Two Decades of the ACL Anthology: Development, Impact, and Open Challenges

Marcel Bollmann
Linköping University
marcel.bollmann@liu.se

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Arne Köhn
New Work SE
arne.koehn@new-work.se

Matt Post
Microsoft
mattpost@microsoft.com

Abstract

The ACL Anthology is a prime resource for research papers within computational linguistics and natural language processing, while continuing to be an open-source and community-driven project. Since Gildea et al. (2018) reported on its state and planned directions, the Anthology has seen major technical changes. We discuss what led to these changes and how they impact long-term maintainability and community engagement, describe which open-source data and software tools the Anthology currently provides, and provide a survey of literature that has used the Anthology as a main data source.

1 Introduction

The ACL Anthology[1] is a repository for scientific contributions within computational linguistics and natural language processing maintained by the Association for Computational Linguistics (ACL). It currently hosts over 88k papers from from relevant conferences and journals within the field, including both ACL-sponsored and non-ACL venues, nearly 400 in total, a growth of almost 70% since 2019 (see Figure 1). It also includes many related materials such as software, posters, slides, and recordings of talks. All papers and materials are open-access, provided to the world without barrier under various open licenses.[2]

Development of the ACL Anthology takes place in a public repository,[3] which contains (i) metadata for all items in the Anthology, in XML and YAML formats; (ii) code for accessing and transforming this metadata, in form of a Python library and scripts; and (iii) code and templates for generating the ACL Anthology website. All code is made available under the permissive Apache-2.0 license.[4] Development has been almost entirely volunteer-driven, though since 2021 the ACL has funded assistants who have contributed to ingestions at the rate of about 20 hours a month.

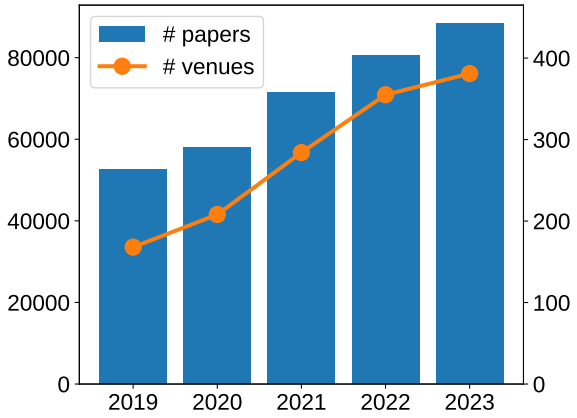

Figure 1: Growth of the ACL Anthology since 2019 as measured by the number of papers (left y-axis) and venues (right y-axis).

In this paper, we first describe the metadata currently provided by the ACL Anthology and efforts to improve it (§2); the technical framework and development of the website (§3); as well as a Python library for accessing data from the Anthology (§4). We then look at the impact the ACL Anthology has had on other open-source software projects and provide a survey of known datasets and studies that rely on the Anthology as a main data source (§5). Finally, we discuss lessons and challenges (§6) as well as future directions that we would like to see realized for the Anthology (§7), for which we rely on help from the community.

[1] https://aclanthology.org/

[2] ACL materials ingested since 2016 are CC BY 4.0; https://creativecommons.org/licenses/by/4.0/

[3] https://github.com/acl-org/acl-anthology/

[4] https://opensource.org/license/apache-2-0/

Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), pages 83–94

```
<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>
```

Figure 2: XML metadata for the paper by Anderson et al. (2012) as stored in the ACL Anthology repository.

2 Publication Metadata

At the core of the Anthology repository is the metadata on the publications it hosts. Here, we describe the different kinds of data provided by the Anthology, which efforts we have taken to enrich this data and ensure correctness, and how new materials are added to the Anthology.

2.1 Organization

At a high level, the papers in the Anthology are organized into *collections* of *volumes*. A *volume* is a set of related papers that would traditionally have been bound and published as a physical book. A *collection* is a group of volumes that were published at the same time under the same *venue*. Each collection is saved to a file in the data directory, e.g., `data/xml/2022.acl.xml` for ACL main proceedings volumes from 2022.

Each volume in a collection has a list of metadata, including its book title and its list of editors, which are typically the program chairs of a conference. It also notes the month, year, and address of the event of the event where the associated event was presented.[5] Volumes also identify their associated venues, and can be linked to multiple venues to denote joint events (cf. §2.5).

