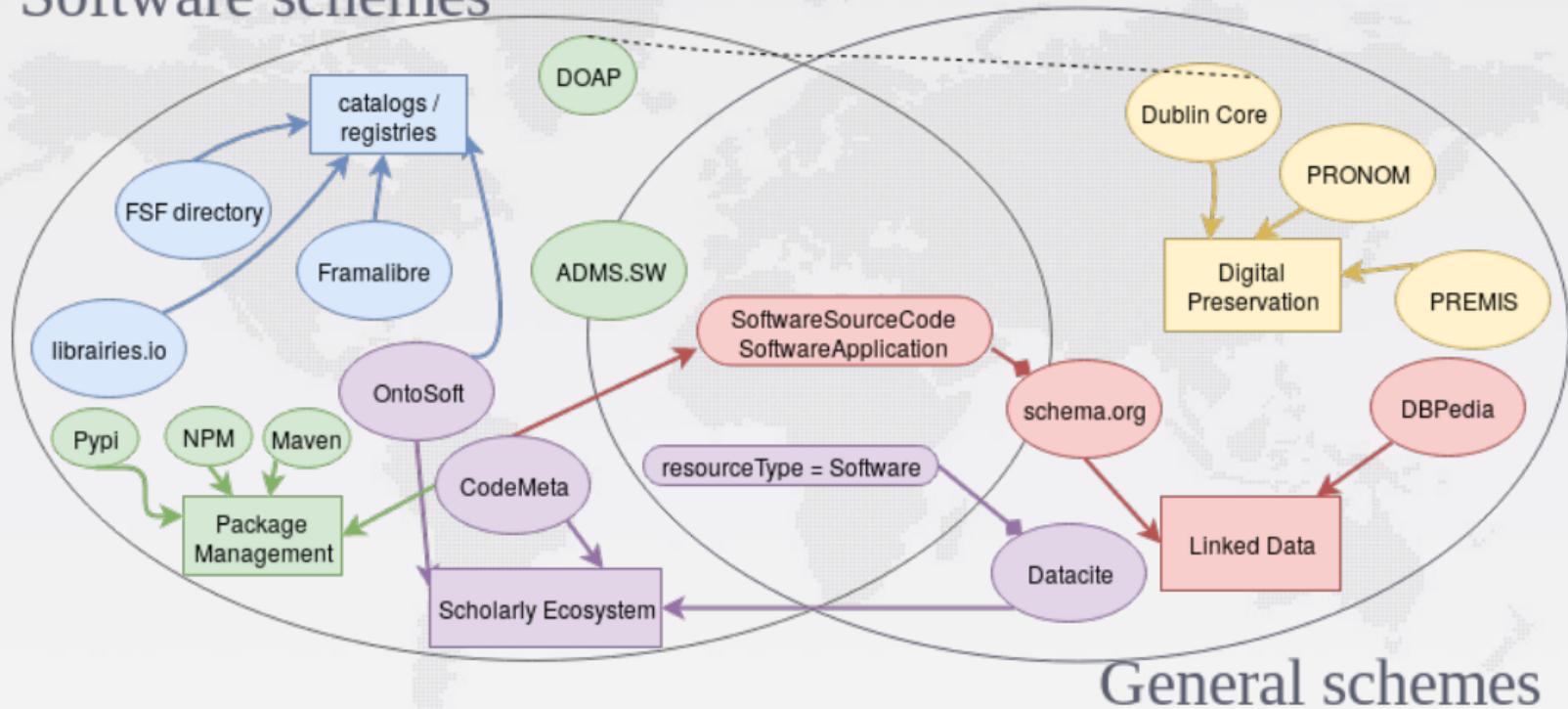
General schemes

The landscape of software ontologies

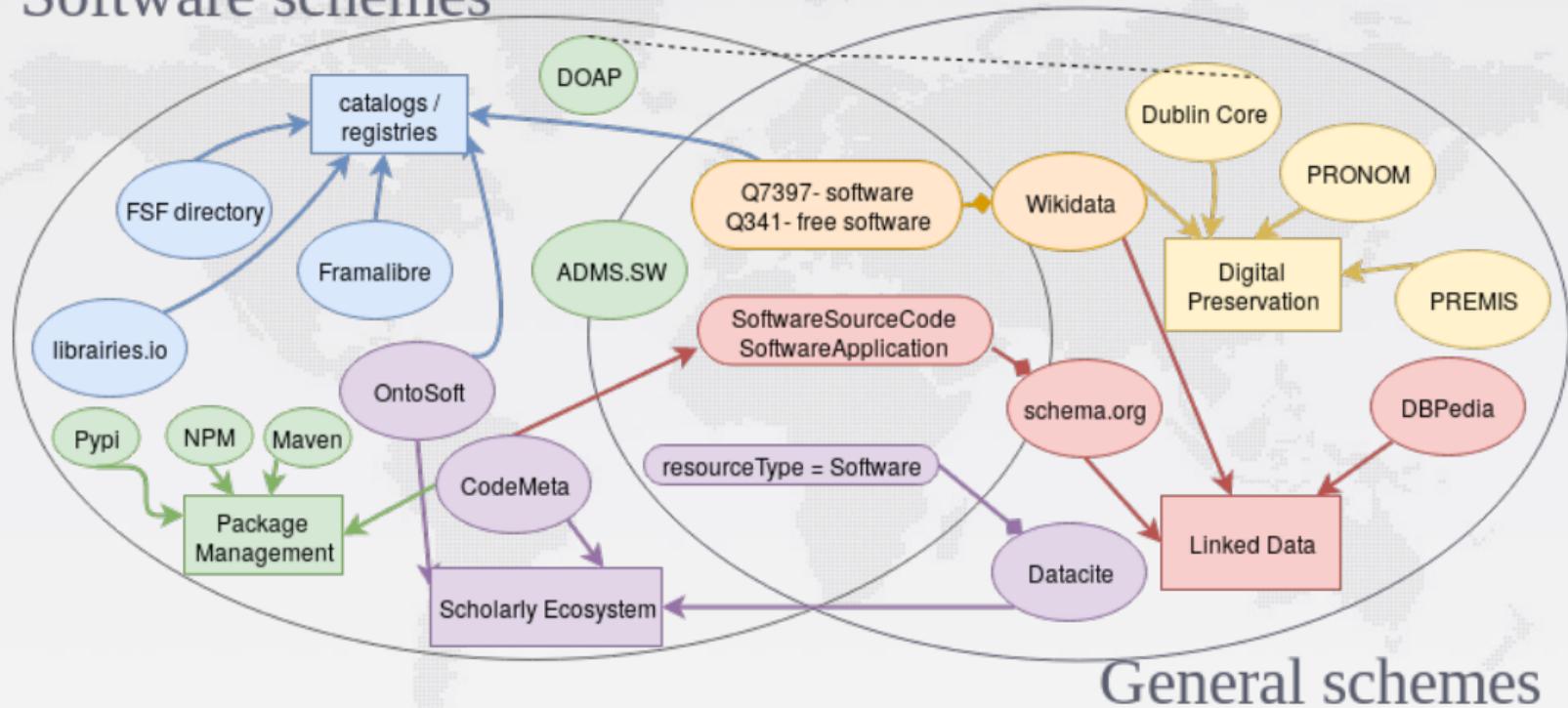
Software schemes

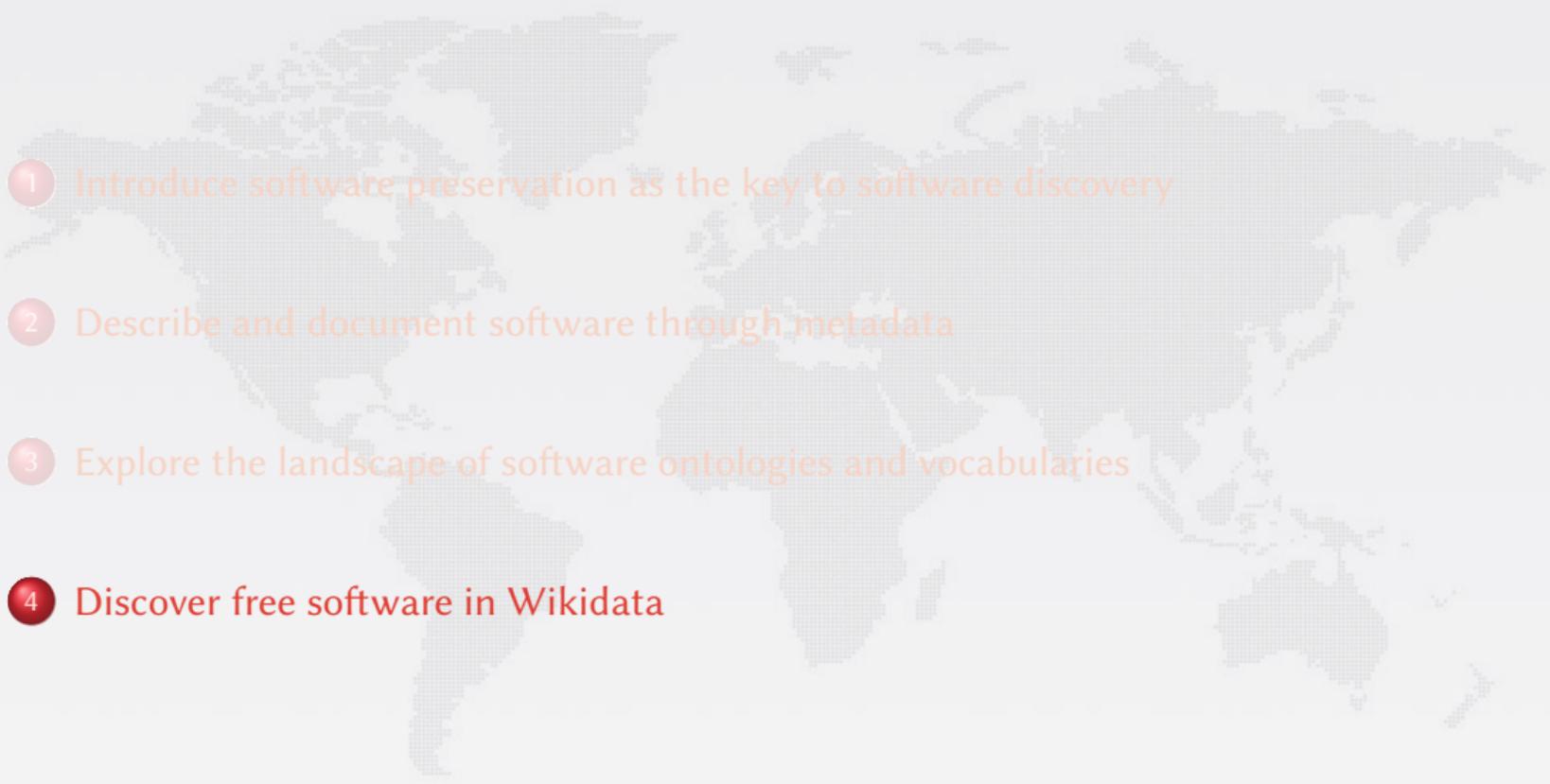


Software schemes



Software schemes



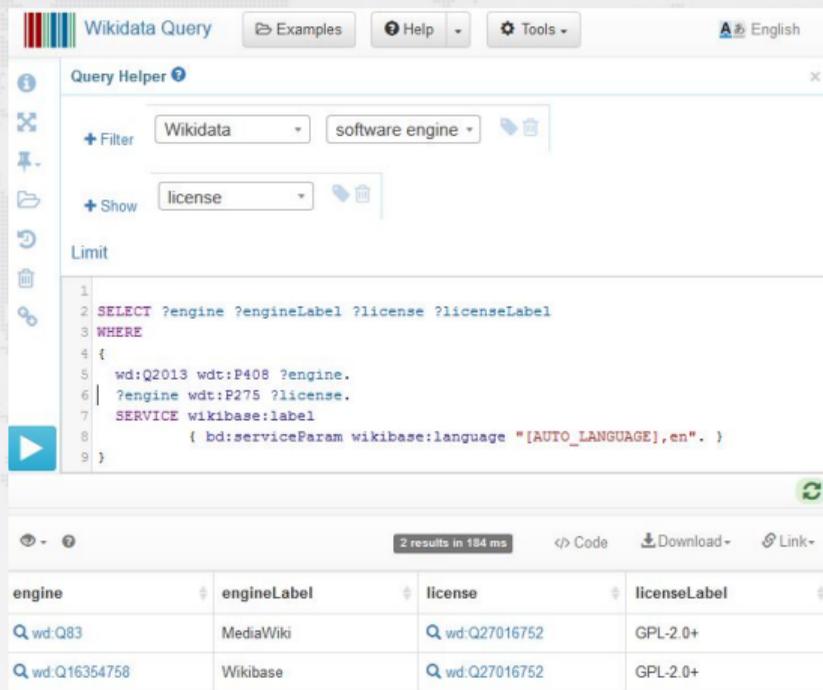
- 
- 1 Introduce software preservation as the key to software discovery
 - 2 Describe and document software through metadata
 - 3 Explore the landscape of software ontologies and vocabularies
 - 4 Discover free software in Wikidata

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



The screenshot shows the Wikidata Query interface. At the top, there are navigation buttons for 'Examples', 'Help', and 'Tools', along with a language selector set to 'English'. Below this is the 'Query Helper' section, which includes a filter set to 'Wikidata' and a software engine dropdown set to 'software engine'. The 'Show' section is set to 'license'. A 'Limit' section is also present. The main area displays a SPARQL query:

```
1
2 SELECT ?engine ?engineLabel ?license ?licenseLabel
3 WHERE
4 {
5   wd:Q2013 wdt:P408 ?engine.
6   ?engine wdt:P275 ?license.
7   SERVICE wikibase:label
8     { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
9 }
```

Below the query, the results are displayed in a table format. The table has four columns: 'engine', 'engineLabel', 'license', and 'licenseLabel'. There are two results shown:

