Integrating User-Generated Content in the ACL Anthology

Praveen Bysani
Web IR / NLP Group (WING)
National University of Singapore
13 Computing Link, Singapore 117590
bpraveen@comp.nus.edu.sg

Min-Yen Kan
Web IR / NLP Group (WING)
National University of Singapore
13 Computing Link, Singapore 117590
kanmy@comp.nus.edu.sg

Abstract

The ACL Anthology was revamped in 2012 to its second major version, encompassing faceted navigation, social media use, as well as author- and reader-generated content and comments on published work as part of the revised frontend user interface. At the backend, the Anthology was updated to incorporate its publication records into a database. We describe the ACL Anthology's previous legacy, redesign and revamp process and technologies, and its resulting functionality.

1 Introduction

To most of its users, the ACL Anthology[1] is a useful open-access repository of scholarly articles on the topics of computational linguistics and natural language processing. The liberal use and access policy granted by the Association of Computational Linguistics (ACL) to the authors of works published by the ACL makes discovery, access, and use of its research results easily available to both members and the general readership. The ACL Anthology initiative has contributed to the success of this mission, both as an archiving and dissemination vehicle for published works.

Started as a means to collect and preserve articles published by the ACL in 2001, the Anthology has since matured and now has well-defined workflows for its core missions. In 2009, the Anthology staff embarked to expand the Anthology's mission to meet two specific goals: on the backend, to enforce a proper data model onto the publication metadata; on the frontend, to expand the scope of the Anthology to encompass services that would best serve its constituents. Where possible, we adopted widely-deployed open source software, customizing it for the Anthology where needed.

With respect to the backend, the revamp adopted a database model to describe the publication metadata, implemented using MySQL. On top of this database layer, we chose Ruby on Rails as the application framework to interact with the data, and built suitable web interfaces to support both administrative and end-users. The backend also needed to support resource discovery by automated agents, and metadata export to sites that ingest ACL metadata.

With respect to the frontend, the Anthology website needed to meet the rising expectations in search and discovery of documents both by content and by fielded metadata. To satisfy both, we incorporated a faceted browsing interface that exposes metadata facets to the user. These metadata fields can be used to restrict subsequent browsing and searching actions to the values specified (e.g., *Year = 2001–2011*). Aside from resource discovery, the frontend also incorporated changes to support the workflow of readers and authors. We added both per-author and per-publication webpages. The publication pages invite the public to define content for the Anthology: anyone can report errors in the metadata, authors can supply revisions and errata, software and dataset links post-publication, readers can discuss the papers using the commenting framework

Praveen Bysani's work was supported from the National Research Foundations grant no. R-252-000-325-279.

[1] http://aclweb.org/anthology/; beta version 2 currently at http://aclantho3.herokuapp.com/.

Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 83–87,
Jeju, Republic of Korea, 10 July 2012.

in the system, and automated agents can use NLP and CL technology to extract, process and post information related to individual papers.

2 Revamp Design

Prior to our revamp, the Anthology's basic mission was to transcribe the metadata of ACL proceedings into a suitable form for the Web. To ensure widespread adoption, a simple XML format for the requisite metadata of *author* and *title* was created, with each ACL event's publication chair providing a single XML file describing the publications in each event and the details of the event (e.g., the volume's *booktitle* and *year*). Other fields were optional and could be included in the XML. The Anthology editor further added a unique identifier, an *Anthology ID*, for each publication record (e.g., "A00-1001"). Mandatory fields in the XML were extracted by a collection of programs to create the visible HTML pages in the Anthology website and the service export files, used to update the Association of Computing Machinery's (ACM) Portal[2] and the DBLP Computer Science Bibliography[3]. Prior to the revamp, this set of XML files – collected over various years – represented the canonical record of all publication data.

While easing adoption, storing canonical publication metadata as XML is not ideal. As it is stored across multiple files, even simple questions of inventory are hard to answer. As there was no set document type definition, the XML schema and enforcement of mandatory fields varied per document. In the revamp, we migrated the publication data into a database schema shown in Figure 1. The database form allows easy incorporation of additional fields that can be provided post-publication (including the Document Object Identifier, DOI, currently provided by the ACM by mutual agreement). The database structure also promotes publications, venues, and authors to first-class objects, enabling joins and views on the data, such as *paper—author* and *venue—special_interest_group*.The database currently has 21,107 papers, authored by 19,955 authors. These papers encompass one journal, 17 conferences and hundreds of workshops sponsored by 14 SIG groups. The publication years of these papers range from 1965 to 2012.

