



# The ACL Anthology

Current State and Future Directions



Daniel Gildea, **Min-Yen Kan**, Nitin Madnani,  
Christoph Teichmann, Martin Villalba

# What is this presentation **about**?

**The ACL Anthology**  
Current state and future directions

**Team:**  
Daniel Gildea (University of Rochester), Min-Yen Kan (National University of Singapore), Nitin Madnani (Educational Testing Service), Christoph Teichmann (Saarland University), Martin Villalba (Saarland University).

The ACL Anthology 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. It offers search of indexed papers, author-specific pages, and services that can be embedded within pages. It is Open Source, and maintained on a volunteer basis.

**A brief history of the Anthology**  
2001: The Anthology is proposed as a project to the ACL Executive by Steven Bird.  
2002: First version of the Anthology, with Steven Bird taking the position of Editor.  
2008: Min-Yen Kan takes the role of Editor.  
2012: A second version of the Anthology is released. Maintenance continues to this day.  
2017: After several years hosted in the University of Singapore, the Anthology relocates to Saarland University.

**Current state of the Anthology**  
43k hosted papers  
4.5k daily hits

The Anthology is a community project run by volunteers, coordinated by the Editor. Hosting has been typically provided by Universities free of charge and all code is publicly available on GitHub: <https://github.com/acl-org/acl-anthology>

**Future proofing the Anthology**  
With the Anthology in a stable state, it is time to plan ahead. We identified three main areas of work:  
- Documenting every aspect of the system, to simplify onboarding of new members when project members change.  
- Update/replace all outdated dependencies.  
- Reinforce our development process to quickly detect bugs and/or data inconsistencies.

**Challenges for the Community**  
We want the Anthology to grow beyond a repository of scientific papers. We invite the community to contribute ideas and implementations on all areas. We suggest two possible first projects:  
- Add anonymous pre-print support, helping authors and preserving double-blind review - either as collaboration with pre-print services like ArXiv, or as an extension  
- Use the paper database to find suitable reviewers for submissions in future conferences

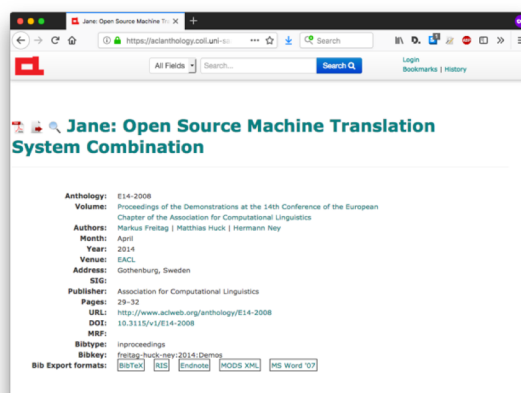
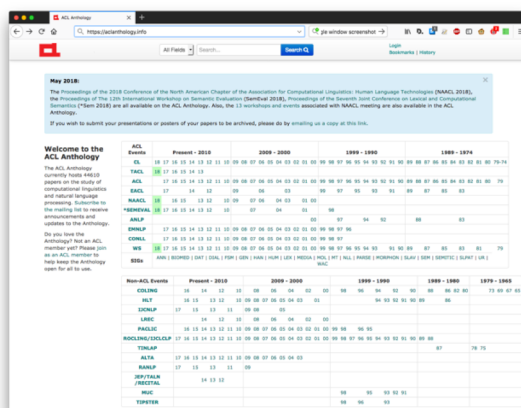
**Join the Anthology team!**  
We are always looking for new volunteers to keep the Anthology running. Programmers (especially Ruby on Rails), system administrators, software engineers... every little bit helps!  
Interested? Know someone? Get in touch with us! Contact any of the authors, or follow us on GitHub!

**Further reading:**  
Anderson, McFarland and Jorahay. Towards a Computational History of the ACL. In: ACL 2016. Proceedings of the ACL Special Workshop 2016 at the Association for Computational Linguistics.  
Bird et al. The ACL Anthology Reference Corpus: A Reference Dataset for Research in Computational Linguistics, Language Resources and Evaluation Conference LREC 2008.

