HELMHOLTZ PROJECT HERMES

HELMHOLTZ RICH METADATA SOFTWARE PUBLICATIONS

HERMES Kickoff Workshop 2021-11-12

SOMEF:
A Metadata Extraction Framework from Software Documentation

Daniel Garijo
@dgarijov

Ontology Engineering Group,
Universidad Politécnica de Madrid
HERMES kick off workshop, Nov 12, 2021

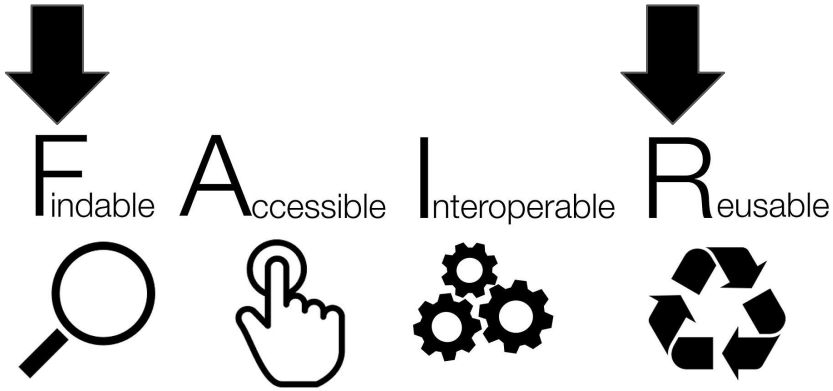
The Importance of Software Metadata

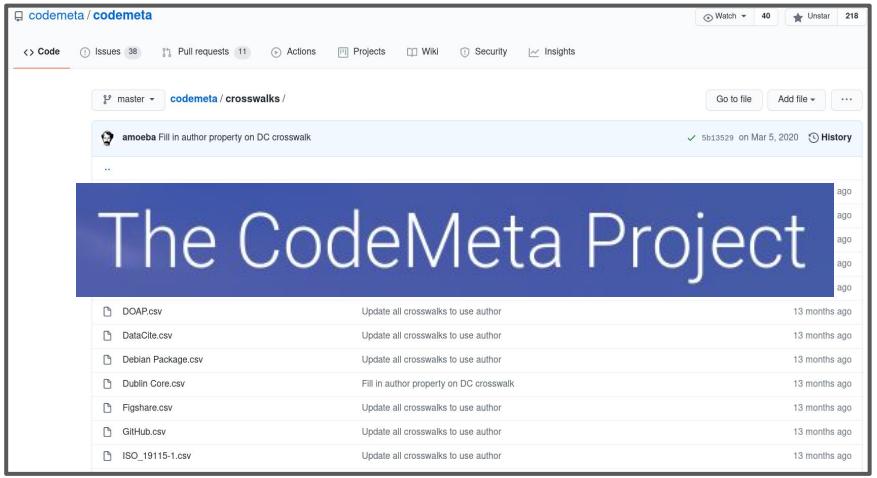
- An increasing amount of software is available online
- Metadata is critical for:
  - **Findability**
  - **Reusability**
  - **Understanding**

zenodo

> 100 M repositories

> 28 M repositories





Current Pipeline for Software Metadata Curation

Develop software & documentation

Create entry in registry

Curation and Validation

Icons from https://icons8.com/

The problems of metadata collection...

Can you please describe your software component with metadata?

I already did! Did you read the project readme?

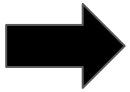
Did you see the online documentation?

Perhaps the you saw the paper?

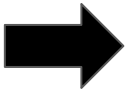
Many domain-specific registries are curated by hand by experts

SOMEF: Automating Software Metadata Extraction

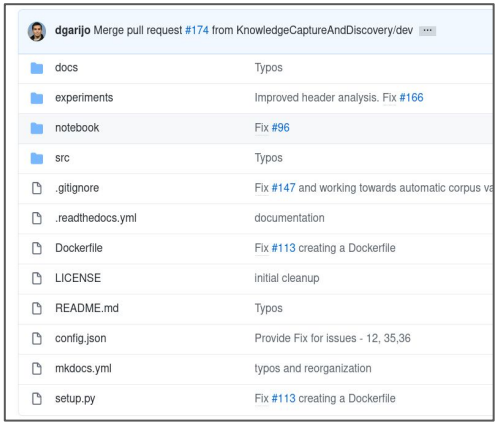
Repository



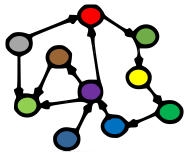
Extraction



Results (Metadata)



