Word to LaTeX for a Large, Multi-Author Scientific Paper

D. W. Ignat

Abstract

Numerous co-authors from diverse locations submitted to a scientific journal a manuscript for large review article in many sections, each formatted in MS Word. Journal policy for reviews, which attract no page charges, required a translation to LaTeX, including the transformation of section-based references to a non-repetitive article-based list. Saving Word files in RTF format and using `rtf2latex2e` accomplished the basic translation, and then a `perl` program was used to get the references into acceptable condition. This approach to conversion succeeded and may be useful to others.

1   Introduction

Twelve authors from five countries and ten research institutions proposed to the *Nuclear Fusion* journal (*NF*) of the International Atomic Energy Agency (IAEA) in Vienna, Austria, a review paper with six sections plus a glossary.

This unusually large manuscript had some hundred thousand words and a thousand references. The sections had different lead authors, so that the references of each section were independent of those in other sections, while often repetitive among sections.

The IAEA gave review papers the privilege of waived publication charges ($150/page), but required authors to ease the publisher's costs by submitting

manuscripts of reviews in LaTeX, the journal's typesetting system. Therefore, a considerable financial incentive appeared for finding a somewhat automated transformation of all the Word sources into a unified LaTeX source.

I was the editor of IAEA's *NF* from mid-1996 to mid-2002 with primary responsibility for the refereeing system and the development of the journal. Previous experience in Unix and and LaTeX for my own research brought an unofficial role as adviser to the IAEA production office on shell scripts, LaTeX, regular expressions, `perl`, and web mounting.

Since the paper appeared valuable from the point of view of journal development, and at the same time a challenge in computer processing, I became particularly interested, and encouraged the authors find ways to satisfy the IAEA requirement: a LaTeX manuscript to better support refereeing and eventual publication.

In the end, the paper in question [1] was published in *NF* and was very well received by the research community, at great credit to the co-authors and also good for *NF*.

When the recent call for papers at PracTex came in it occurred to me that the story might be interesting for this audience.

2   Translation from Word to LaTeX

At the time of submission (mid-1999) the IAEA and *NF* had investigated with a consultant conversions from Word to LaTeX, but had not found a satisfactory solution. One of the twelve authors suggested "rtf2latex2e" by Ujwal Setlur Sathyam (now Ujwal Setlur) and Scott Prahl, following the Word-native RTF (Rich Text Format) writer.

Here is Microsoft's description of RTF from `msdn.microsoft.com`,

"The Rich Text Format (RTF) Specification provides a format for text and graphics interchange that can be used with different output devices, operating environments, and operating systems. RTF uses the ANSI, PC-8, Macintosh, or IBM PC character set to control the representation and formatting of a document, both on the screen and in print. With the RTF Specification, documents created under different operating systems and with different software applications can be transferred between those operating systems and applications."

`rtf2latex2e` uses the RTF reader by Paul DuBois and converts RTF files to LaTeX2e. Some features are: detects text style (bold, italic, etc.); reads embedded

figures; reads tables; converts embedded MathType; converts most Greek and math symbols; reads footnotes; translates hyperlinks. It should compile on any platform that supports a C compiler. Versions for Macintosh, Unix-type systems, and Windows are available. The distribution, issued under the terms of the GNU General Public License as published by the Free Software Foundation, comes with example `.rtf` files.

The current, and final, version of `rtf2latex2e` can be found on the Comprehensive TEX Archive Network, `ctan.org/tex-archive/support/rtf2latex2e` and at `sourceforge.net`.

The result of translation gave the expected

`\documentclass{article}`

`\begin{document}`

`\section*{1. INTRODUCTION}`

and looked good regarding mathematics and tables, but left all citations as footnotes with the expected chaos with repeated references. A typical reference (of the thousands) appeared many times with different chapter-based numbers.

The footnotes were rendered, for example, as `[1.\footnote{[1.] Author, A., Some Journal \textbf{36} (1997) 123.}]` in section 1, but in generally the same way in section 2, except that the "`[1.`" became "`[2.`".