2.2 Paper Metadata

For all papers hosted on the Anthology, corresponding metadata can be found in the `data/xml/` folder of the repository. Figure 2 gives an example for the metadata of a single paper; it will contain, at a minimum, the *title* and *bibkey* (i.e., the bibliographic key as found in the official BibTeX export[6]) of the paper. *Authors* of a paper are stored with first and last name components clearly marked up so as to aid in formatting them correctly in bibliographic information. All files hosted on the Anthology server, such as a paper's PDF, will also have a *hash* attribute in the metadata, which is a simple CRC-32 checksum that can be used to verify any files downloaded from the Anthology. The full set of tags and attributes in the XML metadata is documented in the form of a RELAX NG schema (Clark, 2002).[7]

Paper metadata stored this way can be processed with any XML processing software or library. As the Anthology website is built from these XML files, the data is guaranteed to be identical to what users see online. GitHub CI checks automatically validate the XML files against the schema, ensuring that they always conform to the tags and attributes defined there.

Language	Est. Count
French	2195
Chinese	716
Portuguese	68
Swedish	34
Norwegian	33
Danish	32
German	7

Table 1: Estimated counts of non-English papers.

Languages At the time of this writing, the Anthology contains 85,324 papers with a `<title>` tag. The majority of these are written in English, but a few other languages are also represented. Such papers can be annotated with a `<language>` tag. As this tag is not yet systematically applied, we rely on heuristics to obtain estimated counts of non-English papers, shown in Table 1. The major venues for these papers are the JEP, RECITAL, TAL, and TALN venues (French), and the ROCLING, IJCLCLP, and CCL venues (Chinese).

Data Types Whereas older versions of the Anthology hosted only PDFs of papers/volumes and

[5] These fields were originally intended to note the date of publication and the publisher address, but have morphed in purpose over the years.

[6] https://aclanthology.org/anthology.bib.gz

[7] https://github.com/acl-org/acl-anthology/blob/master/data/xml/schema.rnc

their metadata, papers now support richer supplementary content, including slides, video, software, and data downloads. In case the paper itself needs to be corrected subsequent to publication, there is support for adding revisions of or errata for the original paper, as well as for retractions and removals.

Fixed-casing in Titles One of the most important parts of the site is its BibTeX export functionality. In the paper metadata input by authors upon submission and provided by publication venues to the Anthology for ingestion, many paper titles typically feature *title casing*—that is, capitalization of all content words. However, ACL bibliography style files call for *sentence casing*, in which only proper names (words that would be customarily capitalized even outside of a title) are capitalized, along with the beginning of the title. In order to avoid sloppy lowercasing of languages and other proper names, it is necessary to detect which letters in the original title should have their original casing protected in BibTeX entries, and which should be subject to alteration by the stylesheet. The Anthology codebase implements a set of heuristics based on wordlists to determine which characters should be flagged as *fixed-case* per English spelling conventions.[8] Approximately 45% of titles in the data contain at least one fixed-case designation.

The current heuristics were implemented in 2020, informed by reviewing the data in the Anthology at the time, and the wordlists are updated from time to time as new proper names are encountered. The main components of the heuristics are:

- The `truelist`, a set of 13k words and phrases that should have fixed capitals. The list was seeded with words commonly capitalized in abstracts, and augmented from gazetteers of names of languages and geopolitical entities, as well as manual additions. Salient entries that are not languages or places include "ACL Anthology", "Abstract Meaning Representation", "Carnegie", "Chinese Discourse Treebank", "Viterbi", and "Wizard of Oz".

- Lists of several adjectives and nouns commonly occurring as part of names whose capitalization should match the rest of the name. These are mostly geographic terms like "North", place descriptors like "University" and "Island", and "Ancient" and "Modern" (common modifiers in language names).

- General spelling rules, the most important of which are: (i) Any word with a capital letter in a non-initial position (e.g., "TextTiling", "QA") is marked as fixed-case. (ii) Any tokenized word consisting of a single uppercase letter other than "A", "K" or "N", or a single uppercase letter plus ".", is also fixed-case.

These rules are applied at ingestion time and marked with `<fixed-case>` tags in the XML. Skimming through the XML titles in recent proceedings, we find that errors are rare.[9] There are thus no plans to incorporate more sophisticated named entity recognition software.

2.3 Author Metadata

The Anthology website also provides author pages, which compile all items authored (or edited) by a given person. In contrast to paper metadata, information about authors is only *indirectly* stored in the XML, in the form of names attached to paper entries. This poses two challenges: (i) *names* need to be mapped to *identities*—this involves both merging, as the same person can have published under different names, and disambiguation, as different people can have the same name; and (ii) person identities need to be *indexed* in order to provide a mapping from people to their papers.

