Software Ontologies and Metadata Schemes

Morane Gruenpeter

Inria, Software Heritage

September 19, 2017

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Preliminary questions

“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

Preliminary questions

"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 Ontology

- What is software ?
- With what terms should we describe a *software artifact*?
- What about *software source code* ?

Metadata about Software Source Code

Software metadata objectives

manage, share, discover, archive *software source code*

Use cases

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

LOV- Linked open vocabularies

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

Where is the metadata available ?

catalogs and registries

- libraries.io
- OpenHub
- OntoSoft

Publisher's repositories

- GitHub
- Bitbucket
- SourceForge

Where is the metadata available ?

catalogs and registries

- libraries.io
- OpenHub
- OntoSoft

Publisher's repositories

- GitHub
- Bitbucket
- SourceForge

advantages and drawbacks

	registries	repositories
accuracy	- not created by author	+ added by authors/maintainers
completeness	+ very detailed	- not a priority
longevity	- depends on registry	- depends on publisher

Where is the metadata available ?

in the *software source code* itself

- package management file
- CITATION file
- .About file
- codemeta.json file

Where is the metadata available ?

in the *software source code* itself

- package management file
- CITATION file
- .About file
- codemeta.json file

advantages and drawbacks

	metadata file
accuracy	+ created by author and evolves with code
completeness	+ freedom of vocabulary and terms used
longevity	+ not dependent on platform (repository or registry)

Where is the metadata available ?

in the *software source code* itself

- package management file
- CITATION file
- .About file
- codemeta.json file

advantages and drawbacks

	metadata file
accuracy	+ created by author and evolves with code
completeness	+ freedom of vocabulary and terms used
longevity	+ not dependent on platform (repository or registry)

Bottomline: to insure the archival of metadata, keep it in the data

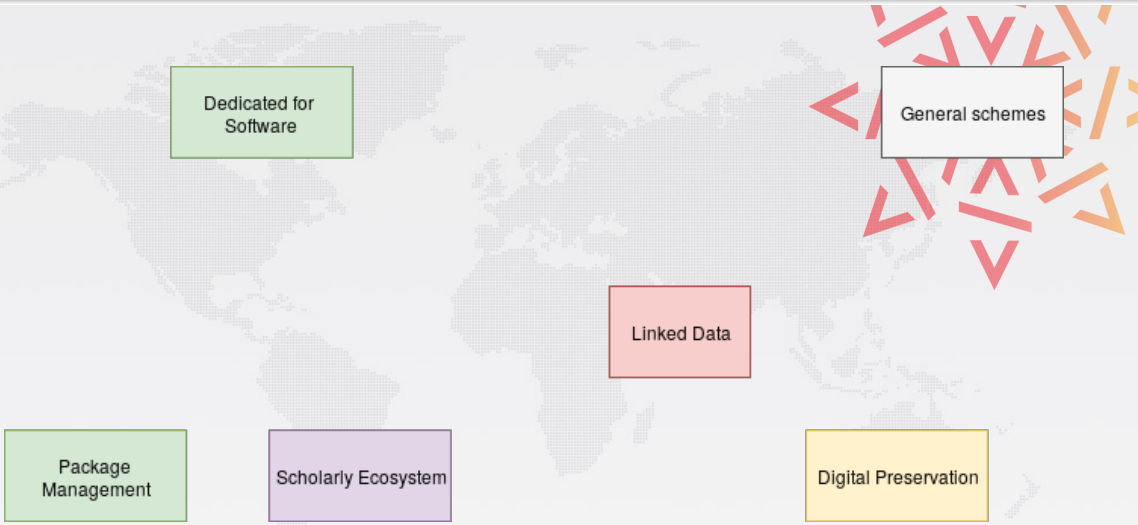
This is not *software source code*

Ceci n'est pas une pipe.

