Database publishing with LuaTeX and the speedata Publisher

Database Publishing is the repetitive (semi) automatic transformation from a data source to some kind of output (HTML, PDF, epub,…). A common task is to have an excel sheet, a product information management system or a webshop database and generate reports, data sheets, product catalogs or other kind of PDF documents out of it. Database publishing is often equal to 'InDesign Publishing' with the help of some plugin that automates the task of pulling data from the database into the document. The user can (and must) make the resulting document more beautiful.

There are several alternatives to this approach, especially when you need 100% unattended workflows. Each alternative has advantages and of course drawbacks. 1) ConTeXt fills this gap nicely, but requires a very knowledgable programmer. 2) Many times users write some perl or python scripts that reads the database contents and produces some kind of output, perhaps LaTeX code that must be run with PdfLaTeX. This is a fast approach, but tends to get very hackerish after some time. 3) There is a standardized way of transforming XML to PDF called XSL-FO. This w3c standard has the big advantage that many tools exist to help the user in the task of publishing. But XSL-FO is very limited in its capabilities to produce reasonable documents.

A common demand in high volume output is to optimize page usage. As an example: imagine you have six products in a group but a page only fits five. The software system should be able to re-arrange the products and change a few parameters (image size, text size, text length), so that all six products fit on the same page and thus a whole page saved. The aforementioned systems are either very demanding on the programming side or just not capable of optimizations like these.

The product of our company is filling in this gap. It provides a way to transform XML (and thus any data) to PDF. It has a specialized input language designed for the purpose of laying out elements on a page, and it has all functionality of a modern programming language (variables, loops, if-then-else switches). It can put text and graphical elements on a virtual page that is used for any kind of layout optimization. These virtual pages can be removed and re-typeset with different parameters and only the 'best' page will make it to the PDF. As there is no control language for this kind of application yet, the system is inspired by the standards HTML (table layout), XPath (accessing the data and running specialized functions) and XSLT (accessing document nodes, programming constructs).

The software (called 'speedata Publisher') is written in Lua and makes heavy use of the LuaTeX engine. We use TeX to break paragraphs into lines, arrange the programmatically created boxes and glue for layout of complex tables and to write clean PDF. The publisher is open source software (AGPL) and runs under the three major operating systems (Linux, Windows, Mac OS X). The documentation is mostly still in German, although we are currently translating the documentation into english.

Patrick Gundlach