Name Merging and Disambiguation There are two forms of name ambiguity: (i) individual authors may publish papers under different variants of their name, and (ii) a particular name variant may be used by more than one person. To resolve (i), we compile a YAML metadata file (`data/yaml/name_variants.yaml`) to collapse known variants under a canonical representation, which is used for the author's page on the website. Additionally, we merge names automatically if they only differ in diacritics (e.g., *José* vs. *Jose*), as we find that in practice they almost always refer to the same person. To address (ii), or false positives from the merge heuristic, an *ID* can be assigned to create an author identity. This ID can then be used in a paper's `<author>` tag to link it to that identity. Figure 3 contains examples.

[8] https://github.com/acl-org/acl-anthology/tree/master/bin/fixedcase

[9] In the ACL 2023 proceedings, an example of a false positive is the English word "Even", which is also the name of a language. A false negative is "New Yorker".

```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
```

Figure 3: Example of YAML metadata for merging multiple surface forms of a name under a single canonical representation (*James H. Martin*), and for identifying a person with an ambiguous name (*Yang Liu*).

We typically use the Ph.D. institution as the disambiguator in the ID, but plan to move to an ORCID representation.

Author Indexing Finding metadata for a specific paper in the XML files is easy: for example, given a paper with the Anthology ID "2020.acl-main.699," its metadata will be located in the `2020.acl.xml` file under a `<volume>` with ID "main" and a `<paper>` with ID "699." This XML block could be retrieved with a single XPath expression. However, to find all papers authored by a given person, it is necessary to parse *all* XML files and look for instances of their name plus respective variants. This motivates the need for building an *index* that maps people to their (co-)authored papers. To avoid data redundancy, we do not store such an index in the repository directly, but rather compute it dynamically through a Python library specifically made for the Anthology, which we describe in §4.

2.4 Event Metadata

An *event* is a set of otherwise unrelated Anthology volumes that were presented together at a conference. Events are inferred from collections: each non-journal collection is assumed to have been presented in the real world. Additionally, an `<event>` block in a collection's XML file allows to note other volumes that were associated with that event (e.g., colocated workshop volumes).

Until 2023, events had no explicit representation. The Anthology now has the ability to represent event metadata and link to materials related to it, such as the conference handbook and videos from plenary talks and meetings. Importantly, we will also be able to generate citations for these materials. Completion of this work is planned for this year.

2.5 Venue and SIG Metadata

Every volume is linked to a venue using a `<venue>` tag in the volume's metadata. Every venue has its own file under `data/yaml/venues` listing key information about that venue, including its name (e.g., "Conference on Machine Translation"), its acronym ("WMT"), and tags determining whether it belongs to ACL and whether it is displayed on the main page. A URL-friendly "slug" containing only lowercased, alphanumeric characters is constructed from the venue acronym. Similarly, workshop volumes can be associated with an ACL Special Interest Group (SIG). Information on these can be found under `data/yaml/sigs`. Each SIG file lists all the volumes associated with that SIG.

2.6 Ingestion of New Materials

Ingestion refers to the process of importing new materials in the Anthology. A single ingestion typically includes the main volumes of a conference (e.g., ACL long and short papers, tutorials, system demonstrations, and its student research workshop) together with its colocated workshops. Publication chairs compile the proceedings in a single, formatted directory and submit them to the Anthology, where they will be assembled in a branch of the GitHub repository and submitted as a pull request for review by the Anthology team.

As of 2022, the preferred ingestion format is ACLPUB2,[10] a modernization of ACLPUB.[11] The Anthology has also developed scripts for a number of other ingestion formats, including for the TACL and CL journals, for the MT Archive, and for a generic TSV format. These scripts can be found under `bin/ingest_*.py`.

3 Website

Most users interact with the ACL Anthology through its website. Here, we describe how the website is built, the technical developments it has seen in recent years, and new features we have introduced for the community.

3.1 Static Rewrite

The Anthology website underwent a major rewrite in early 2019, switching from a dynamic to a fully static site. Gildea et al. (2018) describe the technical framework of the Anthology prior to this

[10] https://github.com/rycolab/aclpub2
[11] https://github.com/acl-org/ACLPUB

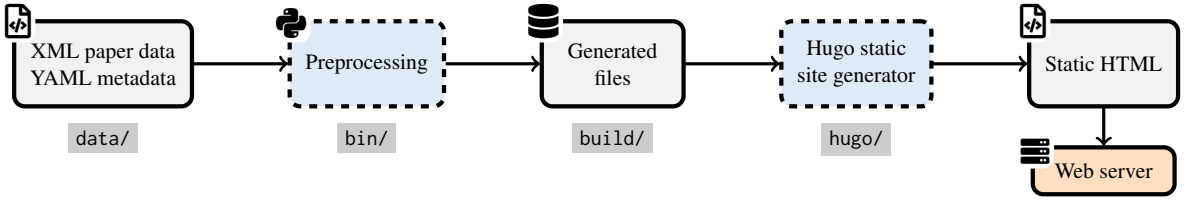

Figure 4: Simplified illustration of the ACL Anthology build pipeline, with the folders where relevant files can be found in the official repository (at https://github.com/acl-org/acl-anthology/).

rewrite, which consisted of a Ruby on Rails application powered by a PostgreSQL database and the Apache Solr search platform. New data could be added to the Anthology via an intermediate XML format containing all bibliographic metadata for proceedings volumes and the contained papers. These XML files had to be ingested into the PostgreSQL database and indexed by the Solr engine.