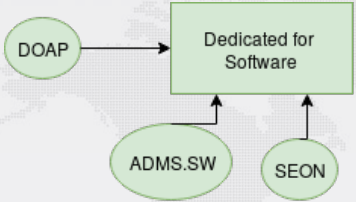
Software Citation Principles (FORCE11's 2015 conference and WG)

- Importance : first class citizen in the scholarly ecosystem
- Credit and attribution : authors, maintainer
- Unique identification: points to a unique, specific software version (DOI, Git SHA1 hash, etc..)
- Persistence : identification beyond the lifespan of the software (swh-id)
- Accessibility: url, publisher
- Specificity : version, environment

Landscape of Software Ontologies



Landscape of Software Ontologies



General schemes

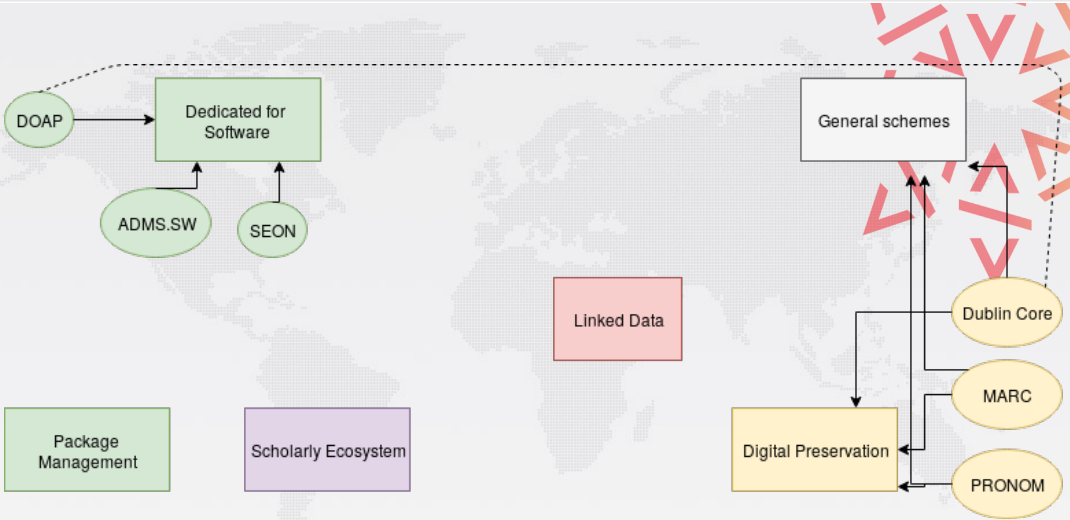
Linked Data

Package Management

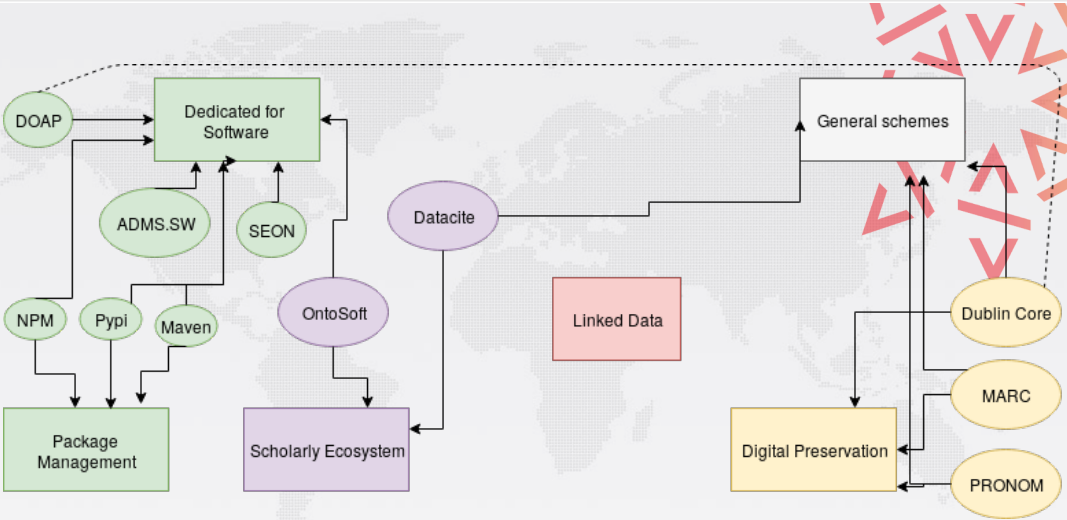
Scholarly Ecosystem

Digital Preservation

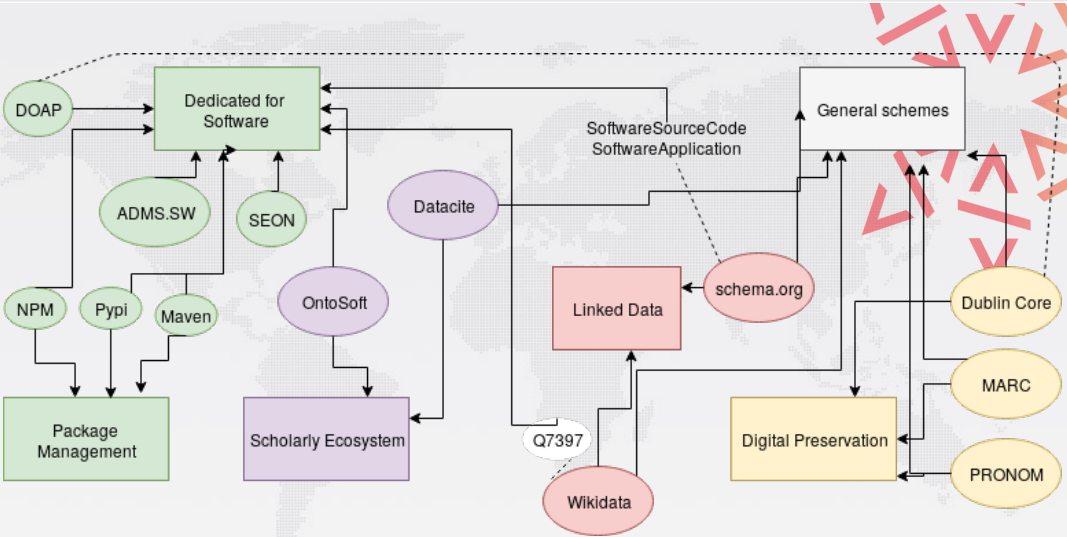
Landscape of Software Ontologies