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

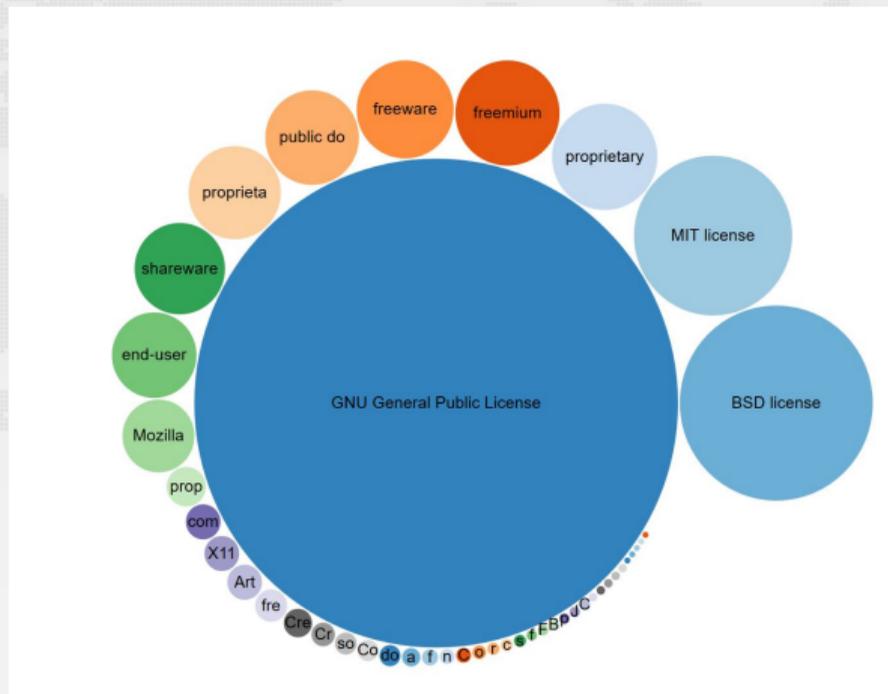


Figure: Try this query!

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

```
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 "en". }  
8 }
```

Figure: Try this query!

Create an image grid of Gnu/Linux distributions

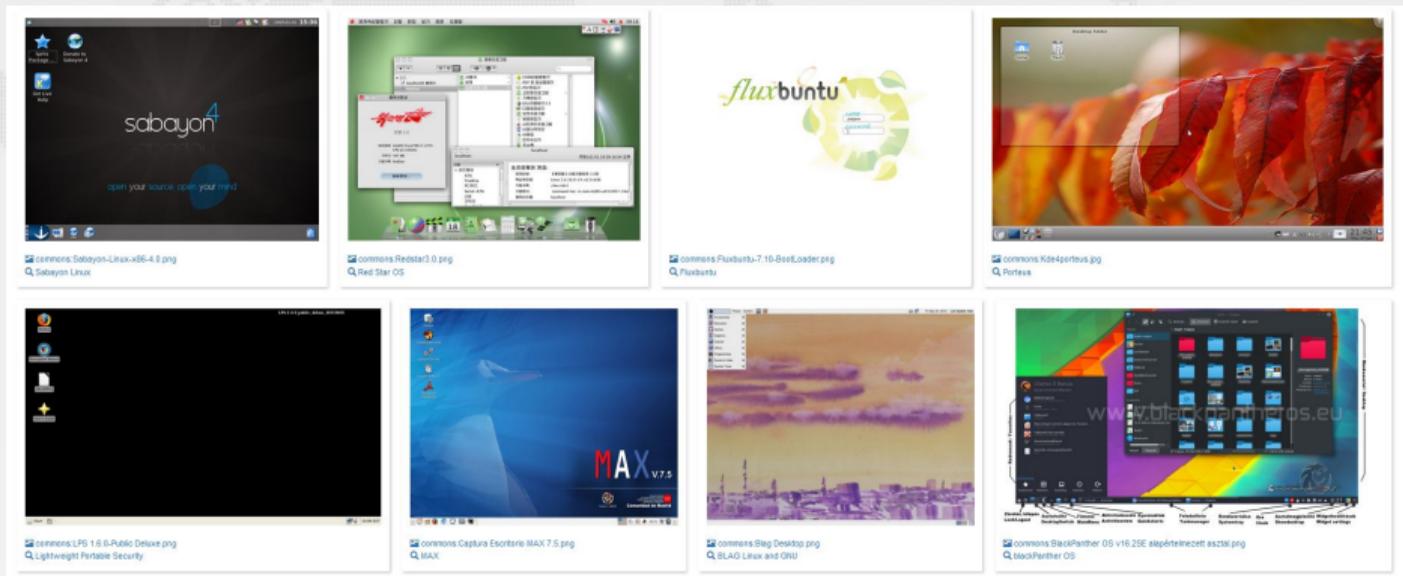


Figure: Try this query!