This new static site generation approach is illustrated in Figure 4. The XML files with the bibliographic metadata (cf. §2.2) now constitute the primary data source for the Anthology; there is no derived database. Some additional metadata, e.g. for publication venues (cf. §2.5) and for disambiguating author names (cf. §2.3), is also stored in YAML files. For building the website, the data files are first processed by a number of Python scripts using an internal Python library (described in §4), which generate page stubs for the site generator and convert the bibliographic data to a number of export formats (e.g. BibTeX). Afterwards, the website is built using the site generator Hugo,[12] resulting in entirely static HTML files. A CI/CD pipeline on GitHub automatically builds and uploads these to the production web server. To provide search functionality, the website embeds a custom search using Google's Programmable Search Engine (PSE).[13] Notably, the PDF files of the papers, as well as any supplementary material, are *not* part of the repository or the build pipeline as they constitute an enormous volume of data;[14] they are currently copied over manually to the web server.

Advantages Using a static website offers many performance benefits. Since there is no database backend that needs to be queried, the user experience when browsing the website is considerably faster. Building the website locally is also faster: with the former approach, seeding the database would take "at least 30 minutes,"[15] while a full build of the static website takes around nine minutes on a modern laptop.[16] Most of this speed comes from using Hugo as the site generator, which takes 133 seconds to generate 165k HTML pages (i.e., ≈0.8 ms per page).

The complexity of running the Anthology is also greatly reduced, which both makes maintenance easier (the production system only needs a web server and no other software) and lowers the barrier of entry, as potential contributors do not need to set up any services to test their contributions. The complete Anthology can be built and served with a single `make` call, provided that Python and Hugo are installed on the system. Generating a static website means that it is trivial to serve copies of the Anthology, which is extensively used in the continuous integration (CI) pipeline; every pull request is rendered to a preview website and the effect on the production system can be easily checked.

Search Functionality Providing search functionality on the website is difficult in our simplified static setting.[17] Google's PSE provides search within both HTML pages and PDF files, but also comes with a number of drawbacks: (i) customization options are limited; e.g., there is no way to display a paper's landing page and its PDF as a single item in the search results, even though they logically belong together; (ii) no immediate control over the indexing of new or changed items, leading to delays of updates being reflected in the search; (iii) region-blocking of Google Search making the search unusable for affected users, e.g. in China. Addressing these issues is challenging and most likely requires either (i) re-introducing a dynamic

[12] https://gohugo.io/
[13] https://programmablesearchengine.google.com/
[14] 86,819 files consuming 87 GB.
[15] cf. https://github.com/acl-org/acl-anthology/blob/4a751ac/README.md?plain=1#L58; also note that at this point in time, the Anthology had less than half the number of papers it has now.
[16] Tested on AMD Ryzen 7 Pro 5850U, 12 GB RAM, with Python 3.11.4 and Hugo v0.115.3.
[17] For further discussion, also see https://github.com/acl-org/acl-anthology/issues/165.

server component to host our own search platform, which comes with an increased maintenance burden; or (ii) using a commercial search-as-a-service platform, which comes with a financial cost.

3.2 Features

Citation Export Formats Each paper's page provides a number of user-friendly citation formats beyond BibTeX, including Markdown and EndNote. These are available for download or for one-click copying. We also provide preformatted long and informal textual citation formats.

GitHub Issue Tracker Corrections to the metadata can be requested by anyone by opening a Github issue. We make extensive use of issue templates to facilitate opening and handling a range of common corrections and suggestions from users.

Zotero Integration Zotero[18] is a popular open source reference manager that allows users to import scholarly content and metadata as they browse the web. Content from the Anthology is easily imported into Zotero, with metadata parsed from BibTeX files.[19]

Papers With Code Integration Papers With Code (PWC)[20] is a website that links papers to related datasets and code repositories. For papers in the ACL Anthology, data is fetched from PWC through an API and automatically merged into the Anthology metadata; dataset and code links provided by PWC are subsequently displayed on the Anthology website.

