

# Integrating User-Generated Content in the ACL Anthology

Praveen Bysani

**Min-Yen Kan** (*ACL Anthology Editor*)



Search the Anthology  via Google | via Searchbench @ DFKI | via AAN @ UMich

The ACL Anthology currently hosts over 21,200 papers on the study of computational linguistics and natural language processing. [Subscribe to the mailing list](#) to receive announcements and updates to the Anthology.

**NEW** Jul 2012: The [Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics](#) and the [12 associated workshops](#) are now available in the Anthology. The [Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue \(SIGDIAL 2012\)](#), are also now available.

**NEW** The beta version of the new ACL Anthology goes live. Try it out and give us your feedback!

## 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](#) **NEW** [12](#)

**EACL:** [Intro](#) [83](#) [85](#) [87](#) [89](#) [91](#) [93](#) [95](#) [97](#) [99](#) [03](#) [06](#) [09](#) [12](#)

**NAACL:** [Intro](#) [00](#) [01](#) [03](#) [04](#) [06](#) [07](#) [09](#) [10](#) [12](#)

**\*Sem/  
SemEval:** [98](#) [01](#) [04](#) [07](#) [10](#) [12](#)

**ANLP:** [Intro](#) [83](#) [88](#) [92](#) [94](#) [97](#) [00](#)

**EMNLP:** [96](#) [97](#) [98](#) [99](#) [00](#) [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](#) [11](#) **NEW** [12](#)

**SIGs:** **NEW** [ANN](#) [BIOMED](#) [DAT](#) **NEW** [DIAL](#) [FSM](#) [GEN](#) [HAN](#) **NEW** [LEX](#) [MEDIA](#) [MOL](#) **NEW** [MT](#) [NLL](#) **NEW** [PARSE](#) [MORPHON](#) **NEW** [SEM](#) [SEMITIC](#) **NEW** [SLPAT](#) [WAC](#)

## Other Events

**COLING:** [65](#) [67](#) [69](#) [73](#) [80](#) [82](#) [84](#) [86](#) [88](#) [90](#) [92](#) [94](#) [96](#) [98](#) [00](#) [02](#) [04](#) [06](#) [08](#) [10](#)

**HLT:** [86](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [01](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [12](#)

**IJCNLP:** [05](#) [08](#) [09](#) [11](#)

**LREC:** [00](#) [02](#) [04](#) [06](#) [08](#) [10](#)

**PACLIC** [95](#) [96](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

**Rocling** [Intro](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

**TINLAP:** [75](#) [78](#) [87](#)

**Donors Needed:** [COLING-65](#), any missing COLING

**ALTA** [Intro](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

**RANLP** [09](#) [11](#)

**MUC:** [91](#) [92](#) [93](#) [95](#) [98](#)

**Tipster:** [93](#) [96](#) [98](#)

**In Progress:** Finite String

\*: denotes a joint meeting

» [Toggle Notes](#)

