The ACL Anthology: Current State and Future Directions

Daniel Gildea
Department of Computer Science
University of Rochester
gildea@cs.rochester.edu

Min-Yen Kan
School of Computing
National University of Singapore
kanmy@comp.nus.edu.sg

Nitin Madnani
Educational Testing Service
Princeton, NJ
nmadnani@ets.org

Christoph Teichmann and **Martín Villalba**
Department of Language Science and Technology
Saarland University
villalba@coli.uni-saarland.de

Abstract

The Association of Computational Linguistic's Anthology is the open source archive, and the main source for computational linguistics and natural language processing's scientific literature. The ACL Anthology is currently maintained exclusively by community volunteers and has to be available and up-to-date at all times. We first discuss the current, open source approach used to achieve this, and then discuss how the planned use of Docker images will improve the Anthology's long-term stability. This change will make it easier for researchers to utilize Anthology data for experimentation. We believe the ACL community can directly benefit from the extension-friendly architecture of the Anthology. We end by issuing an open challenge of reviewer matching we encourage the community to rally towards.

1 Introduction

The ACL Anthology[1] is a service offered by the Association for Computational Linguistics (ACL) allowing open access to the proceedings of all ACL sponsored conferences and journal articles. As a community goodwill gesture, it also hosts third-party computational linguistics literature from sister organizations and their national venues. It offers both text and faceted search of the indexed papers, author-specific pages, and can incorporate third-party metadata and services that can be embedded within pages (Bysani and Kan, 2012). As of this paper, it hosts over 43,000 computational linguistics and natural language processing papers, along with their metadata. Over 4,500 daily requests are served by the Anthology. The code for the Anthology is available at https://github.com/acl-org/acl-anthology under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License[2]. Slightly different from the Anthology source code, ACL also licenses its papers with a more liberal license, supporting Creative Commons Attribution 4.0 International License[3], supporting liberal re-use of papers published with the ACL.

The maintenance of the code and the website is handled through volunteer efforts coordinated by the Anthology editor. Running a key service for the computational linguistics community that needs to be continuously available and updated frequently is one of the main issues in administering the Anthology.

We discuss this issue along with the challenges of running a large scale project on a volunteer basis and its resulting technical debt. As we look towards the future, previous research has shown that it can also be used as a data source to characterize the work and workings of the ACL community (Bird et al., 2008; Vogel and Jurafsky, 2012; Anderson et al., 2012). Extensions to the Anthology that build on this information could make the Anthology an even more valuable resource for the community. We will discuss two possibl eextensions – anonymous pre-prints and support for finding relevant submission reviewers by linking au-

[1] https://aclanthology.info/

[2] https://creativecommons.org/licenses/by-nc-sa/3.0/

[3] https://creativecommons.org/licenses/by/4.0/

Proceedings of Workshop for NLP Open Source Software, pages 23–28
Melbourne, Australia, July 20, 2018.

thors in the Anthology with their research interests and community connections. Beyond being useful in itself, work on such challenges has the potential to motivate the ACL community to further support the Anthology.

2 Current State of the Anthology

The ACL Anthology was proposed as a project to the ACL Executive by Steven Bird at the 2001 ACL conference and first launched in 2002, with a second version developed in 2012, commissioned by the ACL committee. Steven Bird also served as the first editor of the anthology from 2002 to 2007, a post which Min-Yen Kan took over in 2008 and continues to fill as of today. The Anthology provides access to papers in Portable Document Format (PDF) as well as the associated metadata in multiple formats (e.g., BIBTEX and Endnote). For recent papers, authors can also opt include data, notes and open-source software, and may provide Digital Object Identifiers (DOIs) for permalinking the citations within their papers.

The technology behind the current version is detailed in Table 1. As a community project, daily administration and development is handled by volunteers. However, to tackle larger problems with the Anthology which require a more focused effort, the ACL committee has solicited paid assistance. Hosting and bandwidth for the Anthology has historically been provided by universities free of charge. It was hosted at the National University of Singapore until the spring of 2017, when it was migrated to its current home at Saarland University. In the future, hosting duties are planned to fall under the umbrella of the ACL itself, unifying all services under https://www.aclweb.org/portal/.

Framework	Ruby on Rails
Search engine	Solr
Database	PostgreSQL
Web server (Prod./Test)	Nginx / Jetty
Operating System	Debian GNU-Linux

Table 1: Tech stack for the ACL Anthology.