Paper Awards The Anthology marks papers that have received awards from their respective conferences. A page compiling all awarded papers does currently not exist, but could be a future addition.

4 Python Library

Building the Anthology website requires transforming all the metadata described above into a format suitable for generating the HTML pages using Hugo. To do this, we rely on a custom-built Python library that is currently found in the `bin/anthology/` folder of the repository. Besides wrapping access to the XML and YAML data files, the Python library (i) implements author indexing and name disambiguation functionality (cf. §2.3); (ii) converts markup found in paper titles and abstracts into appropriate representations for HTML or LaTeX; (iii) converts a limited subset of LaTeX expressions in paper titles and abstracts into appropriate Unicode and/or HTML representations; (iv) generates bibliographic information (e.g. BibTeX entries).

4.1 Adoption of the Library

Like the other parts of the repository, the library is open-source and free to use for anyone wanting to access the Anthology data. As with the metadata files, the library is used to build the ACL Anthology website, so re-using it is guaranteed to provide data identical to that on the website. In practice, however, we see some obstacles for a wider adoption of this library (e.g. for projects such as those surveyed in §5.3). One is a lack of proper documentation; while Python scripts found in the repository's `bin/` folder can serve as concrete examples for how the library is used, individual functions are often undocumented. This has consequences not just for third-party adoption, but also for maintainability (cf. §6). Another is the partly unintuitive interface of the library; when it was first built in 2019, the top priority was to recreate the exact functionality of the Ruby application that existed at the time (cf. §3.1), and as such, it is mainly geared towards the needs of building the Anthology website. For these reasons, we are in the process of reimplementing this library.

4.2 PyPI Package

Based on the challenges touched upon above (and further discussed in §6), we have begun reimplementing the Python library in a way that (i) is user-friendlier and better documented, and (ii) easily installable (e.g. via `pip`). A fully usable version of this package is already available on PyPI, the main Python package repository, as `acl-anthology-py`.[21] Figure 5 shows some examples of how this library might be used. We are currently working on making this library feature-complete with respect to the functionality needed to replace the old library in the website's build chain (cf. §3.1), but the current version of the library already comes with full API documentation as well as a user guide.[22] Furthermore, it is implemented

[18] https://www.zotero.org/

[19] Implemented by Guy Aglionby: https://github.com/zotero/translators/blob/master/ACLWeb.js

[20] https://paperswithcode.com/

[21] https://pypi.org/project/acl-anthology-py/

[22] Please refer to the PyPI page for the latest link.

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

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

# Find all people named "Dan Klein", and pick the first one (-- as of now, there's only one)
person = anthology.find_people("Klein, Dan")[0]