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

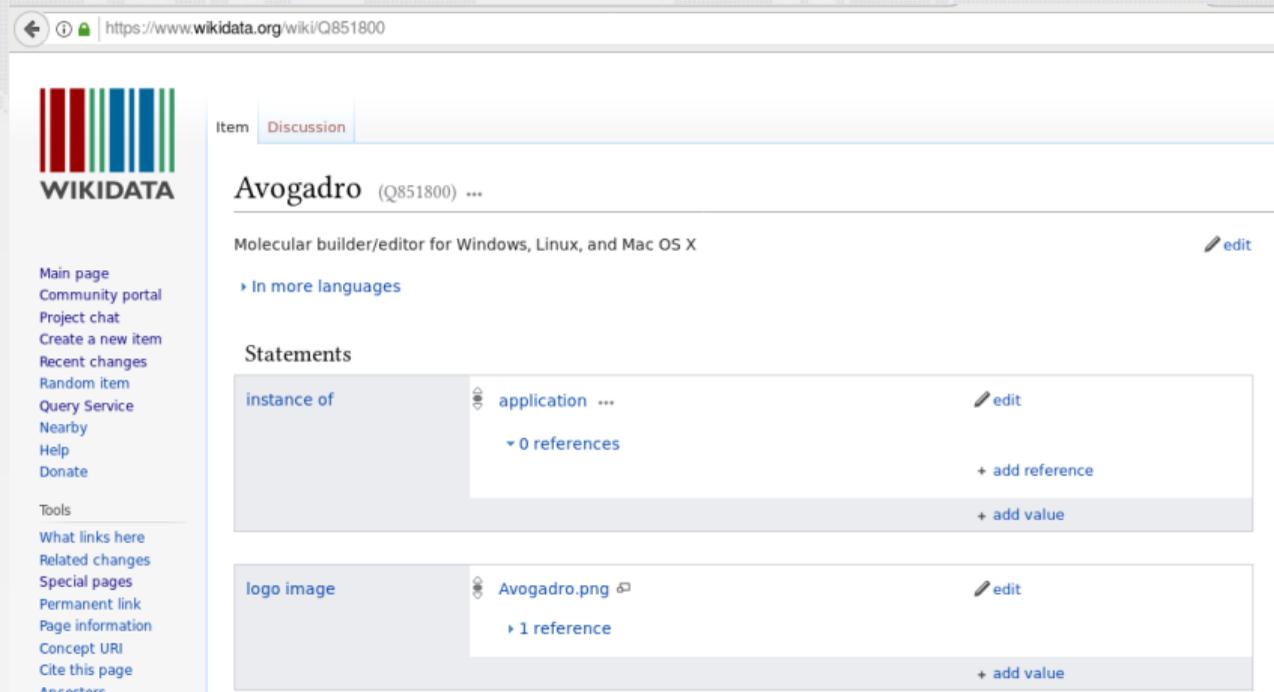
A screenshot of a Wikidata 'External sources' dropdown menu. The menu is titled 'External sources' and contains a list of external identifiers for the NumPy project. Each entry consists of a source name and a corresponding identifier value.

Source	Identifier
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.



The screenshot shows the Wikidata page for Avogadro (Q851800). The page is titled "Avogadro (Q851800) ..." and has a description: "Molecular builder/editor for Windows, Linux, and Mac OS X". There are two statements listed:

- instance of**: application ... (0 references, + add reference, + add value)
- logo image**: Avogadro.png (1 reference, + add value)

The left sidebar contains navigation links such as "Main page", "Community portal", "Project chat", "Create a new item", "Recent changes", "Random item", "Query Service", "Nearby", "Help", "Donate", "Tools", "What links here", "Related changes", "Special pages", "Permanent link", "Page information", "Concept URI", "Cite this page", and "Ancestors".

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  edit
	▶ 1 reference
	+ add value
Free Software Directory entry	 Avogadro  edit
	use science ...
	▶ 1 reference
	+ add value

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



The screenshot shows the FSF Resource Directory website. At the top left is the FSF logo and the text "FREE SOFTWARE DIRECTORY". A dark red navigation bar contains the words "about" and "ca". Below this is a search box with a "Go" button and a "Search" button. The main content area has a "Page" dropdown set to "Discussion" and tabs for "Overview", "Details", and "About this entry". The "About this entry" tab is active, showing the title "Avogadro" and a link to the main page. The text describes Avogadro as an advanced molecule editor and visualizer. A bulleted list highlights its features: International (translations), Intuitive (easy for students and researchers), Fast (multi-threaded rendering), Extensible (plugin architecture), and Flexible (Open Babel import).