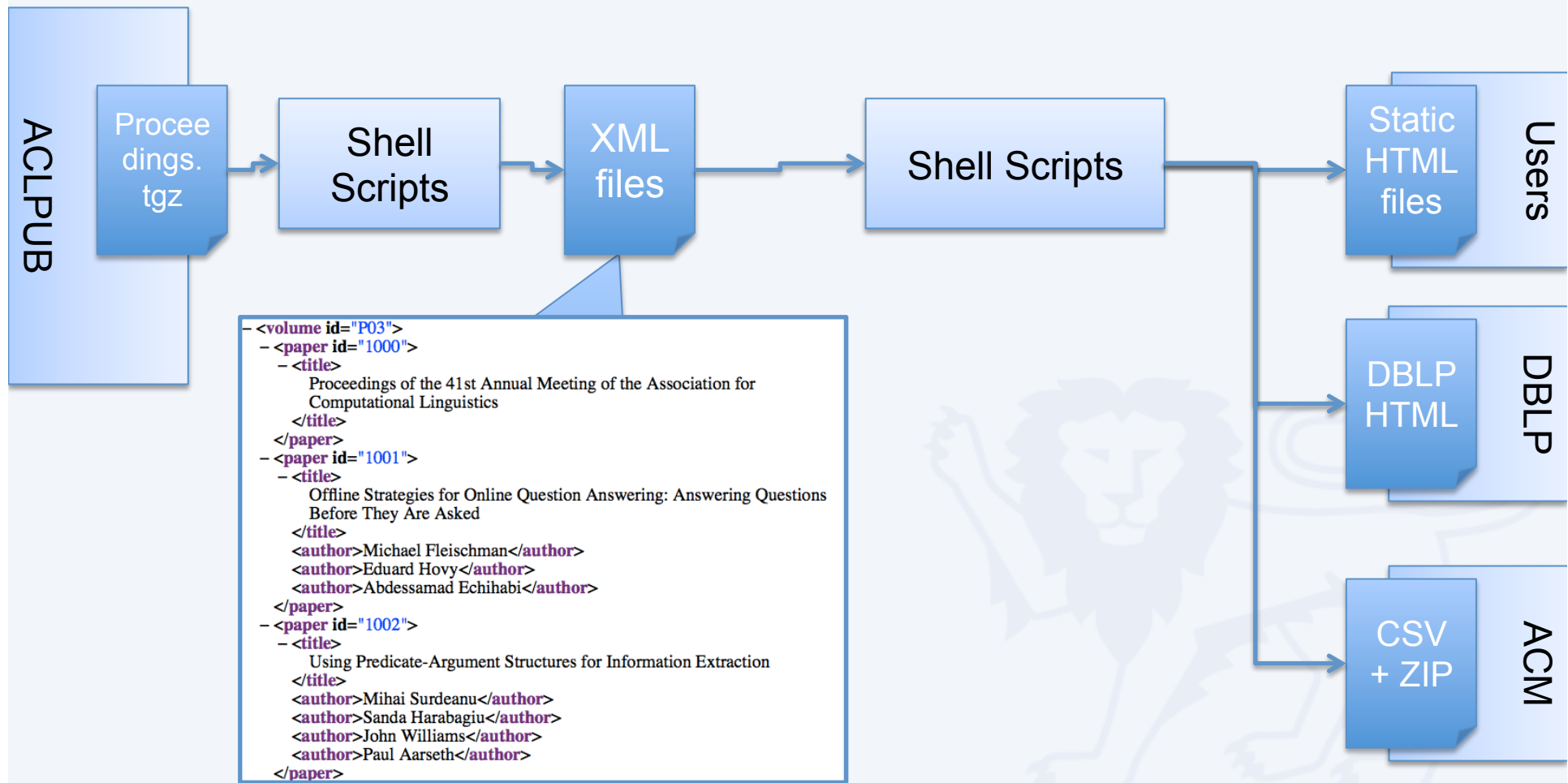
**Join the Association for Computational Linguistics (ACL):** Benefits include discounts on conferences

» [Toggle Copyright and Credits](#)

[Min-Yen Kan](#) (Editor, 2008-) / [Steven Bird](#) (Editor, 2001-2007)

- Initiative started in 2001
- Mission: to preserve ACL's scientific legacy
- Open-access repository

