Integrating User-Generated Content in the ACL Anthology

Praveen Bysani
Min-Yen Kan (ACL Anthology Editor)
Initiative started in 2001
Mission: to preserve ACL’s scientific legacy
Open-access repository
Version 1 Architecture

ACL PUB

Proceedings tgz → Shell Scripts → XML files → Shell Scripts → Static HTML files

DBLP HTML

CSV + ZIP

ACM
Anthology XML format

• Central data structure for ACL publications
  – Describe metadata for each publication
  – Unique identifier (ACL Anthology ID)
  – The canonical record

• But:
  – No DTD, often doesn’t validate as XML
  – Hard to inventory or ask useful questions of the data

→ Probably has outlived its usefulness
Anthology Version 2

Have already met the earlier mission goals, so it’s time to …

Revise our mission

– Use a proper data model
  Allows more first class objects (venues, authors as well as papers)
– Use contemporary technology
  Ruby on Rails (MVC architecture for data-backed web application)
  Faceted Search (via Blacklight and Solr)
– Incorporate user content
Enables join views (e.g., SIGs to venues, Authors to co-authors)
Version 2 (Beta) Architecture

XML files → Rake Tasks → Ruby on Rails → Apache Solr

Solr Indices

Blacklight → HTML / AJAX

Users
Usability – Front End Changes

- Faceted Search and Discovery
- Author, Venue and Publication pages
- Post-publication revision by authors
- Incorporate readership feedback
- Integrate programmatic contributions

Authors

Readers

Programmatic Agents
Welcome to the ACL Anthology

ACL Anthology currently hosts over 19,200 papers on the study of computational linguistics. Subscribe to the mailing list to receive announcements and updates to the Anthology.

Recent News:
- The June issue of the Computational Linguistics Journal, is now available.
- Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, is now available

Popular Papers this week
- Measuring Text Reuse
- Proceedings of the Third International Workshop on Paraphrasing (IWP2005)
- Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts
- Sentence Boundary Detection and the Problem with the U.S.
- Instance Based Lexical Entailment for Ontology Population
- Automatic Verb Classification Based on Statistical Distributions of Argument Structure

Popular Authors this week
- Olga Batiukova
- Sheng Li
- Akihiro Tamura
- R. Brodersen
- Satoshi Sekine
- Lance Ramshaw
- Anil Kumar Singh

ACL Events
- Journal: Intro FS MT&CL 74-79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12
- ACL: Intro 79 80 81 82 83 84* 85 86 87 88 89 90 91 92 93 94 95 96 97* 98* 99 00 01 02 03 04 05 06 07 08 09 10 11
- EACL: Intro 83 85 87 89 91 93 95 97* 99 03 06 09
- NAACL: Intro 00* 01 03 04 06* 07* 09* 10*
- SemEval: 98 01 04 07 10
- ANLP: Intro 83 89 92 94 97 00*
- EMNLP: 01 02 03 04 05 06 07* 08 09 10 11
- Workshops: 90 91 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10
- SIGs: ANNI BIOMED DAT DIAL FSM GEN HAN LEX MEDIA MOL MT NLL PARSE MORPHON SEM SEMITIC WAC

Other Events
- Faceted Search / Browse
- Use Statistics
- Auxiliary Content
Faceted Search and Browsing

Facets are properties of first class objects: (e.g., author, year, venue for papers)

Allows complex user queries, drill-down, and transparent cross-over between searching and browsing.
Faceted Search and Browsing

- Filters for many venues, auxiliary content
- Extensible to add other filters (e.g., using a dataset, editorship, target processing language)
Readership: Use Statistics

Popular Papers this week

- Measuring Text Reuse
- Proceedings of the Third International Workshop on Paraphrasing (IWP2005)
- Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts
- Sentence Boundary Detection and the Problem with the U.S.
- Instance Based Lexical Entailment for Ontology Population
- Automatic Verb Classification Based on Statistical Distributions of Argument Structure

Popular Authors this week

- Olga Batiukova
- Sheng Li
- Akihiro Tamura
- R. Brodersen
- Satoshi Sekine
- Lance Ramshaw
- Anil Kumar Singh

Incorporate readership usage about downloads and page views
Author (and Venue) pages

Dan Gillick

Popular Co-Authors

Dan Klein (2)
Benoit Favre (1)
James Zhang (1)
John DeNero (1)
Taylor Berg-Kirkpatrick (1)
Yang Liu (1)

Venue

others (2)
acl (1)
aaacl (1)
wmt (1)

Publications

2011

- Jointly Learning to Extract and Compress

2010

- Non-Expert Evaluation of Summarization Systems is Risky

2009

- Sentence Boundary Detection and the Problem with the U.S.
- A Scalable Global Model for Summarization

2006

- Why Generative Phrase Models Underperform Surface Heuristics

With popular co-authors and publication venues
Bibliographic Metadata Export
(supports more formats via MODS Bibliographic metadata exchange format)

Social Media Integration

User Bibliographic Data Correction and Contribution

Social Commenting

Programmatic Contributed Data (AAN webpage shown as an example)

Per-Publication View

ACL Anthology Network

Anthology ID: P11-1100
Author(s): Ziheng Lin; Hwee Tou Ng; Min-Yen
Year: 2011
Event: ACL

Bib Export Formats: ☐ Bib ☐ Ris ☐ EndNote
Erratum: None

Edit this paper's metadata

Add New Comment

Contributed Information

aal_page
The public is invited to:

- Report errors in the metadata
- Supply revisions and errata, software dataset links post-publication
- Discuss the papers using the commenting framework
Programmatic contributions

*We are most excited about the opportunities here!*

- Apply NLP to our own works
- Allow automated agents to publish supplementary material for a paper

- Agents provide information in an XML format

- Currently support per-paper contributions such as **text**, **hyperlinks** and **embedded webpages**

```xml
<paper id="P11-1100">
  <content name="keywords" type="text">
    <item>
      discourse, implicit reference, text coherence, readability
    </item>
  </content>
</paper>
```
Impact and usage of V 2 interface

- Analyzed application and search logs over 5 days
- Received ~16k page views, compared to ~7k views on original website

Notable ratio (16%) of longer visits support that new features encouraged more user engagement
• **Majority of requests (68%) use faceting feature**
  • Supports claim that faceted browsing is preferable to search choice in casual contexts

• **Average response time is 0.73 seconds, average load time 5.6 seconds**
  • Need better scaling on database, current work to migrate to cloud architecture (Amazon EC2)
Conclusion and outlook

Collaborate with the community to incorporate programmatic contributions into the Anthology
• Especially R50 contributions!

Taking back search from Google’s custom search also means that our search logs can be provided to our own community for research

Tighter integration with other ACL resources such as the ACL Anthology Network and the ACL Searchbench