Effective Creation of Self-Referencing Citation Records

System SelfBib

Tomáš Čapek and Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xcapek1@aurora.fi.muni.cz, sojka@fi.muni.cz

Abstract. Acquiring citation records from online resources has become a popular approach to building a bibliography for one's publication. LaTeX document preparation system is the most popular platform for typesetting publications in academia. It uses BibTeX as a tool used to describe and process lists of references. In this article we present a simple method that allows the automatic creation of a full self-referencing citation record for a collection of papers typeset and published within one proceedings of a conference. This greatly facilitates access to the bibliography entries for anyone who wishes to use them as part of their own publication.

1 Introduction

Mathematicians, engineers, philosophers, lawyers, linguists, economists and other scholars all appreciate quick access to other people's research not only in terms of its actual content but also to get bibliography entries, and especially if they are using LaTeX and BibTeX for typesetting. With the growth of widely accessible citation databases and search engines that aggregate scholarly literature there is a need to not only retrieve the information it contains but also to provide information about the publications we produce. For example, the automatic parsing of publications provided by Google Scholar does not always correctly identify all necessary bibliographic data and sometimes even mixes different fields up. The same holds for citation extraction services like that offered by Mendeley.[1] To prevent incorrect metadata records, it is the responsibility of each author, editor or publisher, depending on the scope of the publication, to present their own online publications in such a way that avoids the need for guessing on the part of the indexing engine. There are already established channels for the big players (CrossRef, Google Scholar, Elsevier, Springer, Thompson Reuters) to exchange, validate and match paper metadata. Metadata are often retyped, or produced semi-automatically, which is still error-prone. The optimum point is when the metadata are *generated* during the preparation and typesetting of the publication. In this setup, with

[1] http://www.mendeley.com/bibliography-maker-database-generator/

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 97–102, 2010.

batch typesetting systems, such as LaTeX, metadata which appear in the final version of the publication, end up in a metadata record without any human interference. This idea is most likely already employed in commercial systems such as the one by Elsevier and others [1], but we are not aware of any "poor man's solution" for authors and editors. Based on our experience preparing more than twenty multi-author books and proceedings, we have designed and implemented the system *SelfBib* that automates the production of metadata records as a by-product of typesetting multi-author volumes and proceedings. Accurate and timely accessible citation records help to better identify paper duplicates appearing on the Internet, increase the ease of citation and to a degree also the citation rate.

In our view, the best practise is to typeset a book or a proceedings in a single run with a single LaTeX source file via a set of utility scripts. We describe the main aspects of this approach in Section 2. We show how to easily enhance the typesetting process work flow to provide a full and accurate self-referencing citation record with *SelfBib* in Section 3. Finally, we evaluate the "*SelfBib* approach" and its application in Section 4 and wrap up in Section 5.

2 Prerequisities for Typesetting

The main task of a proceedings' editor is to collect and unify the heterogenous papers contributed by authors. More often than not, authoring instructions even allows the use of different systems (Word or TeX) which makes enforcing the publisher's format a very tedious and time-consuming task. In a research setup, it is often expected that editors also provide a table of contents, author or subject index. This could be hardly achieved automatically without typesetting the whole volume in a single (LaTeX) run – otherwise it implies a lot of manual work with any last minute edit. A prerequisite for automated processing of a complete volume is having all contributions converted (or at least their metadata) into a uniform format, or having the metadata collected into one place. Some supporting systems, such as A. Voronkov's *easychair* do provide rudimentary support for editing the Table of Contents pages, but this does not produce a reliable product, especially when working under the pressure of deadlines.

We recommend working with the uniform format, LaTeX, as it is stable, reliable, and widely used by the scientific community. Most metadata are already tagged in the primary source files (`\title`, `\author`) and others are available during the typesetting (e.g. page numbers). The plain (non-binary) format of LaTeX also allows a high degree of automation, and the unification into one format greatly increases the uniformity of the typeset volume. Good and consistent markup then allows many innovative uses, generating multiple indexes (author, name, subject), hypertext linking across the volume and multiple output formats [2], features usually available for monographs only.

We have designed and implemented a system that allows the typesetting of individual articles and the whole volume in one LaTeX run, in parallel from the same files. During the LaTeX run, additional information is written by standard

and custom macros into an auxiliary (.aux) file. This information is sufficient to build full metadata records for the contributed papers and the whole volume. A script is then run on an auxiliary file, which parses and processes the data into the required formats such as BibTeX (see Section 3).

A typical work flow starts with papers being typeset individually, and then the source completeness is checked. Papers are assigned reference numbers, usually by the supporting reviewing system, and files are renamed using a naming scheme based on these unique paper reference numbers. The reference number is used for multiple purposes, e.g. for naming the directory of the paper, for the name of the root paper's TEX source, for prefixing label names in the paper so that they are unique across the whole volume, etc. This naming scheme allows the editing of the tree of LATEX source files to be partially automated. Several scripts have been developed to facilitate the editing process.

The metadata record of publication item contains data of three kinds:

- data provided by authors (title, list and order of authors and their affiliations, abstract,...)
- data supplied by publishers (publisher name, publishing date, ISBN,...)
- data created during typesetting (page numbers)

The author metadata are already tagged in the primary sources, and can be grabbed from there. The publisher's metadata are usually the last items to be typeset, and with good typesetting conventions they are also defined and tagged unambiguously in the LATEX source files of the publication. The idea is to collect all these data during the final LATEX run and create the full metadata records automatically, as a by-product of the volume production.

