A Large-scale Dataset of (Open Source) License Text Variants

MSR 2022

Stefano Zacchiroli
Télécom Paris, Polytechnic Institute of Paris
stefano.zacchiroli@telecom-paris.fr
upsilon.cc/zack | @zacchiro | mastodon.xyz/@zacchiro

Motivations

- Free/**Open Source** Software (FOSS) is everywhere in IT products
- Many different **FOSS licenses** exist, including their **variants**, e.g.:
  - license exceptions
  - additional clauses
  - ad-hoc modifications
  - …

  and have been studied in ESE/MSR research.

How do we study the *full corpus* of FOSS license variants?

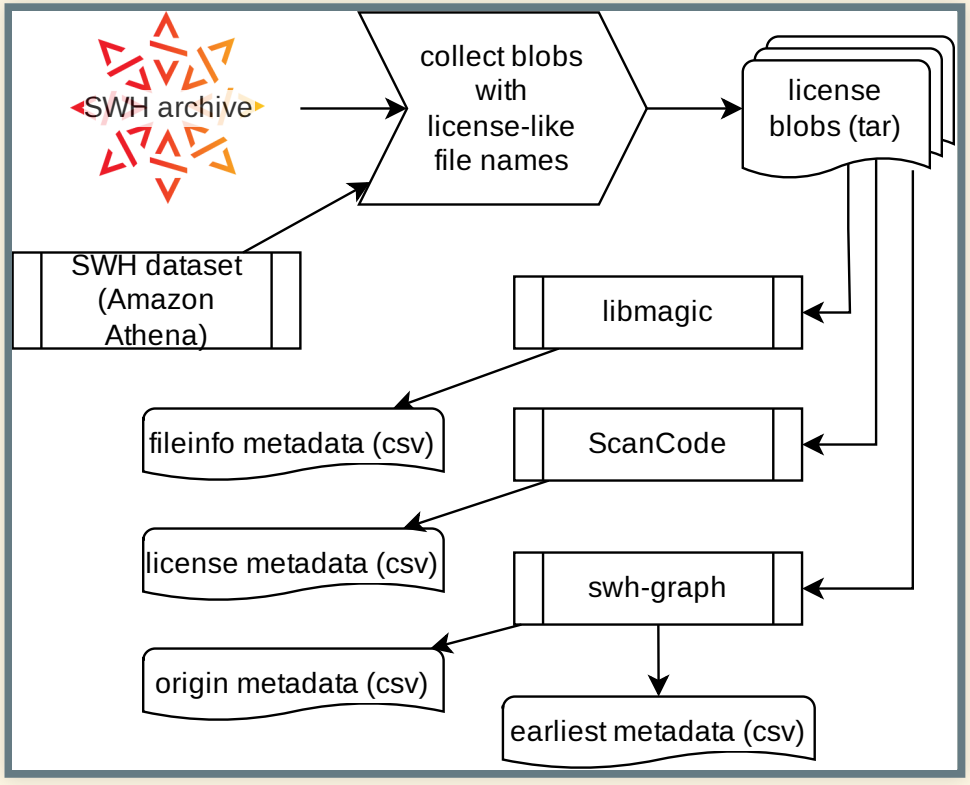
Contribution

We introduce a large-scale dataset of **6.5 million unique license files**, collected from more than **150 million public development projects**.

