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

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
Dept. of Language, Science, and Technology
Saarland University
{cteichmann|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**.

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- Update/replace all outdated dependencies.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.

Backlog | In research | In progress | Done

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

Web Server NGINX — Database server PostgreSQL — Framework Ruby on Rails — Search engine Solr

Operating System Debian Stable

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!

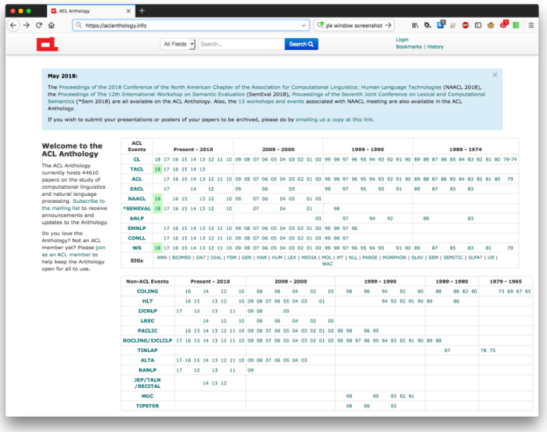
Further reading:

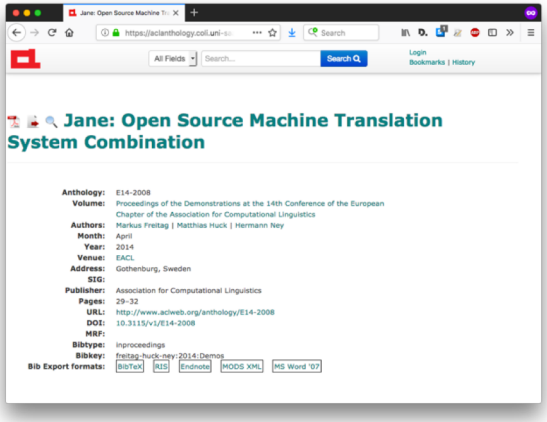
Anderson, McFarland and Jurafsky: *Towards a Computational History of the ACL: 1998-2008.* Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries.