The most important task is the importing, indexing and provisioning of newly accepted papers from recent conference proceedings and journal issues. The original Anthology defined an XML format for simple bibliographic metadata, which has been extended to support the more recent features of associated software, posters, videos and datasets that accompany the scholarly publications. Providing the XML for new materials is an semi-automated process that is largely integrated with the various mechanisms for managing ACL conference submissions and printed proceedings. It is straightforward for ACL events that utilize the licensed START conference management software[4], as an established software pipeline builds upon the artefacts used for creation of the final publications themselves. After the accepted papers are finalized, START produces an archive file of camera-ready PDF files and author-provided metadata such as the title, author list, and abstract for each paper. These files are processed by a set of scripts in START maintained by ACL publication chairs in order to assign page numbers to papers, and to produce a PDF proceedings volume for each conference complete with a table of contents, author index, and other front matter. These scripts also produce bibliographic information that are programmatically transformed into the ACL Anthology's XML format. The Anthology is then updated with the author-provided PDFs and the XML metadata. For importing journal articles and venues not using the START submission system, additional manual work is necessary to construct the Anthology XML. Sanity checks and some manual curation is also necessary to deal with issues such as character encodings and accents in names, multipart family names, and so on. This pipeline has reached a point of high efficiency, but may need to be adapted if the ACL ever considers it necessary to integrate with a different service for conference organization.

3 Running the Anthology as a Community Project

Since the Anthology is not tied to a specific research project or institution, contributors that work on Anthology-related system administration and development tasks have been recruited in response to calls for volunteers at the main ACL conferences. In contrast, new features have been developed by researchers using the ACL Anthology as a resource in their own work, unconnected with the daily operation of the Anthology. Such research deliverables include, for example, the creation of a corpus of research papers (Bird et al., 2008), an author citation network (Radev et al., 2013) or a

[4] https://www.softconf.com/

faceted search engine (Schäfer et al., 2012; Buitelaar et al., 2014). These factors, in combination with the multiple, changing responsibilities and shifting research interests of community members, mean that new volunteers join and leave the Anthology team in unpredictable and sporadic patterns. Preserving knowledge about the Anthology's operational workflow is thus one of the most important challenges for the Anthology.

The Anthology editor has played a key role ensuring the continuity of the entire project. This position has so far always been filled for multiple years, longer than the normal time frame for an ACL officer. The role has been critical in ensuring a smooth transition between volunteers, at the cost of a long term with a heavy workload and a potential single point of failure. In order to tackle both issues, there is currently a concerted effort to improve the documentation of all tasks related to maintaining the Anthology.

As the ACL community and its publishing needs continue to grow, the ACL Executive is considering commercial support for publishing. While this may be suitable for help with daily operations, we strongly advocate the continuation and promotion of a closely-knit volunteer group for development. Passing the responsibilities for the Anthology to a commercial devoid who has no intrinsic interest in the Anthology's scientific contents may end up poorly.

4 Future Proofing the Anthology

All code, documentation, bug reports, and feature requests are hosted at `https://github.com/acl-org/acl-anthology`, along with instructions detailing the steps required to set up an instance of the Anthology and keep it updated with proceedings for new conferences. These instructions have been verified and updated using test builds. We began with the initial documentation provided by experienced contributors to the project and the original developer. New volunteers were then asked to set up and update a new instance of the Anthology on a new server while communicating with more experienced contributors. The documentation was expanded and updated based on the problems and questions encountered during this process. The resulting documentation will likely reduce the learning curve for new volunteers and will make their recruitment easier. It will also make it easier to migrate the Anthology to new servers when the hosting arrangement changes or to create mirrors. The latter is an important future task for the Anthology in order to ensure that alternatives are available if the main Anthology server experiences any downtime.

The current implementation of the Anthology has been extended over the years with minor improvements to functionality and bug fixes. The core code has remained mostly intact from its original version and has proved to be robust and reliable. However, fearing the introduction of bugs and instability (Spolsky, 2000), the maintainers chose to keep the software working in its current state for as long as the technology would allow it, and focus their resources instead on features that would help the community with their research and publication efforts.

This choice is not without its drawbacks. One key problem is the deprecation of dependencies with time. For example, Ruby 2.0 is no longer available in Debian repositories, and SSL support no longer compiles against it by default. These problems can be seen as indicators that delaying upgrades might not be feasible for much longer. Where possible, deprecated libraries are replaced with newer versions. This is the case for the database, web server, and the Java interpreter, all of which have been replaced with little extra effort. When a new version of a library breaks backwards compatibility, the software is either upgraded or frozen in its current version. Ruby (frozen at `2.0.0-p353` via RVM) and Solr are both examples of the latter, with detailed documented instructions to replicate the software environment.