FSF FREE SOFTWARE
D I R E C T O R Y

about ca

Search

Go

Search

Main page
All 16070 packages
Free GNU/Linux distros
Donate to the FSF
FSF store

Interaction

Participation guide
About the FSD
Recent changes
Mailing list
Active users
Backlog

Page Discussion

Overview Details About this entry

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, Italian
- Intuitive: Built to work easily for students and advanced research
- Fast: Supports multi-threaded rendering and computation.
- Extensible: Plugin architecture for developers, including render
- Flexible: Features include Open Babel import of chemical files, crystallography, and biomolecules

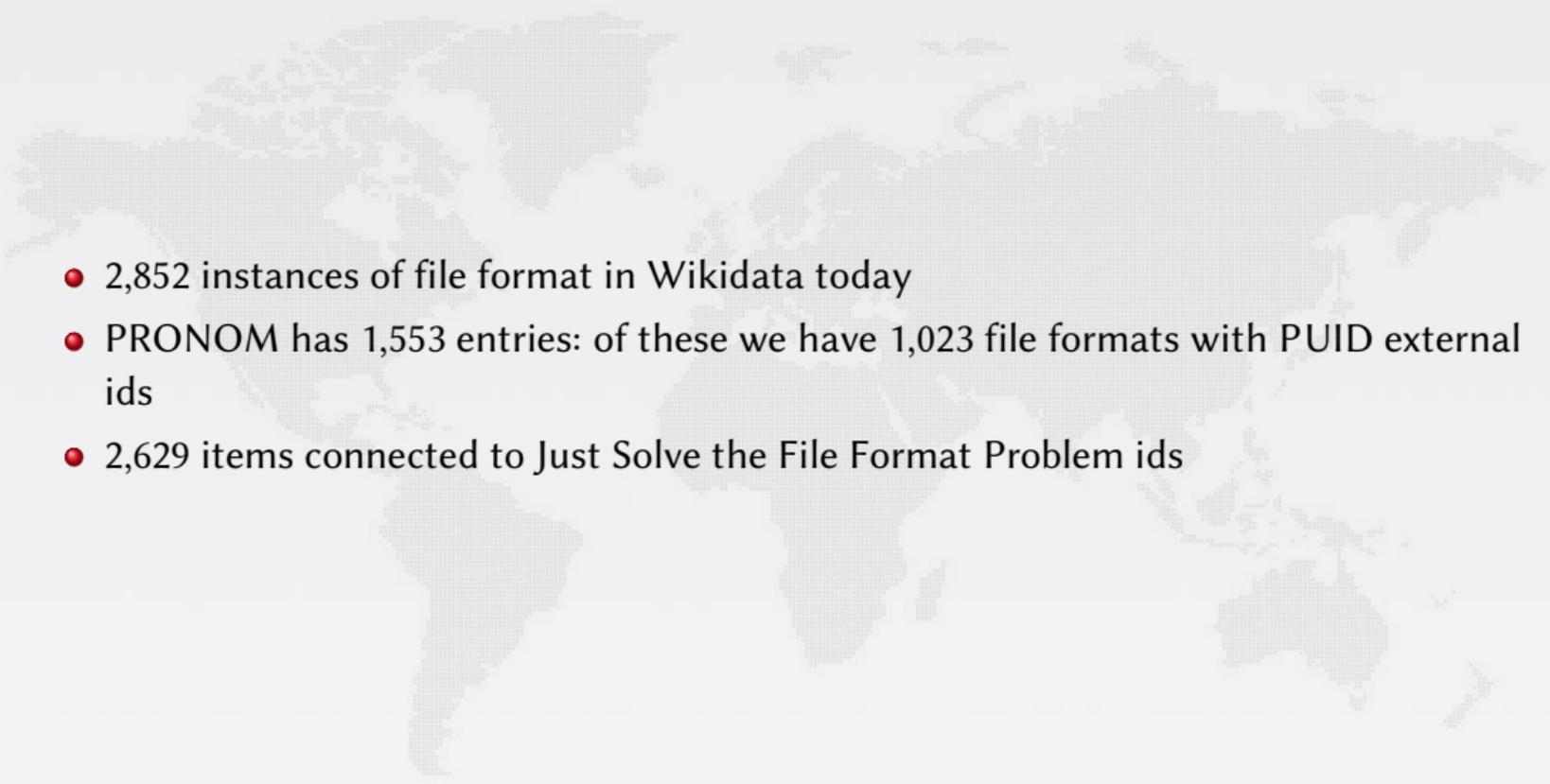
Wikidata crosswalks multiple external IDs

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

software	softwareLabel	deb	fed	gen	arch
Q8063151	ZNC			net-irc/znc	znc
Q203374	Zim	zim	Zim	x11-misc/zim	zim
Q189475	Zend Framework				
Q19612984	Zathura	zathura	zathura	app-text/zathura	zathura
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