**Stack:**  
NOSQL, Web Server (NGINX), Database server (PostgreSQL), Framework (Ruby on Rails), Search engine (Solr).

- Summarize the history and current state of efforts related to the Anthology
- Illustrate the challenges of maintaining a community Project
- Invite the community to extend the capabilities of the Anthology
- Call you to join the Anthology team

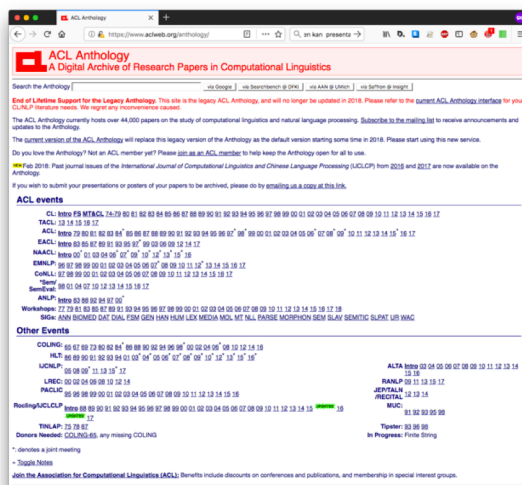
# The Anthology in **summary**



- Open access service for all ACL-Sponsored publications
- Also hosts posters and additional data
- Paper search and author pages
- 45K papers and 4.5K daily hits
- Open Source
- Maintained by volunteers
- New papers added in collaboration with proceedings editors

# A brief **History** of the Anthology

- Proposed in 2001 by Steven Bird
- First version online in 2002, with Steven Bird as editor
- Min-Yen Kan becomes the new editor in 2008
- A new version of the Anthology with extra functionality is released in 2012
- Hosting of the Anthology moves from the National University of Singapore to Saarland University



Steven Bird



Min-Yen Kan



# How to **Future-proof** the Anthology









## **Challenges**

- Limited resources for day-to-day code maintenance
- Dependencies become outdated
- Maintainer churn

## **Solutions**

- Docker container for easier set-up and sandboxing
- Collaborative documentation efforts to ease onboarding
- Migration plan on the pipeline, including upgrades and test cases

# Upcoming major steps

Backlog	 Upgrade and/or migrate outdated dependencies	 Full text search over uploaded papers
In research	 Full test coverage and consistency checks	
In progress	 Docker image for releases	 Add a staging server
Done	 Add index support for popular search engines	 Document and update the installation process
	 Add a test server	

- Hosting the Anthology within the main ACL website
- Recruit a new Anthology editor
- (possibly) pay for extra support for the Anthology

# Exercise: Importing of your slides

The screenshot shows the ACL Anthology website. A blue box at the top left contains the text: "Jun 2018: The Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, the Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), Proceedings of the Seventh Joint Conference on Lexical Semantics (\*Sem 2018) are all available on the ACL Anthology. Also, the 13 workshops and events associated with NAACL meeting are also available in the Anthology. Finally, the Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation from 2017 are also now available in the Anthology. If you wish to submit your presentations or posters of your papers to be archived, please do by emailing us a copy at this link." A blue line points from this text to a larger blue box at the bottom left. This box contains the text: "If you wish to submit your presentations or posters of your papers to be archived, please do by emailing us a copy at this link." The website interface includes a search bar at the top with "All Fields" and "Search Q" buttons, and a "Login" link. Below the search bar is a table of ACL Events with columns for "Present - 2010", "2009 - 2000", and "1999 - 1990". The table lists various events like CL, TACL, ACL, EACL, NAACL, \*SEM Eval, ANLP, EMNLP, CONLL, WIS, and SIGs, with corresponding years in the cells. A "Non-ACL Events" section is also visible at the bottom of the table.

Jun 2018: The Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, the Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval 2018), Proceedings of the Seventh Joint Conference on Lexical Semantics (\*Sem 2018) are all available on the ACL Anthology. Also, the 13 workshops and events associated with NAACL meeting are also available in the Anthology. Finally, the Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation from 2017 are also now available in the Anthology. If you wish to submit your presentations or posters of your papers to be archived, please do by emailing us a copy at this link.