In addition to the production Anthology site, a second version is kept on low-cost cloud servers for testing purposes. This copy has proven useful for testing step-by-step instructions, since rolling back the server to a clean state requires neither authorization nor downtime. It is also used as a staging area, and to do trial imports of new proceedings and for volunteer training.

Security is another major concern: older dependencies increase exposure to unpatched bugs. The Anthology currently does not collect or store personal data, rendering the consequences of a data breach modest. A compromised server, however, presents not only a risk for the maintainers (service downtime, unauthorized applications) but for the community at large, due to the large number of researchers who could be exposed to malicious

scripts. While the former puts the goodwill of the hosting institution at risk, the latter would affect a large portion of the ACL community.

To tackle issues with outdated software, the Anthology volunteer group is working on making the entire Anthology available via a Docker image (Matthias and Kane, 2015). Docker provides a virtualized environment (also known as a *container*) in which software can be run but where, unlike a virtual machine, the underlying operating system resources can be used directly. Containers are typically stateless, allowing system administrators to add and restart services with minimum friction. Hosting a mirror of the Anthology with Docker containers abstracts away the relatively complex server setup and makes it easier to tackle dependency problems independently from future mirror deployments. As a result, hosting institutions can apply their own internal security policies, and the community can benefit from the added robustness via a larger network of mirrors. Development versions of this image are already available at `https://github.com/acl-org/anthology-docker`. When an instance of this Docker container is started, it first downloads all the data necessary to run the Anthology, inclusive of the metadata and source publications (PDF files) for all proceedings hosted within the Anthology. The resulting Anthology instance is a peer of the production site, but completely independent. This makes it possible for member institutions and even interested individual members to easily provide a mirror or experiment with the data in the Anthology.

Freezing software versions has proven useful to keep stability under control, improve documentation practices, and implement long-requested features like search engine indexing. This does not preclude a full software upgrade from being part of our development roadmap. With better test coverage and expanded consistency checks in place, we expect the first successful upgrade tests to be within our reach in the near future.

Docker containers and temporary servers also show great promise for researchers. An isolated, easy-to-replicate software environment reduces friction in transferring tools between researchers usually caused by incompatible software, simplifies the replication of experiments, and limits the data loss due to software bugs. A container-like approach specifying complete environments can also help in distributing code and general research within the community (e.g., CodaLab[5] as used in SemEval competitions). In the future, best practices within the community may encourage researchers to program and experiment within Docker images to aid reproducibility.

The Anthology is currently stable and supports its current, intended use. However, to ensure that the ACL Anthology continues fulfilling its key roles, we call on the members of the ACL to help with both its operational and development goals:

- hosting mirrors of the Anthology and developing policy for mirror management;
- adding and indexing new publications to the Anthology;
- maintaining and updating the code underlying the Anthology;
- extending the capabilities of the Anthology to help tackle new challenges facing the ACL.

5 Challenges for the Anthology

Maintaining community buy-in for the Anthology is necessary to ensure its future. This is best assured by extending the Anthology with useful capabilities that align with research efforts. This is crucially enabled by the liberal licensing scheme that the ACL employs for the publications to empower end users. Research on the history and structure of the NLP community based on this data has already been undertaken (Anderson et al., 2012; Vogel and Jurafsky, 2012).

Anonymous Pre-prints. A current challenge needing attention is the result of the increasing popularity of pre-prints and their role in promoting scientific progress. However, such pre-print systems are not anonymous, interfering with the well-documented gains that author-blinded publications help in combating bias. Through membership polls and subcommitee study, the ACL executive has adopted a recent set of guidelines upholding the value of double-blinded submissions (ACL Executive Committee, 2017).

One solution would be the use of anonymous pre-prints as an option for authors. Currently two ways of implementing this have been discussed: as a collaboration with an existing pre-print service such as arXiv[6] or through hosting pre-prints

[5] `https://worksheets.codalab.org/`
[6] `https://arxiv.org/`

directly within the Anthology. While the latter option would be a challenge to the Anthology – requiring increased resources both for monitoring the submissions and for scaling the system architecture to a larger and less controlled inflow of papers – but could result in better community control of the process, and a greater awareness and feeling of co-ownership of the Anthology and its data among ACL members.