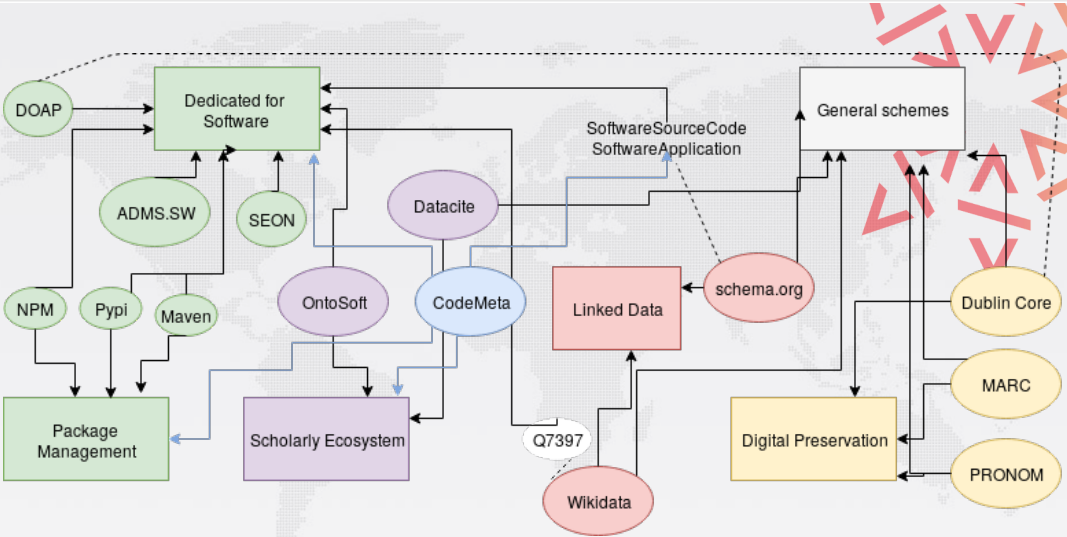
Landscape of Software Ontologies



Landscape of Software Ontologies



Landscape of Software Ontologies



The CodeMeta Initiative

A Rosetta Stone for Metadata in Scientific Software

CodeMeta aims to create a framework {schema, crosswalk, guidlines}
that can be used to standarize the exchange of software metadata

Advantages

- the crosswalk table
- built on schema.org *SoftwareSourceCode*
- an active community

Discussion

CodeMeta -where are the gaps ?

- missing properties
- missing ontologies
- semantic misconceptions

Software Source Code metadata recommendations

- use cases
- best practices / guidlines

Reminder

RDA page

https://www.rd-alliance.org/ig-software-source-code-rda-10th-plenary-meeting

Working document used during the session

http://bit.ly/2wggInQ