- **Readme Analysis**
  - Supervised classification
  - Regular expressions
  - Header analysis
- **File exploration**
  - Notebooks
  - Dockerfiles
  - Documentation
- **GitHub API**



https://github.com/KnowledgeCaptureAndDiscovery/somef/

SOMEF: Supervised classification

- Paragraph-based text classification
- Four main categories:
  - Installation, citation, description, invocation.
- Binary classification problem

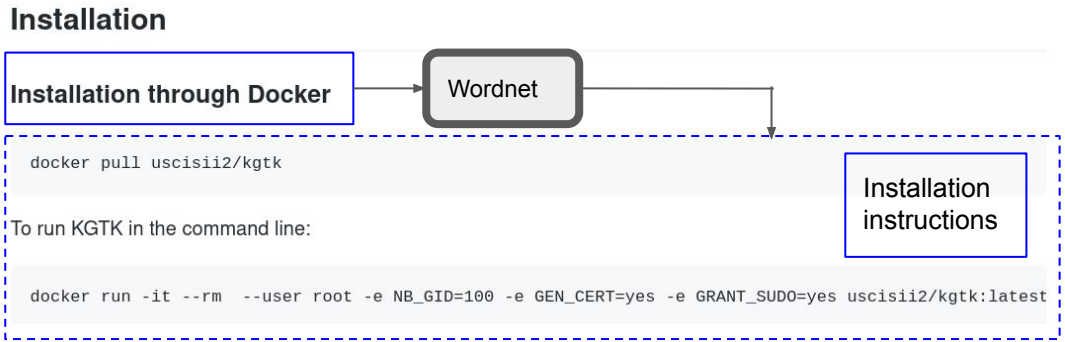
Truth Value	Category	Apprx. Ratio	Count
True	Description	0.5	275
False	Installation	0.125	68
	Invocation	0.125	68
	Citation	0.125	68
	Treebank	0.125	68
	Total	1.0	547

Classifier	Best pipeline	Precision	Recall	F-Measure
Description	CountVectorizer + LogisticRegression	0.85	0.79	0.82
Installation	TFIDFVectorizer + StochasticGradientDescent	0.92	0.9	0.91
Invocation	CountVectorizer + NaiveBayes	0.88	0.9	0.89
Citation	CountVectorizer + NaiveBayes	0.89	0.98	0.93

