Software Ontologies and Metadata Schemes

Morane Gruenpeter

Inria, Software Heritage

September 19, 2017
“Ontologies are agreements, made in a social context, to accomplish some objectives. It’s important to understand those objectives, and be guided by them.”

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


**Software Ontology**

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

“How 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**

<table>
<thead>
<tr>
<th></th>
<th>registries</th>
<th>repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>- not created by author</td>
<td>+ added by authors/maintainers</td>
</tr>
<tr>
<td>completeness</td>
<td>+ very detailed</td>
<td>- not a priority</td>
</tr>
<tr>
<td>longevity</td>
<td>- depends on registry</td>
<td>- depends on publisher</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th>Metadata file</th>
<th>Accuracy</th>
<th>+ created by author and evolves with code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Completeness</td>
<td>+ freedom of vocabulary and terms used</td>
</tr>
<tr>
<td></td>
<td>Longevity</td>
<td>+ not dependent on platform (repository or registry)</td>
</tr>
</tbody>
</table>
Where is the metadata available?

In the *software source code* itself:
- package management file
- CITATION file
- .About file
- codemeta.json file

Advantages and drawbacks:

<table>
<thead>
<tr>
<th>metadata file</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
</tr>
<tr>
<td>+ created by author and evolves with code</td>
</tr>
<tr>
<td>completeness</td>
</tr>
<tr>
<td>+ freedom of vocabulary and terms used</td>
</tr>
<tr>
<td>longevity</td>
</tr>
<tr>
<td>+ not dependent on platform (repository or registry)</td>
</tr>
</tbody>
</table>

**Bottomline:** To insure the archival of metadata, keep it in the data.
This is not *software source code*

*Ceci n'est pas une pipe.*
The Software Ontology *Touchstone*

### 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

DOAP → Dedicated for Software
ADMS.SW → SEON
NPM → Pypi → Maven

Package Management

OntoSoft → Datacite

Scholarly Ecosystem

Linked Data

General schemes

Digital Preservation

Dublin Core
MARC
PRONOM
A Rosetta Stone for Metadata in Scientific Software

CodeMeta aims to create a framework \{schema, crosswalk, guidelines\} 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 / guidelines
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