ACL ANTHOLOGY Report, July 2006
Steven Bird

The ACL Anthology is a digital archive of research papers in computational linguistics, sponsored by the CL community, and freely available to all. It includes the Computational Linguistics journal, and proceedings of many conferences and workshops including: ACL, EACL, NAACL, ANLP, TINLAP, COLING, HLT, MUC, and Tipster.

The anthology now contains just over 11,000 papers (up from 9365 papers twelve months ago), along with full-text search. Most of the papers are also indexed by Citeseer and scholar. google, helping the citation counts of ACL authors. The ACM Digital Library is creating rich metadata and doing full citation linking for all anthology materials.

ADDITIONS OVER LAST 12 MONTHS: Proceedings from ACL-06, EACL-06, HLT-NAACL-06, HLT/EMNLP-05, IJCNLP-05, CL Journal 2004 (always one full year in arrears).

PERSISTENT URLs: The ACL website supports persistent URLs for all papers that are resolved to a copy at the selected mirror site. These URLs have the form http://www.aclweb.org/anthology/P99-1012, and they may be used for citation purposes (e.g. in BibTeX entries).

FUTURE MATERIALS: Jason Eisner and Philipp Koehn have modified the ACL publication software to generate conference CD-ROMs using the same directory layout and file-naming conventions as the anthology. BibTeX files are automatically generated and made available to users. It is now much easier to incorporate new materials into the anthology.
Conference proceedings are published in the anthology at the same time as the conference. The journal and any SIG workshops not held in conjunction with an ACL meeting will continue to require manual processing. Unfortunately this year's EACL meeting did not use the ACL publication software and the materials needed to be processed manually by the EACL publications team after the conference. The instructions for the publication software need to be updated to cover two further tasks: (i) obtaining the workshop identifiers from the editor, and (ii) uploading the materials to the anthology.

ONGOING ACTIVITIES

DIGITAL OBJECT IDENTIFIERS: These are akin to ISBN numbers, but apply to individual papers. In collaboration with the ACM we will assign DOIs to each anthology item in the coming year. The nominal cost for DOI assignment is $1 per article, or $10k for the whole anthology.
The ACM will cover the cost for past materials, while the ACL will cover the cost of DOI assignment for anthology materials from 2006 onwards.

CONVERSION OF PAST CONFERENCE MATERIALS: Several conferences in 2003 and 2004 produced electronic materials that are not compatible with the anthology. These materials are hosted on the site as they appeared on CD-ROM, but they still need to be converted to the anthology layout and naming scheme for consistency.

SITE DESIGN: The site design is simple and functional; user input on improving the design would be welcomed.

REQUESTED MATERIALS: Anyone with hardcopy of MUC 1998 proceedings or a CDROM copy of HLT 2002 is requested to contact the editor.

TOPICAL INDEXING: The existence of persistent URLs makes it easy for individuals and special interest groups to set up annotated bibliographies with pointers to papers in the anthology. Moreover, the community's own text categorization techniques ought to be applied to its own text collection. The anthology site should link to any well-curated, comprehensive categorizations of its content, so that members of the CL community can benefit from them.