Table 2. Best classification results for each metadata category

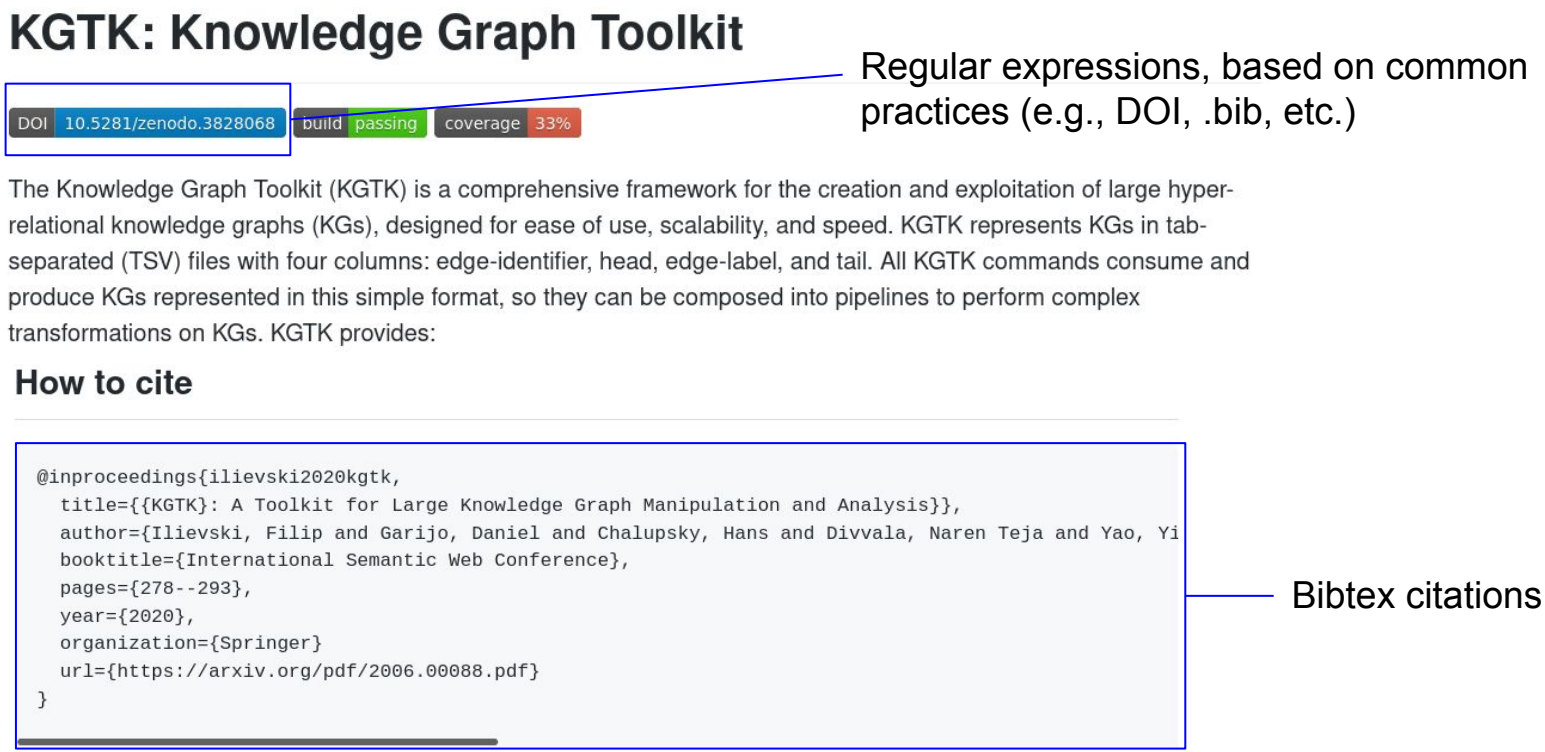
SOMEF: Header Analysis

- Extraction based on frequent header analysis
  - Fuzzy matching based on synsets



Category	Header analysis (F-Measure)
Description	0.68
Installation	0.85
Invocation	0.91
Citation	0.87
Usage	0.68
Documentation	0.95
Requirements	0.93
Support	0.52
License	1

SOMEF: File Exploration and Regular Expressions



Recognizing Metadata Categories

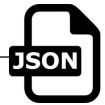
- **Name** (GA)
- **Full title** (RE)
- **Description** (SC, HA)
- **Citation** (SC, RE, HA)
- **Installation instructions** (SC, HA)
- **Invocation** (SC)
- **Usage examples** (HA)
- **Documentation** (HA, FE)
- **Requirements** (HA)
- **Contributors** (HA)
- **FAQ** (HA)
- **Support** (HA, RE)
- **License** (GA, HA)
- **Stars** (GA)

Method used:

- Supervised Classification (SC)
- Header Analysis and Synset comparison (HA)
- File Exploration (FE)
- Regular Expressions (RE)
- GitHub API (GA)

- **Contact** (HA)
- **Download URL** (HA, GA)
- **DOI** (RE)
- **DockerFile** (FE)
- **Notebooks** (FE)
- **Executable notebooks (Binder)** (RE)
- **Owner**: (GA)
- **Keywords** (GA)
- **Source code** (GA)
- **Releases** (GA)
- **Changelog** (GA)
- **Issue tracker** (GA)
- **Programming languages** (GA)
- **Acknowledgements** (HA)
- **Repository status** (RE)
- **Arxiv links** (RE)
- **Scripts** (FE)
- **Contributors** (FE)
- **… (currently expanding more)**