# Version 1 Architecture



## 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

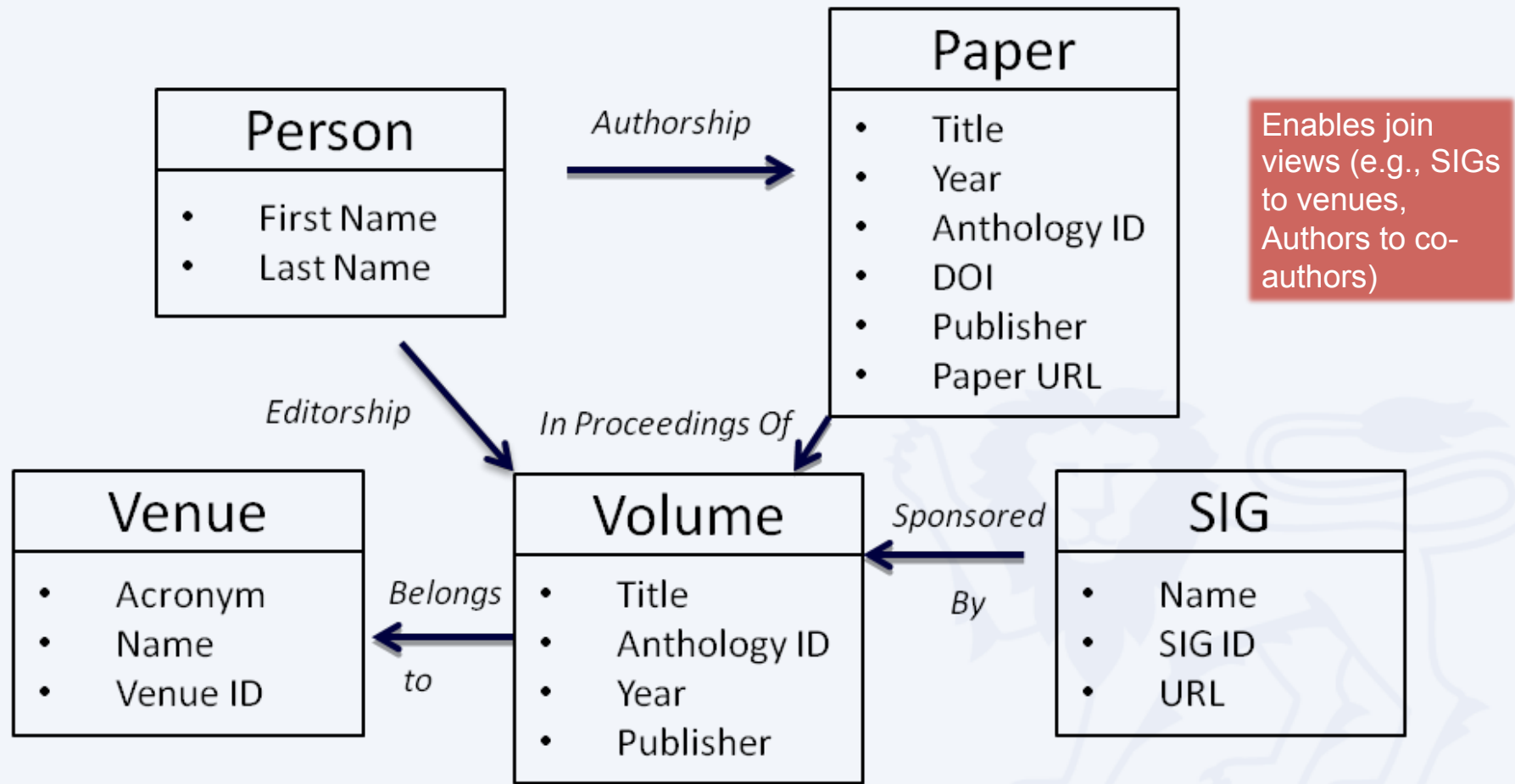
*Have already met the earlier mission goals, so it's time to ...*