# Get a list of URLs to all paper PDFs by a given person
urls = [paper.pdf.url for paper in person.papers()]

# Generate the BibTeX entry of a paper based on its Anthology ID
bibtex = anthology.get("2020.acl-main.699").to_bibtex()
```

Figure 5: Examples illustrating the usage of the `acl-anthology-py` Python library. Please refer to the latest API documentation for the most up-to-date information.

using many "best practices" of software development, including a high test coverage (>90%) and automated CI checks that enforce coding style conventions and type hints.

We hope that this redesigned library will make the ACL Anthology easier to maintain and thus more future-proof. Releasing the library on PyPI should also greatly increase discoverability and, consequently, adoption in related work that wants to access Anthology data.

5 Impact of the ACL Anthology

A lot of research has built on data from the ACL Anthology over the years, but even the technical infrastructure has had impact on other open-source projects. Here, we try to provide an extensive survey of software, datasets, and scientific studies that directly rely on data or code from the Anthology.

5.1 On Open-Source Software

Being open-source and easy to use made it possible for other publication repositories to re-use the ACL Anthology infrastructure. At least three such projects currently exist: (i) the SemDial workshop series,[23] which publishes its proceedings dating back to 2004; (ii) the IR Anthology,[24] which currently collects ≈63k papers from information retrieval venues (Potthast et al., 2021); and (iii) the Global Water Futures archive,[25] which hosts 1.2k publications from their project on addressing water threats. The last project in particular highlights that the codebase of the ACL Anthology has even found adoption outside of computer science domains.

5.2 On Corpora and Datasets

The ACL Anthology Reference Corpus (ACL ARC; Bird et al., 2008) was one of the first efforts to build on the ACL Anthology for academic research, providing the extracted full-text and metadata for 11k papers up to February 2007. The ACL Anthology Network Corpus (AAN; Radev et al., 2009) expands on this by providing citation and collaboration networks. Schäfer et al. (2011) introduce the ACL Anthology Searchbench, which Weitz and Schäfer (2012) build on to provide a "citation browser." Singh et al. (2018) present CL Scholar, an ACL Anthology "knowledge graph miner." Unfortunately, as of now, most of these initiatives appear to be abandoned and/or unavailable; the ANN is still accessible through the broader "All About NLP" project (also AAN; Fabbri et al., 2018).

More recently, the NLP4NLP corpus (Mariani et al., 2019a) incorporates data from the ACL Anthology as part of a dataset of articles in "speech and natural language processing over a period of 50 years (1965–2015)," which Mariani et al. (2022) extend to cover publications until 2020. The NLP Scholar project combines data from the ACL Anthology and Google Scholar in a new dataset (Mohammad, 2020b) and an associated visual exploration tool (Mohammad, 2020c). NLPExplorer (Parmar et al., 2020) offers a curated dataset and web portal[26] with annotations including manually curated topic classification. Finally, the ACL

[23] http://semdial.org/anthology/venues/semdial/
[24] https://ir.webis.de/anthology/
[25] https://gwf-uwaterloo.github.io/gwf-publications/
[26] http://nlpexplorer.org/

OCL corpus (Rohatgi et al., 2023) provides automatically extracted text for ACL Anthology papers currently up to September 2022.

Several other annotated datasets based on subsets of the Anthology exist: Schäfer et al. (2012) present a small corpus of 266 papers annotated with coreference; QasemiZadeh and Schumann (2016) introduce a dataset for terminology extraction and classification; Gábor et al. (2016) add semantic annotation such as entity tagging and relations, which was subsequently used in a SemEval shared task (Gábor et al., 2018); Iwatsuki et al. (2020) build a dataset with formulaic expressions and their communicative functions; van Dongen et al. (2020) add citation information; Hao et al. (2020) introduce a corpus annotated with "future work sentences"; and Hou et al. (2021) present a corpus for entity tagging of tasks, datasets, and evaluation metrics in 30k ACL Anthology papers.

5.3 On Academic Research

The ACL Anthology has frequently been used for scholarly literature analysis within the NLP domain, such as citation analysis. As the Anthology does not provide any data on citations, such studies require either data mining of the PDFs, a combination with data from external sources such as Google Scholar or Semantic Scholar,[27] or the use of a dataset that already provides this, like NLP Scholar (Mohammad, 2020b). Despite this necessary extra step, studies have focused on the ACL Anthology to analyze incoming citations (Mohammad, 2020a), outgoing citations (Bollmann and Elliott, 2020; Singh et al., 2023), or geographic citation gaps (Rungta et al., 2022).Van Dongen et al. (2020) present a model for citation count prediction. Guo et al. (2020) provide a Java API for extracting citation context from academic literature, trained on an annotated dataset based on the Anthology.

Beyond citation analysis, studies have used the Anthology to analyze the gender distribution among authors (Vogel and Jurafsky, 2012), to analyze the influence of industry on academic research (Abdalla et al., 2023), and to track the evolution of research topics and domains over time (Anderson et al., 2012; Omodei et al., 2014; Schumann, 2016; Mariani et al., 2019b; Schopf et al., 2023). Joshi et al. (2020) combine data from the Anthology and the Semantic Scholar API to analyze linguistic diversity in NLP research. Fortuna et al. (2021) analyze 50k Anthology papers to find instances of "NLP for Social Good."

Data from the ACL Anthology has also been used to build and evaluate models for a variety of tasks, such as relation extraction (Schäfer et al., 2008), scientific term mining (Jin et al., 2013), name disambiguation and topic modeling (King et al., 2014), semantic labeling (Schumann and Martínez Alonso, 2018), building search systems (Yoneda et al., 2017; Ding et al., 2020), and analyzing document similarity (Ostendorff et al., 2020).

6 Lessons & Challenges