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

- 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

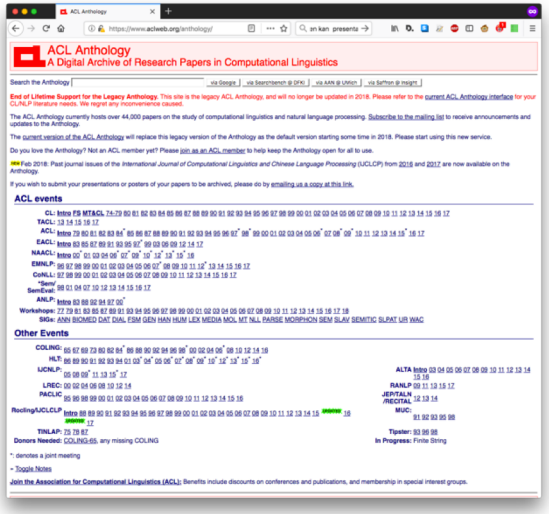
The Anthology in **summary**





- 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



Steven Bird

Min-Yen Kan

- 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

How to Future-proof the Anthology

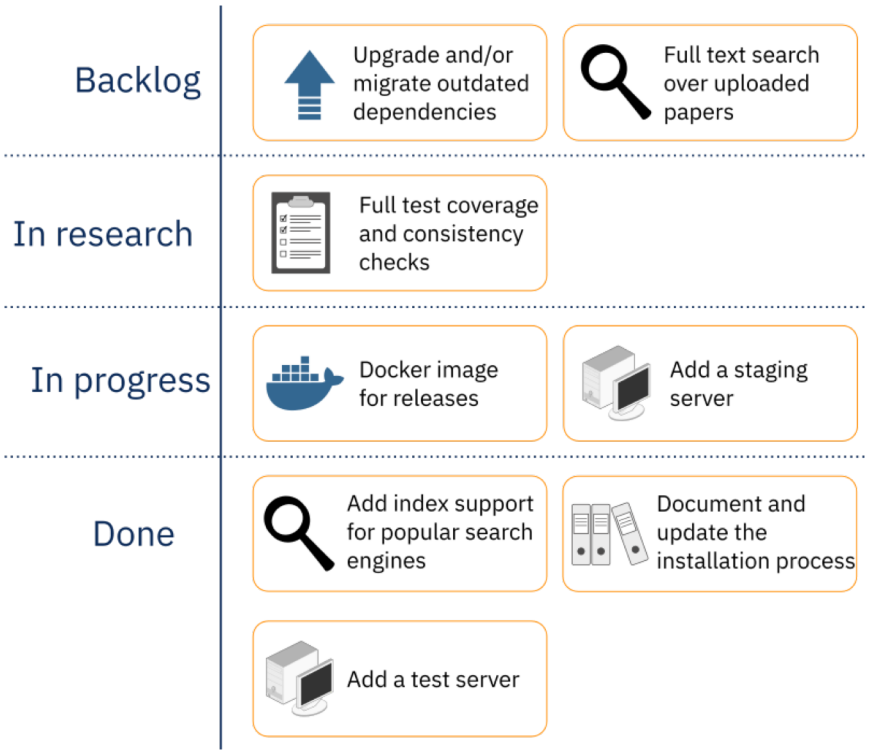
Challenges

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- Dependencies become outdated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

Solutions

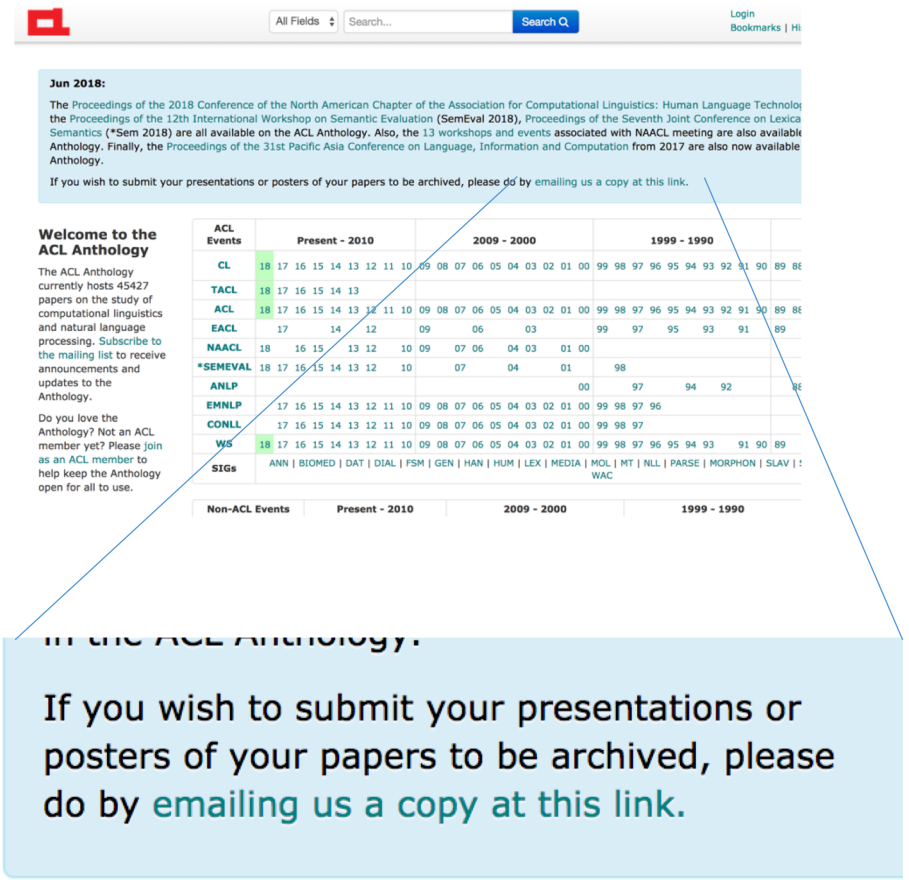
- Docker container for easier set-up and sandboxing
- Collaborative documentation efforts to ease onboarding
- Migration plan on the pipeline, including upgrades and test cases

Upcoming major steps



- 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



- 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)

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
kanmy@comp.nus.edu.sg

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

Christoph Teichmann
Dept. of Language, Science, and Technology
Saarland University
{cteichmann|villalba}@coli.uni-saarland.de

Martín 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

Web Server: NGINX
Operating System: Debian Stable
Database server: PostgreSQL
Framework: Ruby on Rails
Search engine: Solr

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.

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

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: 1998-2008.* Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries.
Bird et. al: *The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic 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**