Reviewer Matching. One key problem with scientific conference and journal organization is in finding suitable reviewers for the peer review process, which is also a key problem for ACL.[7] We believe that we can leverage the ACL Anthology data to support conference organizers in the assignment of potential peer reviewers. There has been a substantial growth in the number of submissions to the main ACL conferences in recent years (Barzilay, 2017), and the ACL has been active in supporting automated approaches to solve the problem (Stent and Ji, 2018) such as the Toronto Paper Matching System (TPMS) (Charlin and Zemel, 2013). However, data for judging the fit between a reviewer and submitted papers are available in the Anthology; i.e., a reviewer's interests and expertise as encoded in their previous publications. Mining and representing such information directly from the Anthology, where data about potential reviewers is already available, makes it unnecessary to upload papers to an external platform, mitigating current low response rates. Measuring overlap between reviewer interests and a submitted paper, based on the reviewer's previous publications, is a problem that the NLP community is ideally suited to solve. Furthermore, the information generated by such a tool could serve conference chairs and journal editors when considering how much weight to assign to a review from specific reviewers. The data required for building such a tool would be both the text and metadata from every submitted paper. While some metadata is already accessible within the Anthology, clean textual content of papers would need to be harvested from the source PDF files, which currently has been partially achieved. (Bird et al., 2008) suggests that the text can generally be extracted using standard tools, with additional processing only necessary for a small fraction of the data. We are aware that clean textual data from the Anthology archives is current on-going interest being investigated by a number of NLP/CL teams within the community.

If such a solution were to be implemented, it would be in the interest of the entire community to have the Anthology maintainers integrate it directly into the Anthology, with support from the original implementers. This has been a problem in the past, where attempts to extend the capabilities of the Anthology with more detailed search and annotation (Schäfer et al., 2011, 2012) were spun off as independent systems to start with and have still not become part of the Anthology service.

We note that these two challenges are synergistically solved. Solving the first challenge will provide the submissions' source text within the Anthology framework and promote better coupling for the second challenge of reviewer matching.

6 Conclusion

The ACL Anthology is a key resource for researchers in the NLP community. We have described the software engineering and maintenance work that goes on behind-the-scenes in order for the Anthology to serve its purpose. This includes ingestion of new papers, maintenance of the Anthology codebase, and the social aspects of recruiting volunteers for this work. The task of training future volunteers and ensuring Anthology uptime is likely to become easier due to improved documentation and simplified server set-up. However, recruitment of new volunteers continues to be an issue.

We invite all community members to download the Anthology images for experimentation, not only for the challenge of automated reviewer assignment, but also for other use cases based on their own research interests. We hope that open challenges and the tasks associated with extending the usefulness of the Anthology will motivate more community members to take interest and become and familiar with its inner workings. We extend an open invitation to anyone interested in the Anthology to get in touch with the members of the team. Our current needs are focused on system administration, software development, database management, and Docker integration, but any kind of experience is welcome.

[7] As intimated through internal discussions with the ACL executive committee.

References

ACL Executive Committee. 2017. Acl policies for submission, review and citation. https://www.aclweb.org/adminwiki/index.php?title=ACL_Policies_for_Submission,_Review_and_Citation. [Online; accessed 05-April-2018].

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a Computational History of the ACL: 19802008. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Regina Barzilay. 2017. Statistics on Submissions and Status Update. https://acl2017.wordpress.com/2017/02/15/statistics-on-submissions-and-status-update/. [Online; accessed 05-April-2018].

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 *Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morocco.

Paul Buitelaar, Georgeta Bordea, and Barry Coughlan. 2014. Hot topics and schisms in NLP: Community and trend analysis with saffron on ACL and LREC proceedings. In *9th Edition of Language Resources and Evaluation Conference (LREC2014)*.

Praveen Bysani and Min-Yen Kan. 2012. Integrating User-Generated Content in the ACL Anthology. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Laurent Charlin and Richard S. Zemel. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.

Karl Matthias and Sean P. Kane. 2015. *Docker: Up & Running*. O'Reilly Media, Inc.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation Journal*.

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*.

Ulrich Schäfer, Jonathon Read, and Stephan Oepen. 2012. The ACL Anthology Searchbench. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Joel Spolsky. 2000. Things You Should Never Do, Part I. https://www.joelonsoftware.com/2000/04/06/things-you-should-never-do-part-i. [Online; accessed 29-March-2018].

Amanda Stent and Heng Ji. 2018. A Review of Reviewer Assignment Methods. https://naacl2018.wordpress.com/2018/01/28/a-review-of-reviewer-assignment-methods. [Online; accessed 29-March-2018].

Adam Vogel and Dan Jurafsky. 2012. He Said, She Said: Gender in the ACL Anthology. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.