Impact of Open Development All development – including changes to the data – happening in public is highly beneficial. By far the most common contributions by non-project members are metadata corrections such as name changes and typos. Some of these come directly as pull requests (≈1k so far), which can often be merged within a day and reduce the workload of the Anthology volunteers; others come as (pre-formatted) GitHub issues (also ≈1k so far), which are then automatically linked to pull requests that are merged monthly.

As everyone can build the Anthology themselves, we occasionally get feedback or suggestions regarding the robustness of the infrastructure on different systems, but overall contributions to the code parts of the project are rare. One notable exception is the integration with Papers With Code (PWC): this functionality was suggested and in large parts contributed by the team behind PWC.

Discoverability and Usability of Open-Source Data and Software There is a discrepancy between the data and tools that the ACL Anthology directly and openly provides and what researchers use in practice. While not all the studies mentioned in §5.3 are explicit about how they obtain data from the Anthology, we do find instances of using web-scraping tools on the Anthology website or using the BibTeX export as the primary data source. In both situations, we would expect that using the XML metadata (cf. §2.2) and/or the Python library (cf. §4) would be a faster and potentially easier[28] way to obtain the same results. The fact that several studies build their own solutions for this points to a problem of *discoverability* or *usability* of the data

[27] https://www.semanticscholar.org/

[28] For example, the BibTeX export contains TeX-encoded characters such as `\"{a}` for the letter ä, while the XML contains Unicode strings.

and tools we provide; i.e., people are either not aware that these data and tools exist, or they find them too hard to use (e.g. due to lack of documentation).

Encouraging Contributions All of the features the ACL Anthology has today are the result of volunteer effort, yet we rarely see new volunteers contributing features to the code. Potentially, one factor in this is (again) the lack of documentation: while there is some information distributed across a GitHub Wiki, code-internal comments, and even the ACL Anthology website,[29] it is neither very accessible nor complete. Such "documentation debt" is known to cause problems for maintainability (e.g., Rios et al., 2020), and is likely to pose a hurdle for potential new contributors as well.

7 Ongoing and Future Work

Automatic Name Disambiguation An increasing problem with the growth of the field is author name disambiguation. As described in §2.3, our XML format supports maintaining separate identities for authors with the same name, but currently these names have to be manually disambiguated at ingestion time. Ingestion input materials often include disambiguating information such as ORCIDs, author affiliation, and so on, but we cannot depend on their presence. A great project (likely publishable) would be to build an automated disambiguation process based on metadata, context (such as co-author lists), and paper content itself.

Incomplete Venue/SIG Information Prior to the new Anthology ID format introduced in 2020, all workshop venues were grouped together under a common `W` prefix, without more specific venue information. Many of these have been manually linked to venues (e.g., all early WMT volumes), but the information is incomplete. These could similarly be linked to SIGs.

Abstracts for Older Papers Since around 2016, papers start systematically including abstracts in the metadata, but most older papers do not have them. Some datasets (e.g., Rohatgi et al., 2023) provide text extracted automatically from the PDFs, potentially enabling us to add abstracts for older papers too. However, some form of quality check of the extracted text appears necessary to maintain a high level of quality of our metadata.

[29] https://aclanthology.org/info/

Copyright The copyright situation with Anthology data has operated under the best efforts of non-expert volunteers, largely from academia. However, the licensing information is imperfect. The ACL owns the copyright to most papers submitted to ACL venues, but there are exceptions where the ACL was only granted a license, such as for papers published by Canadian and British academics, for whom copyright belongs to the English Crown. The ideal situation would be to incorporate this information in paper-level metadata.

Incomplete Older Volumes Many early volumes are missing, and it is not always even known which ones they are.

Front Page Redesign The front page of the Anthology is showing its age, and could use reworking from a graphic designer and/or front-end developer.

8 Conclusion

The ACL Anthology has grown into an invaluable resource for the CL/NLP community. Volunteer contributions are responsible for virtually all of the improvements of the metadata (§2), the Anthology website (§3), and the provided Python library (§4). The survey in §5 highlights the utility and importance of this resource for academic research. To address the ongoing challenges (§6) and future directions (§7), we continue to rely on help from the community. We encourage anyone who is interested in contributing to the ACL Anthology to explore our GitHub repository.[30]

Acknowledgements

We would like to thank everyone who has contributed to the ACL Anthology over the years,[31] and thereby helped to make it into what it is today.

Limitations

The literature survey in §5 was performed by searching for papers with "ACL Anthology" in the title or abstract, as well as inspecting references in and citations of these papers. It is conceivable that there is more relevant work that we missed. Likewise, there may be more open-source projects based off the Anthology that we are not aware of.

[30] https://github.com/acl-org/acl-anthology/

[31] An imperfect overview of contributors can be derived from: https://github.com/acl-org/acl-anthology/graphs/contributors

References

Mohamed Abdalla, Jan Philip Wahle, Terry Lima Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the ACL: 1980–2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Korea. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marcel Bollmann and Desmond Elliott. 2020. On forgetting to cite older papers: An analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7819–7827, Online. Association for Computational Linguistics.

James Clark. 2002. RELAX NG compact syntax. Committee specification, The Organization for the Advancement of Structured Information Standards [OASIS].