## Revise our mission

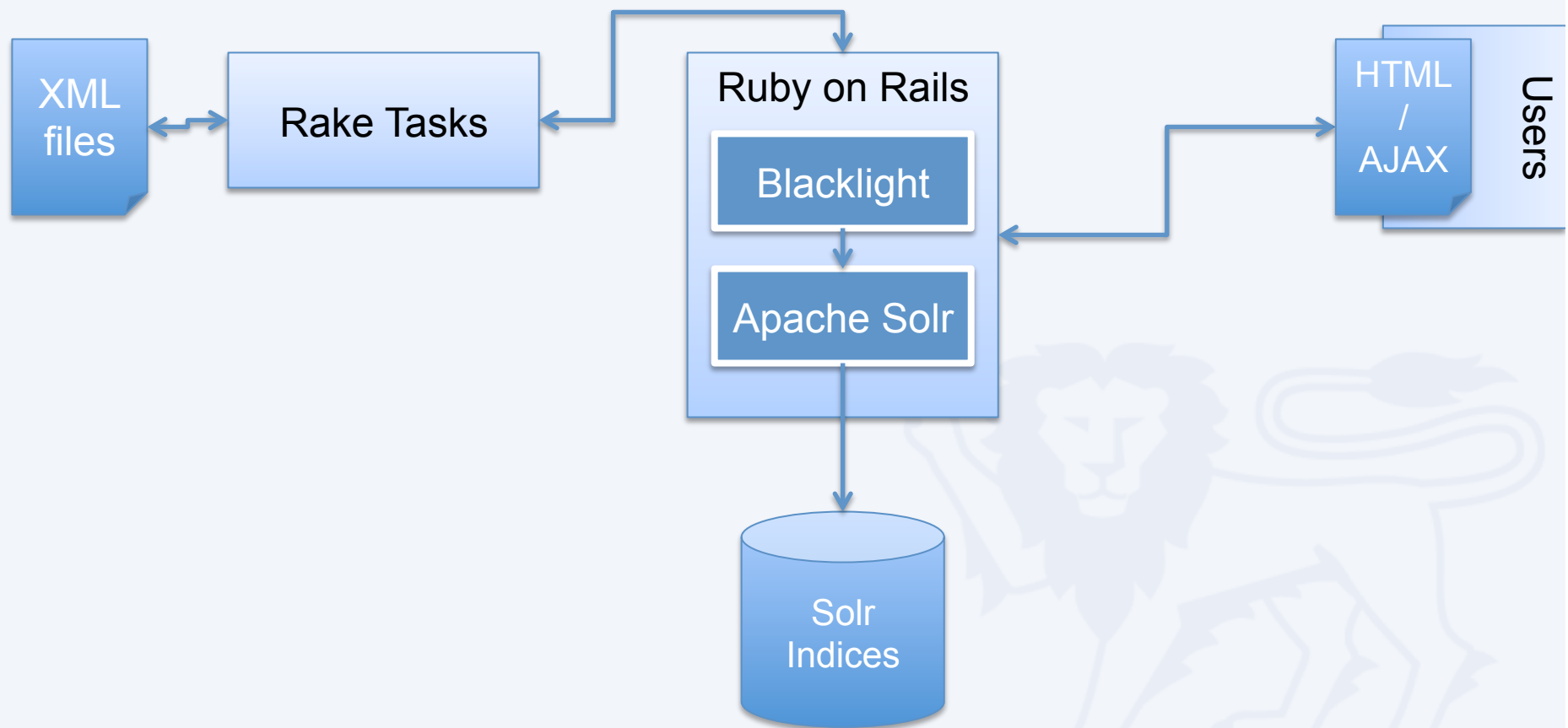
- Use a proper data model
  - Allows more first class objects (venues, authors as well as papers)
- Use contemporary technology
  - Ruby on Rails (MVC architecture for data-backed web application)
  - Faceted Search (via Blacklight and Solr)
- Incorporate user content



# Version 2 Data Model



## Version 2 (Beta) Architecture



## Usability – Front End Changes

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







## ::Faceted Browsing::

Venue

Publication Year

-  [Limit](#)

Current results range from 1965 to 2012

[View distribution](#)

Authors

SIG

- [sigdat \(555\)](#)
- [signll \(519\)](#)
- [siggen \(402\)](#)
- [sigdial \(383\)](#)
- [sighan \(284\)](#)
- [siglex \(269\)](#)
- [sigmt \(220\)](#)
- [sigsem \(191\)](#)
- [sigparse \(106\)](#)
- [sigmorphon \(96\)](#)
- [semitic \(83\)](#)
- [sigann \(68\)](#)
- [sigbiomed \(47\)](#)
- [sigwac \(17\)](#)