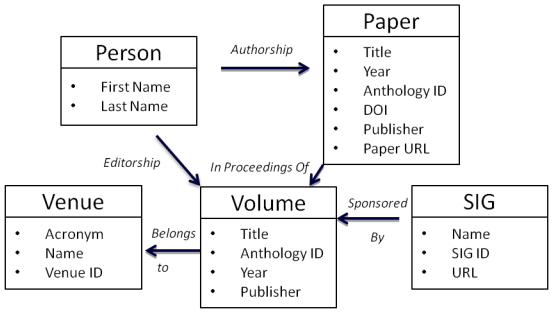

Figure 1: Current database schema for the Anthology.

The database's content is further indexed in inverted search indices using Apache Solr[4]. Solr allows indexing and querying in XML/JSON formats via HTTP requests, powering the frontend website search facility and enabling programmatic search by automated agents in the Anthology's future roadmap. We employ Ruby on Rails (or "Rails", version 3.1), a widely-deployed and mature web development framework, to build the frontend. It follows a Model-View-Controller (MVC) architecture, and favors convention over customization, expediting development and maintenance. Rails provides a closely tied model for basic database interactions, page rendering, web server deployment and provides a platform for integrating plugins for additional functionality. To enable faceted browsing and search, the revamped Anthology integrates the Project Blacklight[5] plugin, which provides the web search interface via our Solr indices. Rails applications can be deployed on many commercial web hosts but not on the current hosting service used by the primary ACL website. We have deployed the new Anthology interface on Heroku, a commercial cloud-based platform that caters to Rails deployment.

3 Frontend Form and Function

Of most interest to Anthology users will be the public website. The remainder of this paper describes

[2] http://dl.acm.org

[3] http://www.informatik.uni-trier.de/~ley/db/

[4] http://lucene.apache.org/solr/

[5] http://projectblacklight.org/, version 3.2

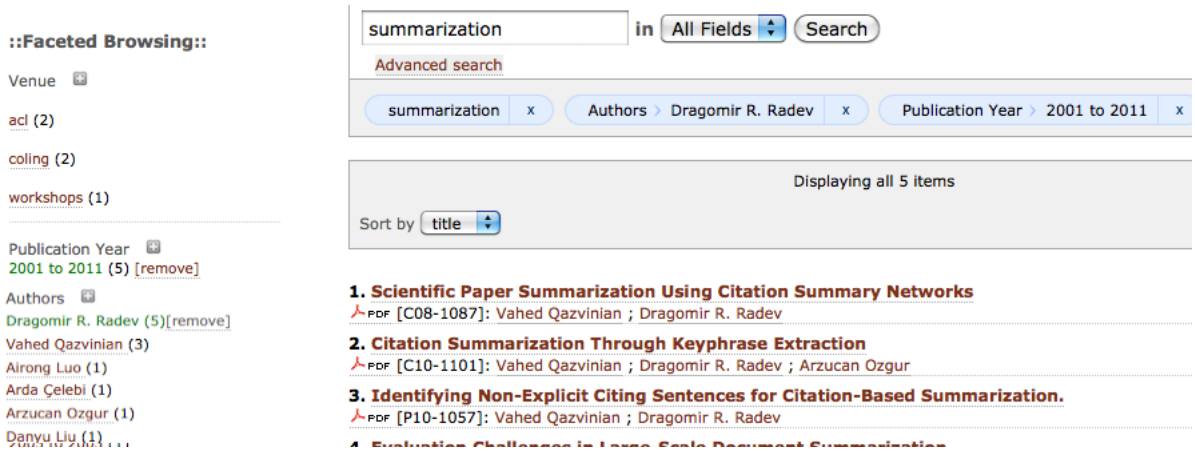

Figure 2: A screenshot of a faceted keyword search, showing additional restrictions on *Author* and *Year* (as a range).

the individual features that have been incorporated in the new interface.

Faceted Browsing: Facets let a paper (or other first-class object, such as authors) be classified along multiple dimensions. Faceted browsing combines both browsing- and search-based navigation: Anthology users can progressively filter the collection in each dimension by selecting a facet and value, and concurrently have the freedom of searching by keyword. It is a prevailing user interface technique in e-commerce sites and catching on in digital libraries.