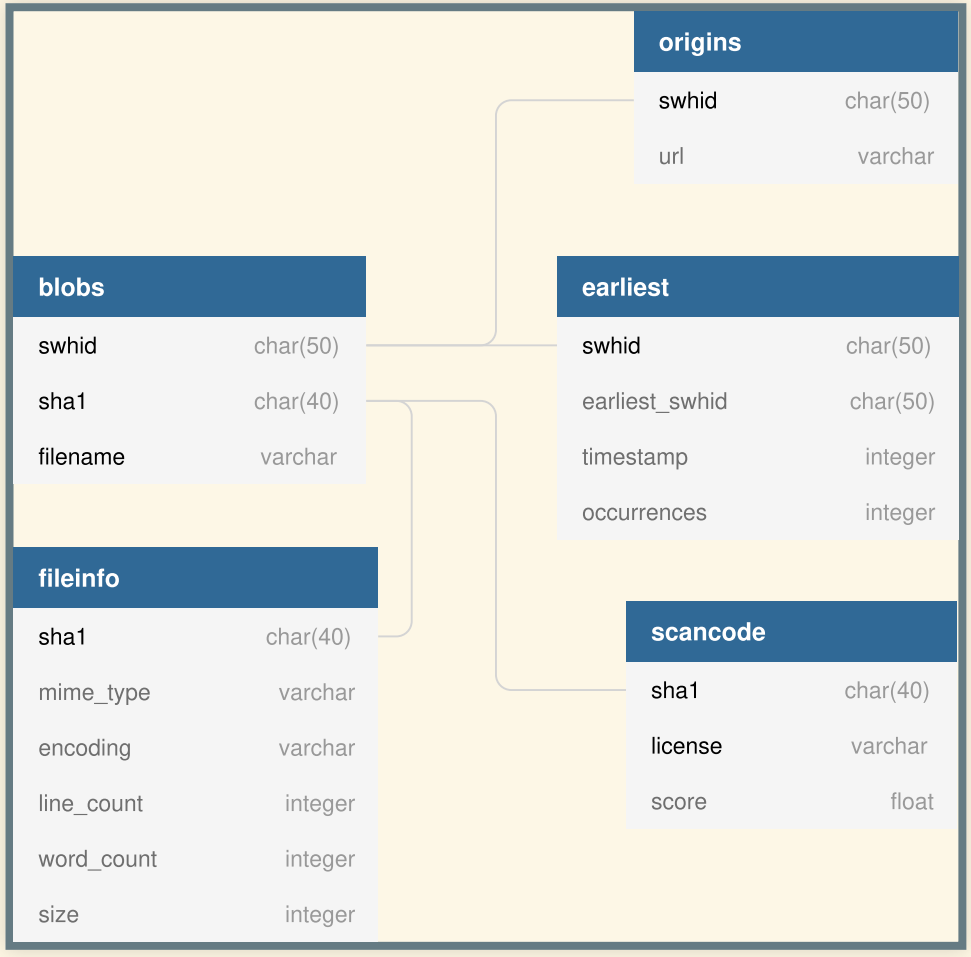
Use cases

1. Large-scale open-source license analysis
2. Training of license-detection tools (industry)
3. Natural Language Processing (NLP) analyses of legal/licensing corpora

Dataset construction



Data model


origins
swhid char(50)
url varchar
blobs
swhid char(50)
sha1 char(40)
filename varchar
earliest
swhid char(50)
earliest_swhid char(50)
timestamp integer
occurrences integer
fileinfo
sha1 char(40)
mime_type varchar
encoding varchar
line_count integer
word_count integer
size integer
scancode
sha1 char(40)
license varchar
score float

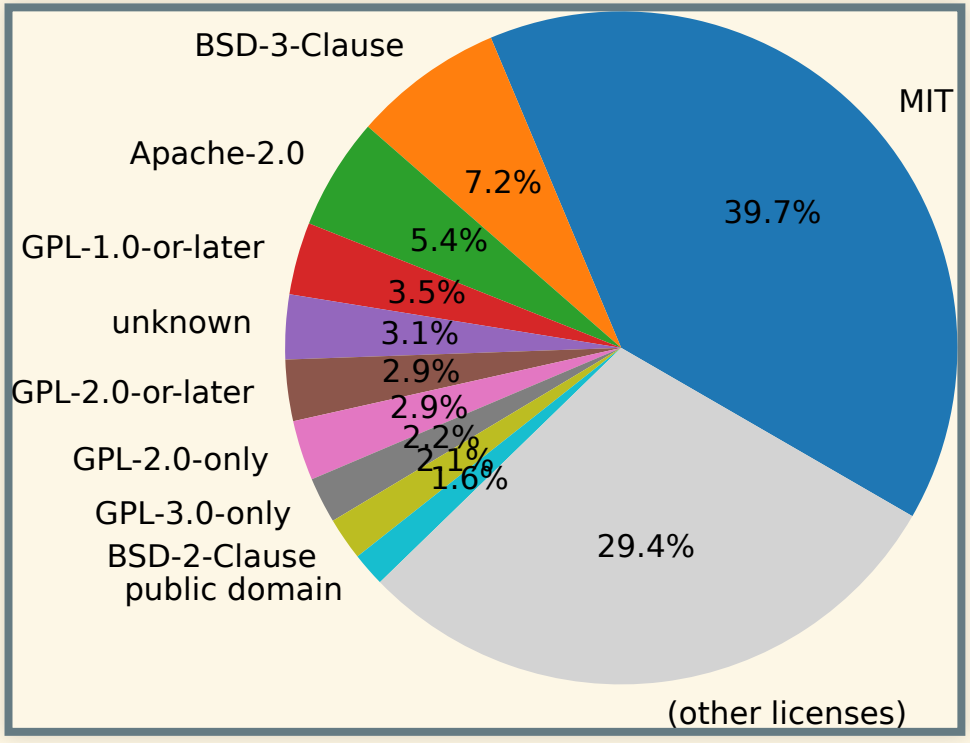
Sample usage (1) — Top words

```
words = pd.read_csv("blobs-wordfreqs.csv").sort_values(by="frequency", ascending=False)
stop_words = stopwords.words('english') + list(string.digits) + list(string.ascii_lowercas
interesting_words = words[~words["word"].isin(stop_words)]
interesting_words.nlargest(100, columns="frequency")
```

Word	Frequency
software	60’539’515
license	47’336’592
copyright	41’946’018
use	28’240’621
work	23’706’422

Sample usage (2) — Top variants by detected license

```
scancode = pd.read_csv("blobs-scancode.csv")
scancode["license"].value_counts().nlargest(10)
```



Sample usage (3) — (Non) textual license files

```
fileinfo = pd.read_csv("blobs-fileinfo.csv")
fileinfo["mime_type"].value_counts()
```

MIME type	Counts
text/plain	5’038’037
text/html	615’632
text/x-php	57’322
text/x-java	50’442
text/xml	43’816

Limitations

- **Noisy real-world data:** deliberate decision of casting a larger net (risking the inclusion of non-license files) + providing information to enable downstream filtering
- **Licenses != open-source licenses**, hence the “(Open Source)” in the title: not all included licenses are FSF/OSI/others-approved

Future work

- **Dataset maintenance:** multiple versions already available at https://annex.softwareheritage.org/public/dataset/license-blobs/
- **Include derived NLP/embedding representations** (for license blobs of textual types)

Learn more

Stefano Zacchiroli.
A Large-scale Dataset of (Open Source) License Text Variants.
In proceedings of The 2022 Mining Software Repositories Conference (MSR 2022), 23-24 May 2022, Pittsburgh, PA, USA. ACM 2022.

- paper DOI: 10.1145/3524842.3528491
- paper preprint: https://arxiv.org/abs/2204.00256
- dataset DOI: 10.5281/zenodo.6379164