Attachments

- [Attachment \(23\)](#)
- [Dataset \(21\)](#)
- [Software \(13\)](#)
- [none \(20,375\)](#)

in [All Fields](#) [Search](#)

[Advanced search](#)

## 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](#)
- [Named Entity Recognition: A Maximum Entropy Approach Using Global Information](#)
- [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](#) [NEW](#) [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](#) [88](#) [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](#)
- SIGs:** [ANN](#) [BIOMED](#) [DAT](#) [DIAL](#) [FSM](#) [GEN](#) [HAN](#) [LEX](#) [MEDIA](#) [MOL](#) [MT](#) [NLL](#) [PARSE](#) [MORPHON](#) [SEM](#) [SEMITIC](#) [WAC](#)

- Faceted Search / Browse
- Use Statistics
- Auxiliary Content

### Other Events

## ::Faceted Browsing::

Venue

acl (349)[remove]

Publication Year

2011 to 2012 (349) [remove]

2011 - 2012

Current results range from 2011 to 2011

2011 to 2011 (349)

2012 to 2012 (0)

Authors

Dan Klein (6)

Yang Liu (6)

Brian Roark (5)

Noah A. Smith (5)

Dan Roth (4)

David Chiang (4)

Eduard Hovy (4)

Hwee Tou Ng (4)

Rada Mihalcea (4)

Regina Barzilay (4)

Timothy Baldwin (4)

Amjad Abu-Jbara (3)

Aria Haghighi (3)

Chris Callison-Burch (3)

Daniel Gildea (3)

Dipanjan Das (3)

Dragomir Radev (3)

Eiichiro Sumita (3)

Francisco Casacuberta (3)

Giorgio Satta (3)

[more »](#)

in All Fields Search

Advanced search

Venue > acl x

Publication Year > 2011 to 2012 x

# Faceted Search and Browsing

Sort by title

Show 50 per page

< Previous

1 2 **3** 4 5 6 7

Next >

### 101. Simple supervised document geolocation with geodesic grids

Select

[P11-1096]: Benjamin Wing ; Jason Baldrige

### 102. Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition

Select

[P11-1097]: Stefan Rüd ; Massimiliano Ciaramita ; Jens Müller ; Hinrich Schütze

### 103. Template-Based Information Extraction without the Templates

Select

[P11-1098]: Nathanael Chambers ; Dan Jurafsky

### 104. Classifying arguments by scheme

Select

[P11-1099]: Vanessa Wei Feng ; Graeme Hirst

### 105. Automatically Evaluating Text Coherence Using Discourse Relations

Select

[P11-1100]: Ziheng Lin ; Hwee Tou Ng ; Min-Yen Kan

### 106. Underspecifying and Predicting Voice for Surface Realisation Ranking

Select

[P11-1101]: Sina Zarrieß ; Aoife Cahill ; Jonas Kuhn

### 107. Recognizing Authority in Dialogue with an Model

[P11-1102]: Elijah Mayfield ; Carolyn Penstein R

### 108. Reordering Metrics for MT

[P11-1103]: Alexandra Birch ; Miles Osborne

### 109. Reordering with Source Language Collocat

[P11-1104]: Zhanyi Liu ; Haifeng Wang ; Hua Wu

### 110. A Joint Sequence Translation Model with I

[P11-1105]: Nadir Durrani ; Helmut Schmid ; Ale

### 111. Integrating surprisal and uncertain-input n formal techniques and empirical results

[P11-1106]: Roger Levy

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

## ::Faceted Browsing::

- Venue
- Publication Year
- Authors
- SIG
- Attachments

- Attachments 
  - Attachment (23)
  - Dataset (21)
  - Software (13)
  - none (20,375)

wmt (272)

Others (4,230)

Publication Year

-

Current results range from 2012

[View distribution](#)

Authors

- [Yuji Matsumoto \(106\)](#)
- [Eduard Hovy \(93\)](#)
- [Ralph Grishman \(89\)](#)
- [Timothy Baldwin \(89\)](#)
- [Chu-Ren Huang \(86\)](#)