The current Anthology defines five facets for papers. 'Author', 'Publication Year', 'Venue', 'Attachments' and 'SIG' (Special Interest Group) of the corresponding volume. The 'Year' facet further exposes an interface for date range filtering, while the 'Attachments' allows the selection of papers with software, errata, revisions and/or datasets easily. The website also has a standard search box that supports complex Boolean queries. Figure 2 illustrates some of these functions in a complex query involving both facets and keyword search. This is an improvement over the previous version that employed Google custom search, which can not leverage our structured data to add filtering functionality. Taking back search from Google's custom search also means that our search logs can be provided to our own community for research, that could enable an improved future Anthology.

Programmatic Contributions: The ACL community is uniquely positioned to enhance the Anthology by applying natural language technology on its own publication output. The ACL Anthology Reference Corpus (Bird et al., 2008) previously standardized a version of the Anthology's articles for comparative benchmarking. We take this idea farther by allowing automated agents to post-process information about any publication directly into the publication's corresponding page. An agent can currently provide per-paper supplementary material in an XML format (shown below) to the editor. After suitable validation as non-spam, the editor can ingest the XML content into the Anthology, incorporating it into the paper's webpage. Such functionality could be used to highlight summarization, information extraction and other applications that can process the text of papers and enrich them.

We use the Anthology ID to uniquely identify the associated paper. Currently the system is provisioned to support supplementary data provided as 1) text (as shown in Figure 3), 2) an embedded webpage, and 3) hyperlinks to websites (similar to how attachments are shown).

```
<paper id="P11-1110">
  <content name="keywords", type="text">
    <item>
      discourse, implicit reference, coherence,
      readability
    </item>
  </content>
</paper>
...
```

Figure 3: Excerpt of a programmatic contribution to the Anthology. The excerpt shows a keyword contribution on paper P11-1110.

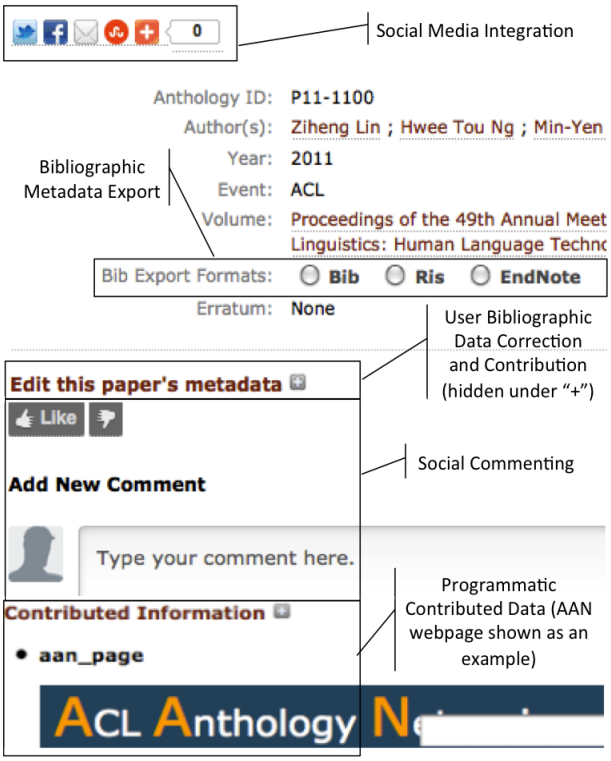

Figure 4: (Compressed) individual publication view with callout highlights of features.

Bibliographic Metadata Export: The previous Anthology exposed bibliographic metadata in BibTeX format, but its production was separate from the canonical XML data. In the revamp, we transform the database field values into the MODS bibliography interchange format. We then integrated the Bibutils[6] software module that exports MODS into four end-user formats: BibTeX, RIS, EndNote and Word. This lessens the effort for users to cite works in the Anthology by matching major bibliography management systems. Our use of Blacklight also enhances this ability, allowing the selection of multiple items to be exported to bibliographic exporting formats or to be shared by email.

User Contributed Data: While social media features are quintessential in today's Web, scholarly digital libraries and academic networks have yet to utilize them productively. One vehicle is to allow the readership to comment on papers and for those comments to become part of the public record. To accomplish this, we integrated a commenting plugin from Disqus[7], which enables users logged into other social media platforms to leave comments.

We also want to tighten the loop between reader feedback and Anthology management. Our revamp allows users to submit corrections and additions to any paper directly through a web form on the individual paper's webpage. Post-publication datasets, corrections to author name's and paper errata can be easily processed in this way. To avoid spam changes, this feature requires the Anthology editor to manually validate the changes. Figure 4 shows the individual publication view, with metadata, bibliographic export, metadata editing, commenting, and user (programmatic) contribution sections.