One task is to transform the `\footnote` style that survives after the Word-rtf-LATEX transformation into the normal `\cite{...}`–`\bibitem{...}` representation of references. More complicated is to detect as identical those references to the same work presented with slight differences; and to detect as distinct those references that are actually different but "look" similar.

The goal was a *unique* citation in the body, such as `\cite{AuthorA36p123}`, and a corresponding entry in the bibliography, such as `\bibitem{AuthorA36p123}` `Author, A., Some Journal, {\bf 36} (1997) 123.`

The power of "`perl`" (Practical Extraction and Reporting Language) and its version of "regular expressions" made order from chaos, and produced material suitable for refereeing, and, eventually, publication.

3   A Sketch of the Process

A "regular expression" (regex) is a generalized string for matching patterns, and possibly replacing whatever is found found with something else. The programs

grep (Global Regular Expression Printer) `sed` (Stream EDitor) and the text editor `emacs`, all of which are part of Unix-like systems, incorporate regex-es.[1]

The compositors at *NF* had the following problem, which was tamed with regular expressions in `emacs`. Journal style treated the names of authors in references as "A.B. Smith" but manuscripts often arrived with treatments such as "SMITH, A. B." or "Smith, A.B." or "A. B. SMITH." Making changes by hand was obviously prone to error, not to mention laborious for those papers needing dozens to hundreds of changes. A little experimentation with a couple keyboard macros on my part greatly simplified a frustrating chore in typesetting.

What follows is a description of regex and `perl`. If you like, you may skip to section 4.

For a flavor of the regex world from the programming point of view:

`/s/Old/New/g` : Old → New globally (g)

`/s/^Old/New/` : Old → New at line start (^)

`/s/(...)Old/NEW\1/` : xyxOld → NEWxyz

In the last example, the string `Old` is sought, but only if it preceded on its line by 3 characters, which are to be remembered by the parentheses ( ) with the label `\1`. Then, `Old` is to be replaced by `NEW` but *followed by* the 3 characters just found (here called `xyz`).

These examples only suggest the full power of searching and replacing available, in particular with `perl`.

A short summary of the regex usages is in "Linux in a Nutshell" [2], and an excellent introduction is in the Wikipedia [3].

For an advanced treatment, see "Mastering Regular Expressions" [4]. The "GNU Emacs Manual" [5] explains using regex-es in editing.

The prime documentation of `perl` is the "Camel book" now in its third edition [6]. The *NF* Office happened to rely mostly on the "Llama book" [7] and the pocket-size Desktop Reference by Johan Vromans [8].

The lead author of *Ref. 1*, Gianfranco Federici, contacted a colleague, Andreas Schott, about the challenge of rationalizing the references. Schott produced a `perl` script `foot2cite.pl` which accomplished the task.

A few years previously the *NF* Office had developed a collection of `bash` [9] shell and `perl` scripts to produce print masters and files for mounting pdf and

---

[1]The tools mentioned above had versions workable under Windows 95, but comments on the capability of later Windows and Macintosh systems are outside the scope of this document.

html [10] articles on the IAEA web server. The LaTeX source of individual articles led to tables of contents and indexes of authors and subjects from individual article source files with the help of native LaTeX markup plus additional markup commands of the local style file. ¿From that experience[2] it appeared interesting to develop an IAEA-local program which could be the base of solutions that might be needed in the future.

Perhaps some readers have heard of the humorous comment that the term `perl` might have at one time been an acronym for "Particularly Eclectic Rubbish Lister." Certainly a casual glance at a `perl` script can scare off even people with some programming experience, especially experience in Fortran. All those odd backslashes changing the meaning of all characters; all those single letters meaning complex and subtle manipulations! It does appear daunting.

One point of this paper is that some significant goals can be accomplished by a `perl` novice after reading through the Llama book [7], putting a compact summary [8] in ones "lap," and trying things, one line at a time. Even a description of the strategy (suppressing details of the compact syntax) can slow down the average reader, so I choose, with the advice of the PracTeX Editor, to move on to the results of the scripting process.

For those interested, some features of the resulting `ref_manip.pl` are described in the following section 5.