On the TEX level, the system consists of

- macros for writing the metadata information into an auxiliary file.
- macros and methodology (naming, tagging, placing local macros) to allow the same files to be used when typesetting a single paper or the whole volume.
- scripting automation (Makefile) to manage the series of typesetting actions and calling the appropriate programs in the right order.

3 SelfBib

SelfBib system consists of several components. The main one is a script (implemented in Ruby programming language [3]), which parses the auxiliary (`.aux`) file from a LATEX run of the whole book and produces well-formed `.bib` file where for each paper within the proceedings the metadata about its title, authors, and first and last pages of the paper in the book are retrieved. In addition to that, a cross-reference key to the primary bibliography entry, which contains information common to all of the papers in the book, is added as well.

In Figure 1 is a sample of the *SelfBib* output consisting of the primary entry and one additional entry for a paper.

```
@proceedings{tsd10conference,
title={{Proceedings of the 13th International
        Conference on Text, Speech and Dialogue---TSD 2010}},
year=2010,
editor={Petr Sojka, Ale{\v s} Hor{\'a}k,
        Ivan Kope{\v c}ek and Karel Pala},
nddress={Brno, Czech Republic},
month=Sep,
publisher={Springer-Verlag},
}

@inproceedings{tsd10conference:100,
    title={{Parsing and Real-World Applications}},
    author={John Carroll},
    pages={2--4},
    crossref={tsd10conference},
}
```

Fig. 1. Sample of *SelfBib* output.

SelfBib has several useful features. For maximum portability, all strings are encoded in 7-bit ASCII so that all entries can be copied as they are, regardless of the language the recipient uses for typesetting. All non-ASCII characters are encoded in LaTeX macros. Also, to ensure that all entries sort correctly, the non-ASCII characters use extended syntax delimited by curly brackets as follows: `{<macro><character>}`. For example, the "š" character is encoded `{\v{s}}`. All frequent variants of accented characters are stored separately in a hash structure and can be extended at will. As an alternative output, *SelfBib* can also provide Google Scholar-compliant HTML meta tags[2] instead of BibTeX entries. The meta tags are useful to include in HTML pages which are dedicated to a single paper. As a result, Google Scholar will always index the metadata as they appeared in the paper without guessing and parsing them from the PDF. This increases the citation matching and lining precision and ensures providing correct bibliography entries.

4 Deployment and Evaluation

When we finish the typesetting of a conference proceedings, there is a variety of ways to promote the self-referencing list of citations for it to be as accessible as possible for anyone who might wish to use one or more entries in their own publication. The most straightforward and natural way is to provide the full reference list for download on the conference homepage. This, however, might not be helpful to users who are unaware of the conference itself and are interested in one particular paper in it, which they have found via a search

[2] `http://scholar.google.com/intl/en/scholar/inclusion.html`

engine. For example, for a paper to appear in Google Scholar results,[3] it needs to be either parsed from the PDF or be accessible in a single (landing) HTML page. For its bibliography record to be accurate, the landing page needs to contain a special set of HTML meta tags[4] that describe the metadata. *SelfBib* can produce bibliography entries in this format as one of its options.

Another way to make the citation list available online is to add the `.bib` file to a online bibliographic database dedicated to a particular field of study. For the field of computer science, the DBLP database is the largest and the most popular resource of bibliographic information[5]. BibTeX format is among those supported that can be used to quickly make the whole citation list available to a large number of scholars via the BibTeX ingestion driver.

Once the accurate and complete metadata item reaches any of the main bibliography citation providers, it tends to be spread via records exchange and matching in systems such as Google Scholar, Mendeley, Bibsonomy, CiteUlike, DBLP, CiteSeer, Crossref and others.

We have generated bibliographic records for twenty proceedings to demonstrate the usefulness of our approach. They are available at the project's web page `http://nlp.fi.muni.cz/projekty/selfbib/bib/`. The system has been proven useful and it significantly facilitates citing bibliography items correctly and efficiently, which in turn potentially increases the citation rate of the papers.

5 Conclusion

In this paper, we have introduced an easy method to enhance the typesetting process of a multi-author volume or an academic proceedings to provide a full and accurate self-referencing list of citations as its by-product. Although our approach can only be used with LaTeX and BibTeX systems, its main advantage is that it is fully automated and quite easy to set up. Depending on the deployment method, the list of citations can make it much easier for anyone compiling a bibliography for their own publication to get access to properly formatted metadata about our publications, or even to help promote our publications by exposing it to a larger number of potential readers.

Acknowledgements This work has been partially supported by the Ministry of Education of CR within the Center of Basic Research LC536 and by the European Union through its Competitiveness and Innovation Programme (Policy Support Programme, "Open access to scientific information", Grant Agreement No. 250503).

[3] http://scholar.google.com/intl/en/scholar/inclusion.htm [4] Google Scholar supports the following tag sets: Highwire Press tags, Eprints tags, BE Press tags and PRISM tags. [5] Primary URL is located at: `http://www.informatik.uni-trier.de/~ley/db/`. Alternate server with limited search capabilities can be found at: `http://dblp.uni-trier.de/` [4]

References

1. Bazargan, K.: LaTeX to MathML and back: A case study of Elsevier journals. In: Proceedings of Practical TeX 2004, TUG (2004).
2. Sojka, P., Růžička, M.: Single-source publishing in multiple formats for different output devices. TUGboat **29**(1) (2008) 118–124.
3. Flanagan, D., Matsumoto, Y.: The Ruby Programming Language. (2008).
4. Ley, M.: DBLP – Some Lessons Learned. PVLDB **2** (2009) 1493–1500.