Author Pages: As a consequence of using Rails, it becomes trivially easy to create pages for other first-class data elements. Currently, we have created webpages per author, as shown in Figure 5. It gives the canonical listing of each author's publications within the Anthology in reverse chronological order and includes a list of the popular co-authors and publication venues. This feature brings the Anthology up to parity with other similar digital libraries. We hope it will spur authors to report publications under different variants of their names so a naming authority for ACL authors can result partially from community effort.

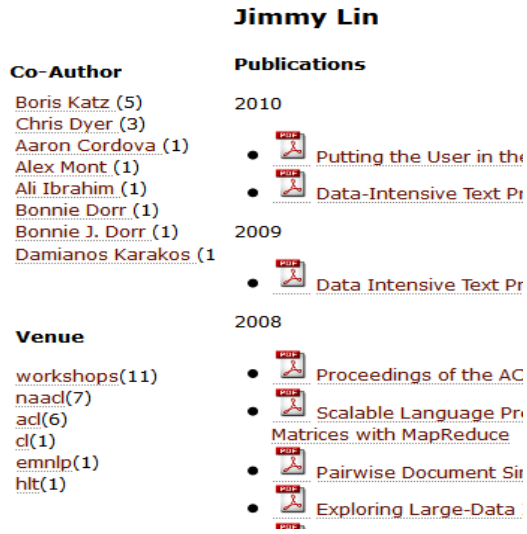

Figure 5: (Compressed) author page with corresponding co-author and venue information.

[6] http://sourceforge.net/p/bibutils/home/Bibutils/

[7] http://www.disqus.com

4 Usage Analysis

The revised Anthology interface is already seeing heavy use. We analyzed the application logs of the new Anthology website over a period of five days to understand the impact and usage of the new features. During this period the website has received 16,930 page requests. This is an increase over the original website, which garnered less than 7,000 page views during the same period. The average response time of the server is 0.73 seconds, while the average load time of a page is measured at 5.6 seconds. This is slow – web usability guidelines suggest load times over 200 milliseconds are suboptimal – but as the website is deployed on the cloud, server response can be easily improved by provisioning additional resources for money. Currently the new Anthology interface is run on a no-cost plan which provides minimal CPU bandwidth to serve the dynamically generated webpages to the readership.

The majority of the requests (11,398) use the new faceting feature; indeed only 30 requests use the traditional search box. The most used facet patterns include "Author, Venue" (51.6%) followed by "Author, Venue, Year" (14.8%). While we believe that it is too early to draw conclusions on user behavior, the overwhelming preference to use facets reveals that faceted browsing is a preferable navigational choice for the bulk of the Anthology users.

3,180 requests reached individual (detailed) publication views, while 2,455 requests accessed author pages. Approximately 62% of the total requests had a visit duration under 10 seconds, but 22% requests last between 11 seconds to 3 minutes, with the remaining 16% sessions being up to 30 minutes in length. The noticeable large ratio of long visits support our belief that the newly-added features encourages more user engagement with the Anthology. Since the website went live, we have received 3 valid requests for metadata changes through the new interface. Up to now, there has not been any use of the social media features, but we believe Anthology users will adopt them in due course.

5 Conclusion and Future Work

S.R. Ranganathan, arguably the father of faceted classification, proposed that "the library is a growing organism" as one of his laws of library science (Ranganathan, 1931). We observe that this is true in the digital context as well.

We will support the legacy ACL Anthology interface until the end of 2012 in parallel with the new interface, gradually phasing in the new interface as the primary one. Our immediate goal is to flesh out the per-author, -venue, -SIG views of the data, and to enable resource discovery via Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al., 2002), an open protocol for harvesting metadata by web crawlers. Our medium term outlook hopes to furtherincorporate grassroot ACL resources such as the ACL Anthology Network (Radev et al., 2009) and the ACL Searchbench (Schäfer et al., 2011).

We are most excited by the ability to incorporate programmatic contributions made by NLP software into the Anthology. We hope that the community makes full use of this ability to showcase the importance of our natural language processing on scholarly data and improve its accessibility and relevance to others.

References

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC'08*.

Carl Lagoze, Hebert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. The open archives initiative protocol for metadata harvesting, version 2.0. http://www.openarchives.org/OAI/2.0/ openarchivesprotocol.htm, June.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology Network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61.

S. R. Ranganathan. 1931. *The Five Laws of Library Science*. Madras Library Association (Madras, India) and Edward Goldston (London, UK).

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology Searchbench. In *Proceedings of the 49th Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 7–13.