4   The Result

The processing into LaTeX of the first draft manuscript created one format completely common to all contributing institutions and authors. With that common form, adjustments in response to the concerns of the *NF* editorial office and referees became easier, as did changes originating with the paper's authors as the review developed. Even so, the refereeing process was extensive, which is not uncommon for articles appearing in *NF*, and particularly articles of such a length.

Independent of what the authors of *Ref. 1* feel about their article and the process of publishing it, the publishing journal and its home organization have interests.

---

[2]The utility of combining LaTeX with scripting languages has been explored recently for the TUG; see for example William M. Richter, "TeX and Scripting Languages" TUGboat, Vol 25, No. 1, p 71 (2004).

The Institute for Scientific Information (ISI) keeps track of an "Impact Factor" (IF) for thousands of journals [11]. The IF is (at least approximately) the number of citations to a journal divided by the number of articles in the period studied. *Nature* and *Science* have IFs in the 20–30 range. The very prestigious *Physical Review Letters* has an IF around 6, and the *Physical Review*, (series A, B, C, D, and E) is typically between 2 and 3. Journals covering plasma physics and nuclear fusion range from 0.5 to 3 or so, and *NF* is consistently the highest in the group. In the six years ending in 2003 *NF* was between 2.2 and 3.4.

According to Google's newly introduced "Scholar" service, articles from all journals covered referenced *Ref. 1* 23 times, which is unusually high for the subfield of science and engineering covered by *NF*. (The time frame was not apparent from the information at Google.)

Records available at the IAEA show that for *Ref. 1* there were 162 downloads in 2003, placing it number 7 in the top 10 downloads for that year; and that the citation rate is roughly double the next most cited article, and far above the average rate.

The numbers quoted above suggest that the research community received *Ref. 1* unusually well, making it a fine credit to each of the authors and to their institutions. The numbers also say that the article had significant positive influence on the IF of *NF*, and therefore a positive influence on the continued success of *NF*. In other words, the appearance of this article was very good for its authors as well as the IAEA and *NF*.

Remembering that publication in *NF*, and at low cost to submitters, required a LaTeX manuscript, one can wonder if all the good news would have happened without `rtf2latex2e` and `perl`. My speculation:

1. the research paper would have come out, if at all, later than it did,
2. it would not have appeared in *NF*,
3. the authors would not have gotten quite the recognition they did,
4. the IAEA and its journal would have a lower IF for the relevant period.

5 Some Details on Rationalizing the References

The idea was to use the "hash" facility in `perl`. Here, a hash is a 1-dimensional array in which the index and the value of the index are both character strings.

6

The 1-to-1 hash `num2cite` connected the Word-original reference number, such as "`1.101`," to a string designed to be unique (except in pathological circumstances) such as "`AuthorA36p123`." For diagnostic purposes the 1-to-1-or-more hash `cite2nums` connected the (uniquely created) `\cite` and `\bibitem` string such as "AuthorA36p123" to the (multiple, in general) original reference numbers such as "1.101"; "2.45".

The multiple LATEX section files produced by `rtf2latex2e` are scanned in sequence for a footnote. If footnote-style text of the nature author-journal-volume-page is found, then an identifier string is made of the first author's last name, first initial, volume number, page number (AuthorA36p123). The text of the reference is entered in a holder for the bibliography under `\bibitem{AuthorA36p123}`, while the footnote is replaced by `\cite{AuthorA36p123}`. Next, the two hashes receive the appropriate entries — such as "1.101" and "AuthorA36p12" — with the help of a counter in the `perl` script. That counter should not become out of synchronization with the footnote numbers given in the paper unless there is is a mistake in the original text.

The references not citing journal articles are detected by the absence of a bolding of an alphanumeric volume number in a footnote. In that case, the `cite` and `bibitem` identifier is formed from the first twenty alpha-numeric characters in the citation, excluding all white space. Again, the text of the reference goes to the bibliography as a `\bibitem`.

If the footnote is to a previously used number, such as [`1.101`] or [`2.202`], then the `num2cite` hash is used to enter the citation with the `\cite` format, without adding anything to the bibliography.

In the script as developed, pre-processing of the raw section files does, for example, the following:

- takes out explicit section numbering

- makes all citations (recall, they are of the `\footnote` type) begin in column one as `\CITE[...]` and occupy one entire (sometimes very long) line

- makes the bolded volume numbers into a particular form that will not confuse later searches for a right brace closing the footnote.

