pdfTEX and XML in the Workflow for the Production of Conference Proceedings

Volker RW Schaa
DANTE e.V.
v.r.w.Schaa@gsi.de

Abstract

Two years ago I was appointed as Proceedings Editor for two conferences. Out of a sudden I had to define aworkflow in order to prepare and take care of all parts of the conference publication data: abstracts and paper submission, as well as maintaining participant data (so-called affiliation data) for a web presentation, the conference volume (Proceedings) as printed paper and CD version.

The two conferences do not differ substantially in size (i.e. number of contributions). The first one (DIPAC2003) had 85 contributions with 300 print pages and 290 authors, the second one (LINAC2004) will cover 450 contributions and 1400 print pages by approximately 1200 authors. However, both are too large to think of a manual production of the proceedings and web pages.

In order to become acquainted with existing conference systems in particle physics, the tools and scripts of PAC/EPAC were examined. These are combined at JACoW, the "Joint Accelerator Conferences on Web" site (http://www.jacow.org), hosted and organized by CERN, the world's largest particle physics laboratory situated in Geneva/Switzerland.

PAC/EPAC conferences comprise approx. 1200 contributions, 4000 pages and 7500 authors. The existing tools and scripts (written in Perl and Visual Basic under Windows) cover only a part of my tasks: hidden field entries of title, subject, authors, and keywords in the pdf files of the contributions, and second, the generation of web pages with meeting, author, and keyword indexes. Printable proceedings are still produced from the single contributions using word processor software (Word, QuarkXPress, etc.). An automated (batch-oriented) processing does not exist.

The workflow combining web presentation and generation of proceedings volume, starts from one XML file. This file is a database export containing all descriptions concerning contributions and participants data. Parameters for directory names of abstracts, contributions, HTML pages, and additional material can be defined in a config file.

The XML file is read, interpreted by a perl script, which generates HTML pages, TEX files, and batch procedures. A TEX control files ensure that from each contribution's pdf file a new pdf file is produced, which then contains all relevant data in the hidden fields, in addition to header and footer information (name and place of the conference, session and paper code, page numbering). For the proceedings volume a further TEX control file is produced.

In the physics world nobody is seriously thinking of dropping indices (or e^+, e^-), let alone special character like α, β, etc. These common symbols are hardly supported on web pages. To overcome this in an 7/8 bit environment, database entries use a (La)TeX notation, and all generated HTML pages are coded in unicode. Thus achieving in addition a proper display of author names with accents and special characters.

The conference web page, which was generated using the presented workflow and the new scripts can be found under the URL http://bel.gsi.de/dipac2003/ showing all contributions and the proceedings volume.