Shane Ding, Edwin Zhang, and Jimmy Lin. 2020. Cydex: Neural search infrastructure for the scholarly literature. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 168–173, Online. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. TutorialBank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–620, Melbourne, Australia. Association for Computational Linguistics.

Paula Fortuna, Laura Pérez-Mayos, Ahmed AbuRa'ed, Juan Soler-Company, and Leo Wanner. 2021. Cartography of natural language processing for social good (NLP4SG): Searching for definitions, statistics and white spots. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 19–26, Online. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL Anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3694–3701, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Chenrui Guo, Haoran Cui, Li Zhang, Jiamin Wang, Wei Lu, and Jian Wu. 2020. SmartCiteCon: Implicit citation context extraction from academic literature using supervised learning. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 21–26, Wuhan, China. Association for Computational Linguistics.

Wenke Hao, Zhicheng Li, Yuchen Qian, Yuzhuo Wang, and Chengzhi Zhang. 2020. The acl fws-rc: A dataset for recognition and classification of sentence about future works. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 261–269, New York, NY, USA. Association for Computing Machinery.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. 2020. An evaluation dataset for identifying communicative functions of sentences in English scholarly papers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the ACL Anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ben King, Rahul Jha, and Dragomir R. Radev. 2014. Heterogeneous networks and their applications: Scientometrics, name disambiguation, and topic modeling. *Transactions of the Association for Computational Linguistics*, 2:1–14.

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2019a. The NLP4NLP corpus (I): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3.

Joseph Mariani, Gil Francopoulo, Patrick Paroubek, and Frédéric Vernier. 2019b. The NLP4NLP corpus (II): 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3.

Joseph Mariani, Gil Francopoulo, Patrick Paroubek, and Frédéric Vernier. 2022. NLP4NLP+5: The deep (r)evolution in speech and language processing. *Frontiers in Research Metrics and Analytics*, 7.

Saif M. Mohammad. 2020a. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.

Saif M. Mohammad. 2020b. NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 868–877, Marseille, France. European Language Resources Association.

Saif M. Mohammad. 2020c. NLP scholar: An interactive visual explorer for natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 232–255, Online. Association for Computational Linguistics.

Elisa Omodei, Jean-Philippe Cointet, and Thierry Poibeau. 2014. Mapping the natural language processing domain: Experiments using the ACL Anthology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2972–2978, Reykjavik, Iceland. European Language Resources Association (ELRA).

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Monarch Parmar, Naman Jain, Pranjali Jain, P Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. NLPExplorer: Exploring the universe of NLP papers. In *Advances in Information Retrieval*, pages 476–480, Cham. Springer International Publishing.

Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. 2021. The Information Retrieval Anthology. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2550–2555, New York, NY, USA. Association for Computing Machinery.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

Nicolli Rios, Leonardo Mendes, Cristina Cerdeiral, Ana Patrícia F. Magalhães, Boris Perez, Darío Correal, Hernán Astudillo, Carolyn Seaman, Clemente Izurieta, Gleison Santos, and Rodrigo Oliveira Spínola. 2020. Hearing the voice of software practitioners on causes, effects, and practices to deal with documentation debt. In *Requirements Engineering: Foundation for Software Quality*, pages 55–70, Cham. Springer International Publishing.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: advancing open science in computational linguistics. arXiv:2305.14996.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 7–13, Portland, Oregon. Association for Computational Linguistics.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL Anthology. In *Proceedings of COLING 2012: Posters*, pages 1059–1070, Mumbai, India. The COLING 2012 Organizing Committee.

Ulrich Schäfer, Hans Uszkoreit, Christian Federmann, Torsten Marek, and Yajing Zhang. 2008. Extracting and querying relations in scientific papers on language technology. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tim Schopf, Karim Arabi, and Florian Matthes. 2023. Exploring the landscape of natural language processing research. arXiv:2307.10652.

Anne-Kathrin Schumann. 2016. Brave new world: Uncovering topical dynamics in the ACL Anthology reference corpus using term life cycle information. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Anne-Kathrin Schumann and Héctor Martínez Alonso. 2018. Automatic annotation of semantic term types in the complete ACL Anthology reference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. Forgotten knowledge: Examining the citational amnesia in NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

Mayank Singh, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, Rajarshi Haldar, Pawan Goyal, and Animesh Mukherjee. 2018. CL scholar: The ACL Anthology knowledge graph miner. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 16–20, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas van Dongen, Gideon Maillette de Buy Wenniger, and Lambert Schomaker. 2020. SChuBERT: Scholarly document chunks with BERT-encoding boost citation count prediction. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 148–157, Online. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Benjamin Weitz and Ulrich Schäfer. 2012. A graphical citation browser for the ACL Anthology. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1718–1722, Istanbul, Turkey. European Language Resources Association (ELRA).

Takuma Yoneda, Koki Mori, Makoto Miwa, and Yutaka Sasaki. 2017. Bib2vec: Embedding-based search system for bibliographic information. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–115, Valencia, Spain. Association for Computational Linguistics.