That pre-processing is no doubt a sign of ignorance of the full power of `perl`, and no doubt extends the execution time. However, execution time is not a practical issue, but being able to construct the script in small pieces that do small, easily

testable, things was very much an issue in the environment of the *NF* editorial and production offices.

The final pass changes `\CITE[...]` into `\cite{...}`, writes the `\bibitem{...}` entries, and, optionally, saves the hashes `num2cite` and `cite2nums` for diagnostics.

There are vulnerabilities. A simple one, which could be programmed around, is that the original footnotes cannot contain inside them the characters `[ ]` or, other than for volume bolding, `{}`. A more difficult vulnerability, practically speaking inevitable, is that truly identical references have to be presented in pretty much the same way. There is probably no automated way to defend against typographical errors in the names, volume or page numbers. (The potential vulnerability to different amounts of white space had a simple defense.) However, an off-line sort of all the `\cite` and `\bibitem` texts would have a good chance of revealing a problem.

6 Acknowledgments

References

[1] G. Federici, C. H. Skinner, J. N. Brooks, J. P. Coad, C. Grisolia, A. A. Haasz, A. Hassanein, V. Philipps, C. S. Pitcher, J. Roth, W. R. Wampler, D. G. Whyte, "Plasma-material interactions in current tokamaks and their implications for next step fusion reactors," *Nucl. Fusion* **41**, No. 12R (2001) 1967-2137.

[2] Jessica Perry Hekman, "Linux in a Nutshell," O'Reilly and Associates, Inc., 1997.

[3] Wikipedia, the Free Encyclopedia, `en.wikipedia.org/wiki/Regular_expression`

[4] Jeffrey E. F. Friedl, "Mastering Regular Expressions," O'Reilly and Associates, Inc., 1997.

[5] "The GNU Emacs Manual," 14$^{\text{th}}$ edition for version 21.3, Free Software Foundation, 2004. Online at `www.gnu.org/software/emacs/manual`

[6] Larry Wall, Tom Christiansen, Jon Orwant, "Programming Perl," (Third Edition) O'Reilly and Associates, Inc., 2000.

[7] Randal L. Schwartz and Tom Christiansen, "Learning Perl," O'Reilly and Associates, Inc., 1997.

[8] Johan Vromans, "Perl 5 Desktop Reference," O'Reilly and Associates, Inc., 1996.

[9] Cameron Newham and Bill Rosenblatt, "Learning the bash Shell," O'Reilly and Associates, Inc., 1995.

[10] Ian H. Hutchinson, "TtH: a TeX to HTML translator," `hutchinson.belmont.ma.us/tth/manual/`

[11] See `http://www.isinet.com/`.

The mapping between the LaTeX code and the compiled document has the virtue of being somewhat more predictable than what a WYSIWYG interface can deliver. There is LyX, of course, which is an excellent WYSIWYM tool for writing scientific documents, and it works great with documents with uncomplicated math. However, I found it unwieldy for writing math that required special packages, TikZ etc. (though I suspect it's just my not being used to the interface). If you don't need to do complicated stuff, I recommend trying LyX. The section/equation/chapter numbering etc. alone makes it pref Preparing Manuscripts using LaTeX. For quickest processing, we prefer to receive final manuscripts from authors in Word docx format. If you wish to use LaTeX, please observe the guidelines on this page in setting up your paper. Make sure that your PDF files are legible on the computer screen to avoid delays in processing. Initial submission: To prepare your manuscript in LaTeX, please follow the LaTeX source guidelines below. â€ Corresponding author.

cls score: 0.70 fus score: 0.75. cls score: 0.78 fus score: 0.39. Figure 1: Visualization of detection results and semantic segmentation feature maps.Â Wordsup: Exploiting word annotations for character based text detection. In Proc. ICCV. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Big-orda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. Icdar 2013 robust reading competition.Â Large kernel mattersimprove semantic segmentation by global convolutional network. In Computer Vision and Pat-tern Recognition (CVPR), 2017 IEEE Conference on, 1743â€" 1751. IEEE. Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015.