[more »](#)

**Venue**

« Previous **Next »** **A-Z Sort** Numerical Sort

- Others (4,230)
- [ACL \(3,336\)](#)
- [COLING \(2,667\)](#)
- [CL \(1,061\)](#)
- [NAACL \(1,044\)](#)
- [EACL \(812\)](#)
- [HLT \(755\)](#)
- [ROCLING \(724\)](#)
- [IJCNLP \(667\)](#)
- [LREC \(621\)](#)
- [EMNLP \(572\)](#)
- [PACLIC \(494\)](#)
- [conll \(387\)](#)
- [ANLP \(344\)](#)
- [SEMEVAL \(323\)](#)
- wmt (272)
- [RANLP \(246\)](#)
- [sigdial \(188\)](#)
- [ALTA \(164\)](#)
- [MUC \(160\)](#)

« Previous **Next »** **A-Z Sort** Numerical Sort

- 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
- Named Entity Recognition: A Maximum Entropy Approach Using Global Information
- 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

## 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\(1\)](#)  
[naacl\(1\)](#)  
[wmt\(1\)](#)


## [Start Over](#)

## Dan Gillick


### Publications

With popular co-authors and publication venues

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](#)



Social Media Integration

Bibliographic Metadata Export  
(supports more formats via MODS Bibliographic metadata exchange format)

Anthology ID: P11-1100  
Author(s): Ziheng Lin ; Hwee Tou Ng ; Min-Yen  
Year: 2011  
Event: ACL  
Volume: Proceedings of the 49th Annual Meet  
Linguistics: Human Language Techno

Bib Export Formats:  Bib  Ris  EndNote  
Erratum: None

# Per-Publication View

User Bibliographic Data Correction and Contribution  
Social Commenting

**Edit this paper's metadata** +

Like

**Add New Comment**

Type your comment here.

**Contributed Information** +

- aan\_page

ACL Anthology News

Programmatic Contributed Data (AAN webpage shown as an example)

# User Contributed Data: Post publication revision, addition

Anthology ID: C10-1021

Author(s): Ying Chen ; Sophia Yat Mei Lee ; Shoushan Li ; Chu-Ren Huang

Year: 2010

Event: Non ACL

Bib Export Formats:  **Bib**  **Ris**  **EndNote**  **Word**

Erratum: None

## Edit this paper's metadata

Please provide your contact details and fill in the details of the necessary changes below. To prevent unauthorized changes, the editors review such requests manually, and will get back to you if the change is approved and reflected in the Anthology.

Your Name:

Your Email:

-----

Anthology ID:

Author(s):

Year:

Event:

Volume:

DOI:

Attachment:

Bib Export Formats:

Dataset\*:

Software\*:

Erratum\*:

Revised Version\*:

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!*

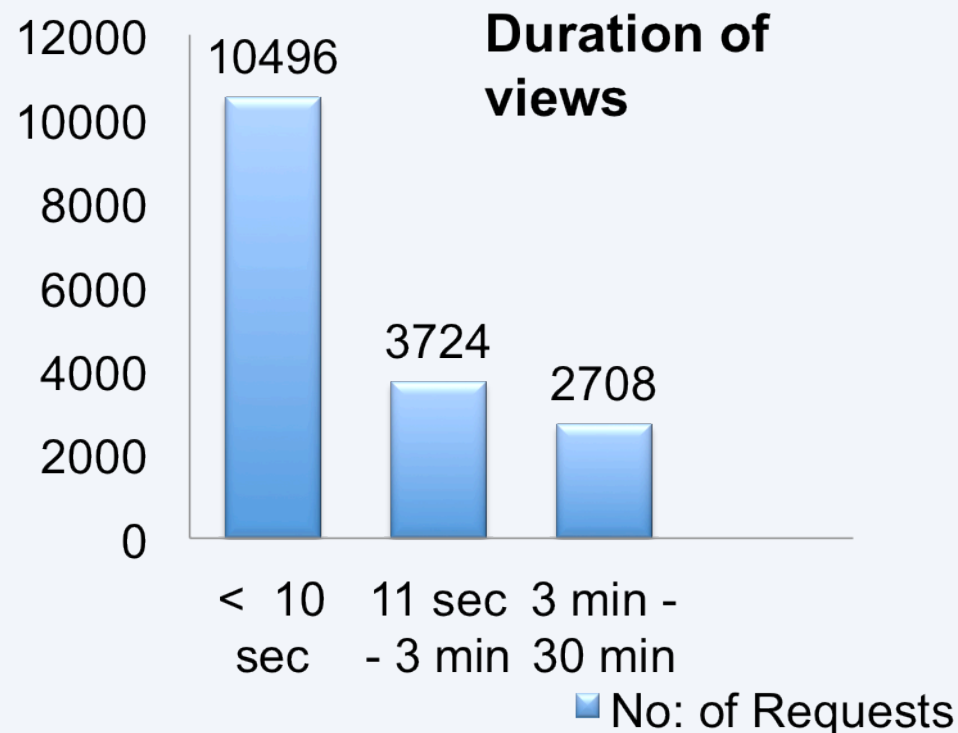
- Apply NLP to our own works
- Allow automated agents to publish supplementary material for a paper
- Agents provide information in an XML format
- Currently support per-paper contributions such as *text*, *hyperlinks* and *embedded webpages*

```
<paper id="P11-1100">  
  -<content name="keywords" type="text">  
    -<item>  
      discourse, implicit reference, text coherence, readability  
    </item>  
  </content>  
</paper>
```

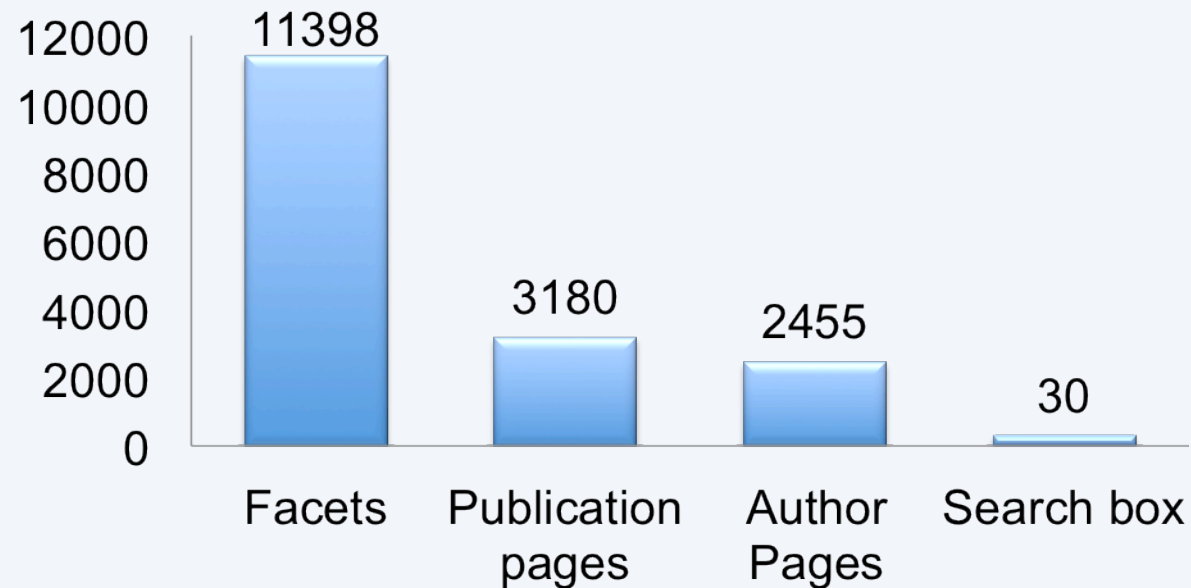


## Impact and usage of V 2 interface

- 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



■ No: of Requests

- **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**