Example Output



CodeMeta


```
description:
  0:
    excerpt:        "WIDOCO helps you to publish and create an enriched and customized documentation of you
                    classes, properties and data properties of the ontology, the OOPS! webservice by María
                    being used. In addition, we use WebVowl to visualize the ontology and have extended Bub
                    documentation of the terms in your ontology (based on [LODE](http://www.essepuntato.it/
                    annotation in JSON-LD snippets of the html produced.\n* Association of a provenance pag
                    means to complete it on the fly when generating your ontology. Check the [best practice
                    WIDOCO.\n* Guidelines on the main sections that your document should have and how to co
                    changelog of differences between the actual and the previous version of the ontology (b
                    them independently and replace only those needed.\n* Content negotiation and serializat
    confidence:
      0:            1
    technique:      "wordnet"
  1:
    excerpt:        "For a complete list of the current improvements and next features, check the project o
    confidence:
      0:            0.8231493588525339
    technique:      "classifier"
  2:
    excerpt:        "Wizard for documenting ontologies. WIDOCO is a step by step generator of HTML template
    confidence:
      0:            1
    technique:      "metadata"
citation:
  0:
    excerpt:        "@inproceedings{garijo2017widoco,\n  title={WIDOCO: a wizard for documenting ontologies
                    organization={Springer, Cham},\n  doi = {10.1007/978-3-319-68204-4_9},\n  funding = {US
    confidence:
      0:            1
    technique:      "classifier"
```

```
@context:           "https://doi.org/10.5063/schema/codemeta-2.0"
@type:              "SoftwareSourceCode"
license:            "https://raw.githubusercontent.com/dgarijo/Widoco/master/LICENSE"
codeRepository:     "git+https://github.com/dgarijo/Widoco.git"
dateCreated:        "2013-07-15"
datePublished:      "2020-12-14"
dateModified:       "2021-03-16"
downloadUrl:        "https://github.com/dgarijo/Widoco/releases"
issueTracker:       "https://github.com/dgarijo/Widoco/issues"
name:               "Widoco"
version:            "v1.4.15_1"
description:
  0:                "Wizard for documenting ontologies. WIDOCO is a step by step generat
  1:                "WIDOCO helps you to publish and create an enriched and customized d
                    the classes, properties and data properties of the ontology, the OOP
                    URI and title being used. In addition, we use WebVowl to visualize t
                    WIDOCO:\n* Automatic documentation of the terms in your ontology (ba
                    /index.html))\n* Automatic annotation in JSON-LD snippets of the htm
                    extraction from the ontology plus the means to complete it on the fl
                    to know more about the terms recognized by WIDOCO.\n* Guidelines on
                    (http://vowl.visualdataweb.org/webvowl/)).\n* Automatic changelog of
                    /)).\n* Separation of the sections of your html page so you can writ
                    practices\n\n"
  2:                "For a complete list of the current improvements and next features,
releaseNotes:       "This pre-release fixes issues regarding namespace prefixes (now the
                    settings in your visualization)\r\n\r\nMore information on the addre
keywords:
  0:                "ontology"
  1:                "wizard"
  2:                "metadata"
  3:                "documentation"
  4:                "ontology-diagram"
  5:                "ontology-evaluation"
```

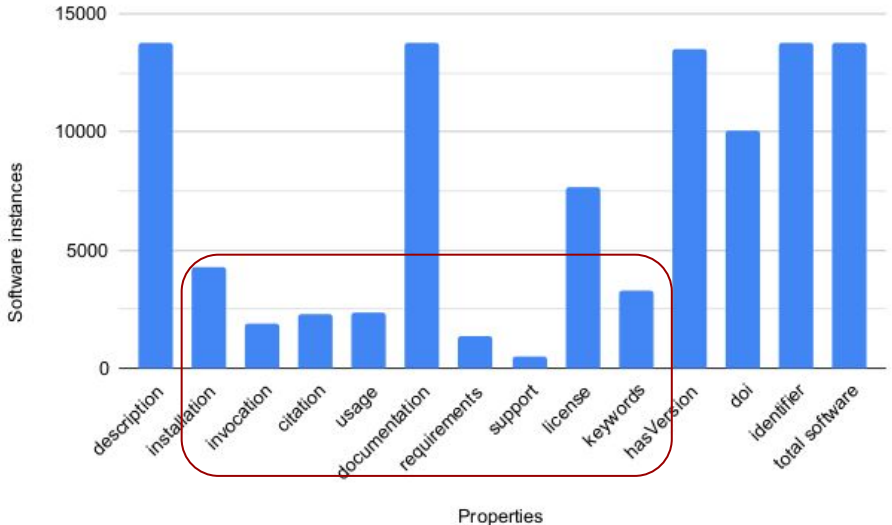
What are we using SOMEF for?