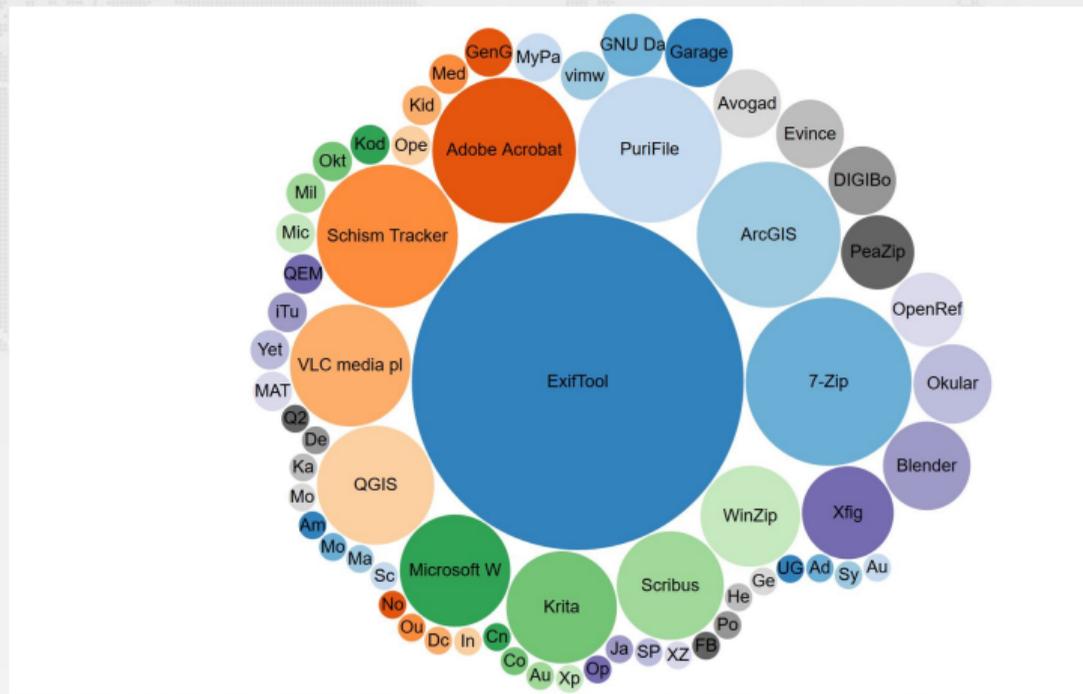
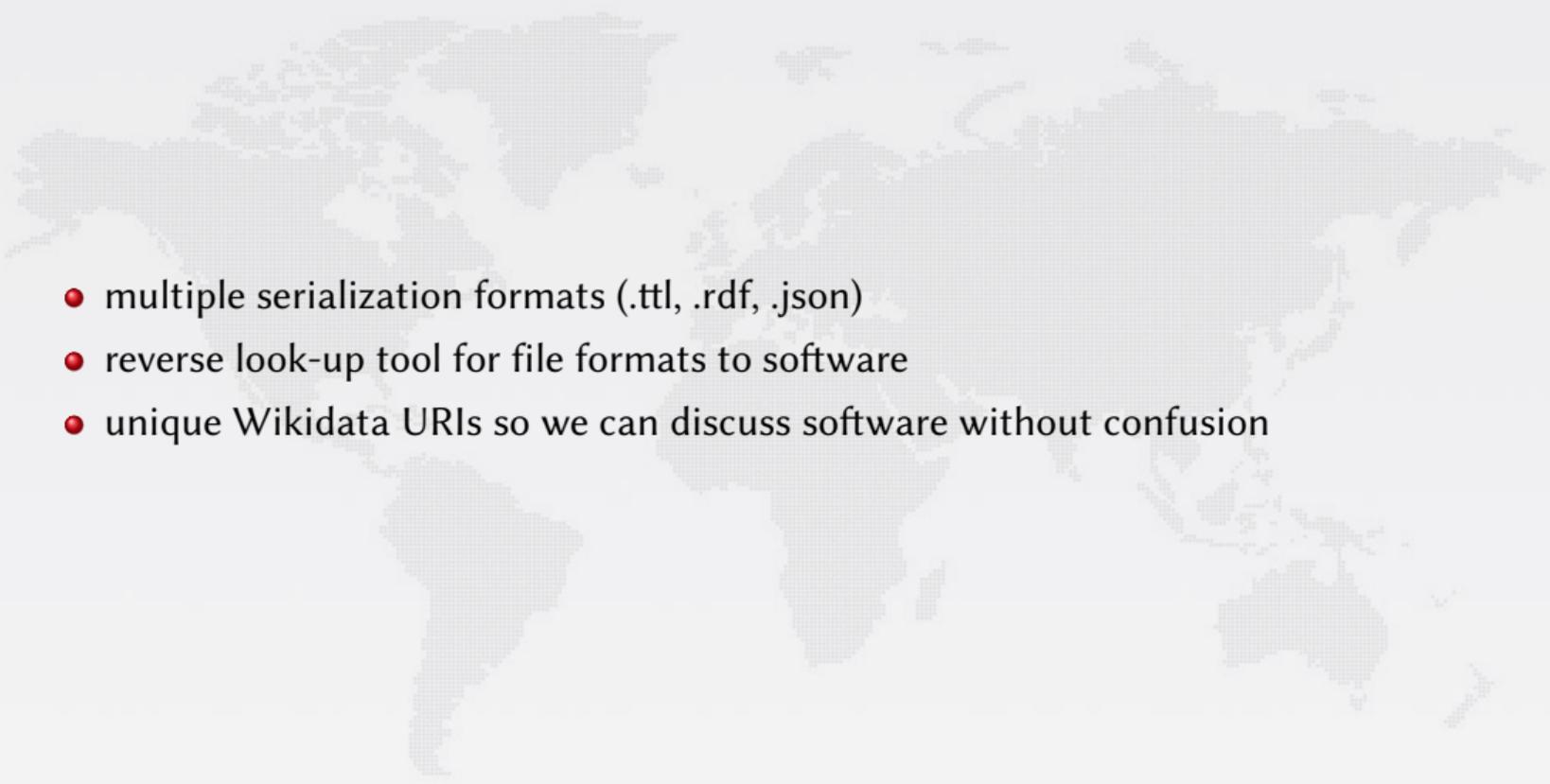
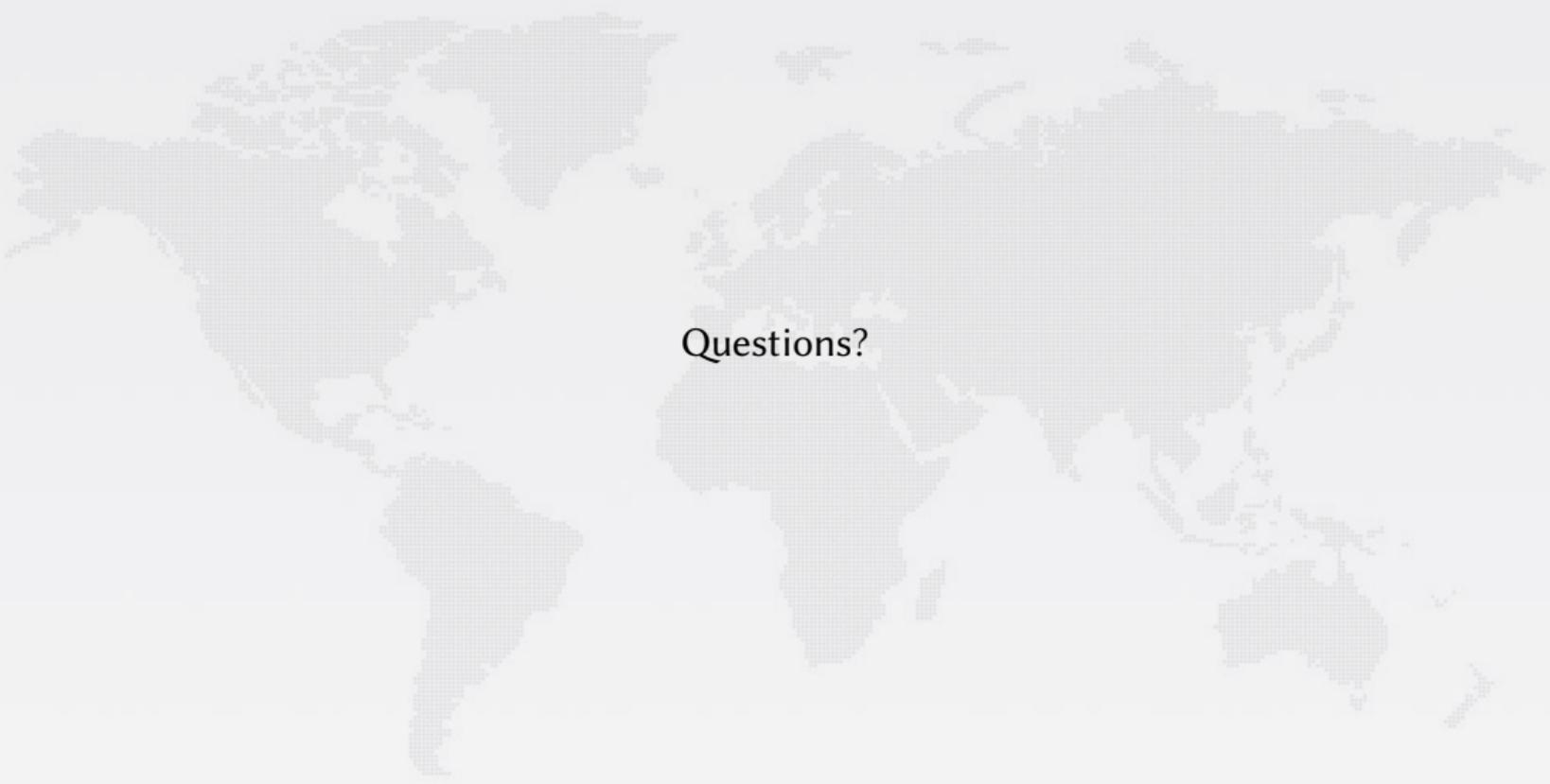


Figure: Try this query!

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

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

Communities

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