Welcome to the ACL Anthology

The ACL Anthology currently hosts 45427 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to receive announcements and updates to the Anthology.

Do you love the Anthology? Not an ACL member yet? Please join as an ACL member to help keep the Anthology open for all to use.

ACL Events	Present - 2010	2009 - 2000	1999 - 1990
CL	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
TACL	18 17 16 15 14 13		
ACL	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
EACL	17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
NAACL	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
*SEM Eval	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
ANLP	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
EMNLP	17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
CONLL	17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
WIS	18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90 89 88	
SIGs	ANN   BIOMED   DAT   DIAL   FSM   GEN   HAN   HUM   LEX   MEDIA   MOL   MT   NLL   PARSE   MORPHON   SLAV   WAC		

Non-ACL Events

Present - 2010

2009 - 2000

1999 - 1990

- We import slides, datasets, videos from your own
- Currently done by email (try it yourself! yes, now)
- Better workflow: pull request against the Anthology XML (à la [csrankings.org](https://csrankings.org))

# Possible **future directions**

- Contains useful information both *for* CL researchers and *about* CL researchers. Useful for identifying suitable reviewers.
- Move focus from day-to-day operations towards development
- Establish a network of mirrors
- Host anonymized pre-prints





# 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  
kanny@comp.nus.edu.sg

**Nitin Madnani**  
Educational Testing Service  
Princeton NJ  
nmadnan@ets.org

**Christoph Teichmann**  
Dept. of Language, Science, and Technology  
Saarland University  
[ctreichmann|villalba]@coli.uni-saarland.de

**Martin Villalba**

The ACL Anthology 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. It offers search of indexed papers, author-specific pages, and services that can be embedded within pages. It is Open Source, and maintained on a volunteer basis.

### A brief history of the Anthology

- 2001 The Anthology is proposed as a project to the ACL Executive by Steven Bird.
- 2002 First version of the Anthology, with Steven Bird taking the position of Editor.
- 2008 Min-Yen Kan takes the role of Editor.
- 2012 A second version of the Anthology is released. Maintenance continues to this day.
- 2017 After several years hosted in the University of Singapore, the Anthology relocates to Saarland University.

### Current state of the Anthology



**43k**  
hosted papers

**4.5k**  
daily hits

The Anthology is a community project run by volunteers, coordinated by the Editor. Hosting has been typically provided by Universities free of charge and all code is publicly available on GitHub:

<https://github.com/acl-org/acl-anthology>



### Future proofing the Anthology

With the Anthology in a stable state, it is time to plan ahead. We identified three main areas of work:

- Documenting every aspect of the system, to simplify onboarding of new members when project members change.
- Update/replace all outdated dependencies.
- Reinforce our development process to quickly detect bugs and/or data inconsistencies.



### Challenges for the Community

We want the Anthology to grow beyond a repository of scientific papers. We invite the community to contribute ideas and implementations on all areas. We suggest two possible first projects:

- Add anonymous pre-print support, helping authors and preserving double-blind review - either as collaboration with pre-print services like ArXiv, or as an extension
- Use the paper database to find suitable reviewers for submissions in future conferences

You can also download our Docker image, including all metadata on ACL papers, and start hacking right away. For references on other cool projects involving the Anthology, make sure to check the **Further Reading** section!

### Join the Anthology team!

We are always looking for new volunteers to keep the Anthology running. Programmers (especially Ruby on Rails!), system administrators, software engineers... every little bit helps!

Interested? Know someone? Get in touch with us! Contact any of the authors, or follow us on GitHub!

### Further reading:

Anderson, McFarland and Jurafsky: Towards a Computational History of the ACL: 1988-2008. Proceedings of the ACL Special Workshop 2012 on Re-examining 50 years of Discoveries.

Bird et al. The ACL Anthology Reference Corpus: A Reference Dataset for Biographic Research in Computational Linguistics, Language Resources and Evaluation Conference (LREC) 2008.



- Comments? Questions?
- Ideas for future directions?
- Interested in joining the Anthology team?

# Come and visit our poster