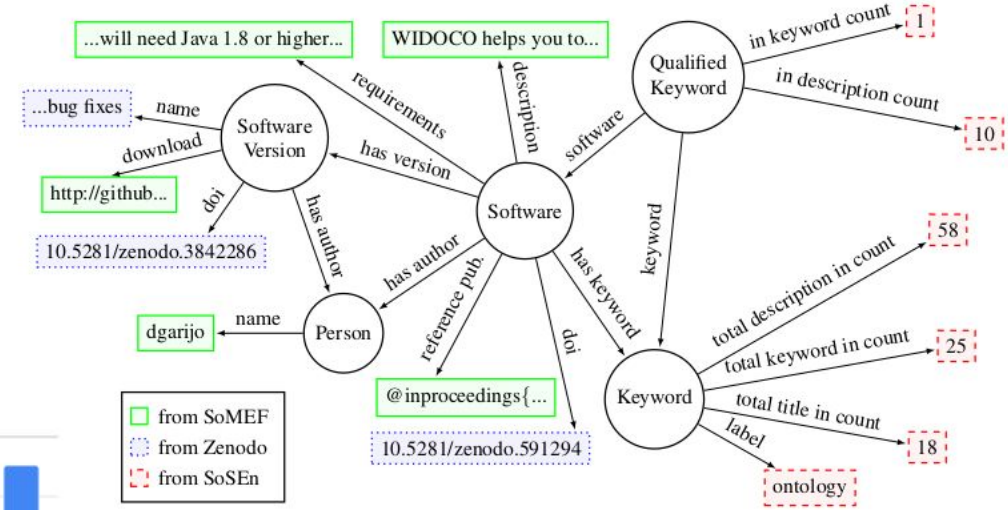
Building Knowledge Graphs of Software Metadata

