Two Decades of the ACL Anthology

Development, Impact, and Open Challenges

Marcel Bollmann[1], Nathan Schneider[2], Arne Köhn[3], Matt Post[4]

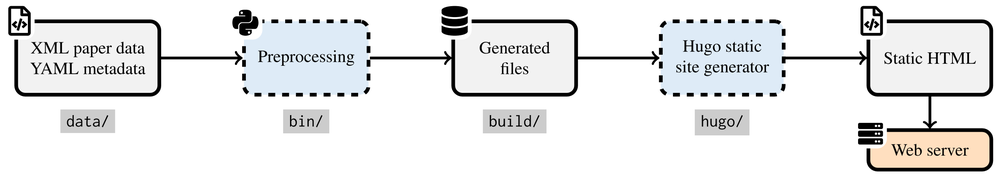
[1]Linköping University, [2]Georgetown University, [3]New Work SE, [4]Microsoft

- The ACL Anthology is a mostly **volunteer-driven** project.
- All development happens in a public **Github repository**.

  github.com/acl-org/acl-anthology/
- You can build the website locally — it's powered by **Python** and **Hugo**!
- All **metadata** is stored in the repo in XML and YAML formats.
  - Includes metadata for name merging and disambiguation, venue metadata, SIG metadata, and more.
- We introduce a new **Python library** for accessing this metadata easily.

  `pip install acl-anthology-py`
- Find the full **documentation** here:

  acl-anthology-py.readthedocs.io/



Simplified illustration of the ACL Anthology build pipeline, with folder names from the official Github repository.

```xml
<paper id="2">
  <title>Towards a Computational History of the <fixed-case>ACL</fixed-case>: 1980-2008</title>
  <author><first>Ashton</first><last>Anderson</last></author>
  <author><first>Dan</first><last>Jurafsky</last></author>
  <author><first>Daniel A.</first><last>McFarland</last></author>
  <pages>13–21</pages>
  <url hash="0fe03143">W12-3202</url>
  <bibkey>anderson-etal-2012-towards</bibkey>
</paper>
```

Example of the XML metadata for a paper (W12-3202) as stored in the ACL Anthology Github repository

```yaml
- canonical: {first: James H., last: Martin}
  variants:
  - {first: James, last: Martin}
- canonical: {first: Yang, last: Liu}
  comment: Edinburgh
  id: yang-liu-edinburgh
- canonical: {first: Yang, last: Liu}
  comment: 刘扬; Ph.D Purdue; ICSI, Dallas, Facebook, Liulishuo, Amazon
  id: yang-liu-icsi
```

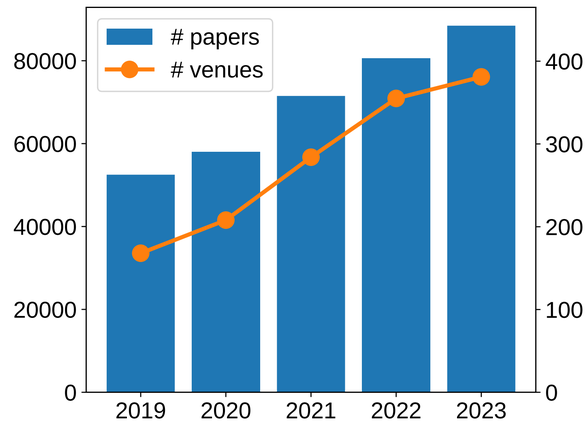
Example of the YAML metadata for name merging ("James Martin") and name disambiguation ("Yang Liu")

```python
# Instantiate the Anthology, automatically fetching data from the official Github repository
from acl_anthology import Anthology
anthology = Anthology.from_repo()

# Find all papers with "ACL Anthology" in the title, and print their BibTeX entries
for paper in anthology.papers():
    if "ACL Anthology" in str(paper.title):
        print(paper.to_bibtex())

# Find all people named "Bill Byrne" and print their ID + URL to all their paper PDFs
for person in anthology.find_people("Byrne, Bill"):
    for paper in person.papers():
        print(person.id, paper.pdf.url)
```

Example illustrating the usage of the acl-anthology-py Python library

ACL Anthology in Numbers

89,149	papers
87	gigabytes of data
388	venues
182	contributors on Github
1,250	pull requests merged
2,174	commits in the past 5 years



Growth of the ACL Anthology since 2019
(left y-axis: papers / right y-axis: venues)

Maybe *you* will contribute the next cool feature?

LINKÖPING UNIVERSITY