> 13k software entries from Zenodo

Assessing the adoption of metadata

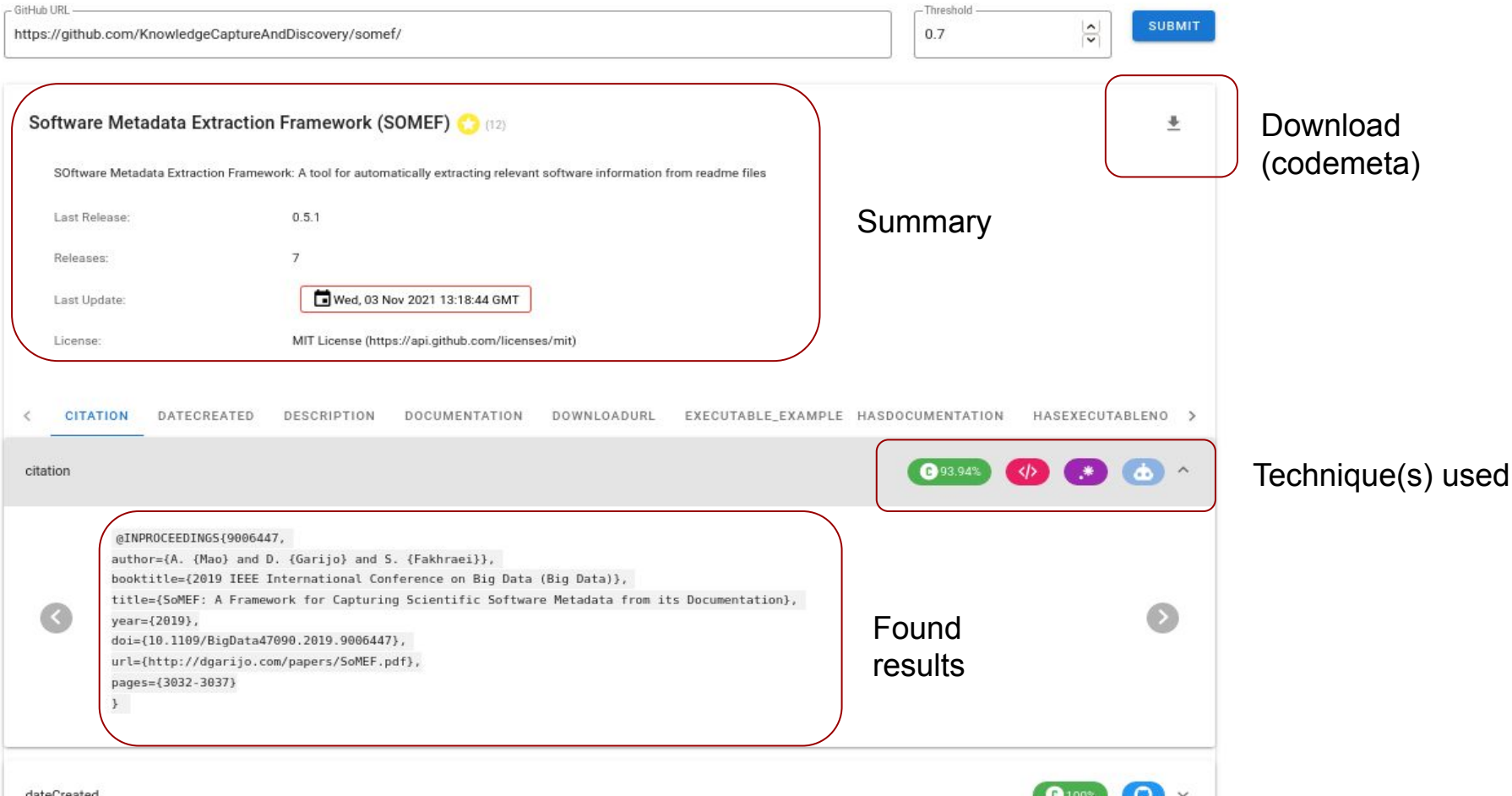
Enabling metadata-based comparison





https://github.com/KnowledgeCaptureAndDiscovery/sosen

Creating applications to spread best practices adoption



Download (codemeta)

Summary

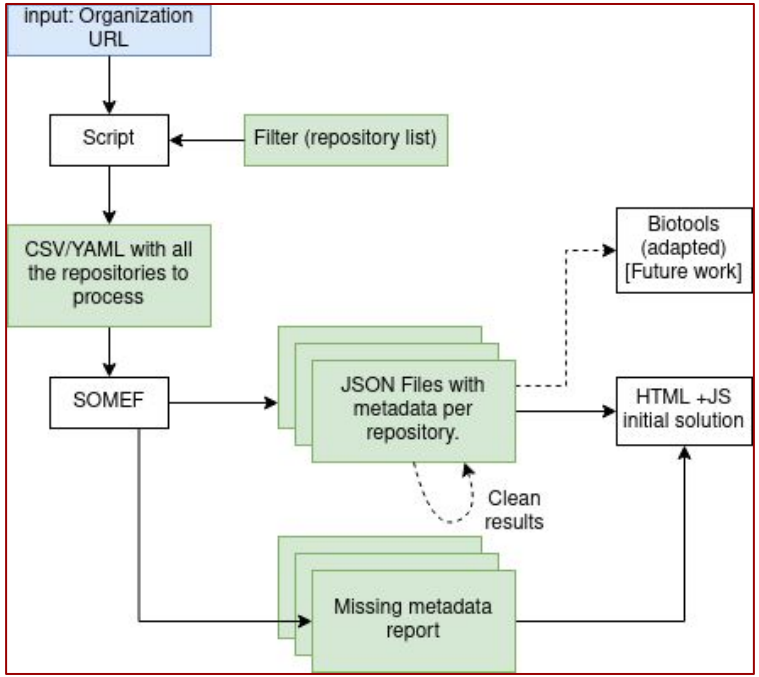
Technique(s) used

Found results

(Work in progress)

- GitHub actions to generate reports of missing metadata

- Semi-automated generation of software catalogs



Beginning to capture software in context

SOMEF helps automatically extract software metadata. Benefits:

- Let authors know what metadata is missing
- Help making software findable
- Capture context of software (setup, configuration, examples)

But there is much more to explore:

- Software classification
- Named Entity Recognition
- Comparison
- Links to publications

Problems? Suggestions? New features?
https://github.com/KnowledgeCaptureAndDiscovery/somef/issues

Thank you!

Where to learn more about project HERMES?


Stephan Druskat, DLR, PI, *@stdruskat*


Oliver Bertuch, FZJ, PI, *@poi_ki_lo_therm*


Guido Juckeland, HZDR, PI, *@GuidoJuckeland*


Oliver Knodel, HZDR, *@olikno1*


Tobias Schlauch, DLR, *@TobiasSchlauch*

- Find us on Twitter
- Write an email to team@software-metadata.pub
- Go to software-metadata.pub