The LaTeXML Daemon: A LaTeX Entrance to the Semantic Web

Deyan Ginev, Heinrich Stamerjohanns, Michael Kohlhase

Computer Science, Jacobs University Bremen
<first initial>.<last name>@jacobs-university.de

Abstract

The language of the TeX/LaTeX typesetting system has become all-pervasive in scientific publications and has proven its utility, convenience and expressivity in its three-decade history. With the advent of the Web 2.0 paradigm, it has also become the primary choice of various technical and scientific social platforms, most prominently online encyclopedias and question-answer forums. The standardization of MathML and OpenMath and the adoption of the former in HTML5, have opened the floodgates for scientific content native to the browser.

The efforts of bringing LaTeX to the web are numerous and have varied in number and approach. Classical scenarios provide hooks to either LaTeX itself or a LaTeX daemon(e.g. [Uni]), incorporating formulas as PDF or PNG, while newer applications pursue a fully native output of XHTML+MathML (see [Sta+09] for an overview). The LaTeXML[Mil] system, and particularly its enhanced branch maintained for the arXMLiv[Sta+10] project, belongs to the second category. Moreover, it takes the effort one step further, being able to create XHTML+MathML+RDFa.

Building on the idea of a "LaTeX daemon", we introduce our own daemon version of LaTeXML, which addresses the problems of *efficient*, *scalable* and *on-the-fly* processing. This enhancement greatly reduces the overhead not only in processing of large-scale corpora, but also when employed as a conversion backend for web services using LaTeX as a frontend language. A state of maturity has been proven by converting one and a half million abstracts from the ZentralblattMATH[Zbl] database, coupled with a stable performance in various installations of the Planetary[Koh+11] system. Thus, we have a starting point from which to approach the Semantic Web.

Next, we equip our system with the capabilities of operating as a web service independent of a file system and of recognizing resources via web-accessible URIs. Furthermore, we add native support for user-embedded metadata, independent of possible vocabularies. When these features are employed in unity, our system acts as a capable conversion backened for web-based authoring systems, scalable to exporting user-defined metadata for add-on semantic services.

An emphasis has been placed on *flexibility*, *versatility* and *ease of use*, in both the setup and the deployment phases. A large range of customization options and a pair of intuitive server-client binaries with detailed documentation, enable out-of-the-box use for a wide range of applications. The system implementation is

based entirely on open web standards and has the full expressivity of the original TEX engine. Additionally, *correctness* and *robustness* are ensured respectively, via the powerful scoping system of LaTeXML and Perl's mastery in localizing both variables and processing flows. The daemon communication is based on sockets, allowing an easy coupling with both local and internet services.

LaTeXML is Public Domain software, and the daemon remains consistently under that license. Currently the software is hosted on the arXMLiv branch of the LaTeXML repository[Gin]. An ongoing collaboration with Bruce Miller will soon result in merging the functionality with the trunk of LaTeXML. Also, a demo page[Arx] has been set up, in order to showcase the features claimed.

Our future work will address reducing the memory and processing footprints in the short term, plus ensuring ironclad security when deployed on the web, in the mid-term. We are simultaneously developing a number of custom libraries that further build on the capabilities discussed herein, trying to fully realize the Semantic Web potential of the framework.

References

[Arx] *arXMLiv: Showcase Demo Page*. URL: http://trac.kwarc.info/arXMLiv/wiki/Demo (visited on 03/03/2011).

[Gin] Deyan Ginev. *LaTeXML: A LaTeX to XML Converter, arXMLiv branch*. URL: https://svn.mathweb.org/repos/LaTeXML/branches/arXMLiv (visited on 03/03/2011).

[Koh+11] Michael Kohlhase et al. "The Planetary System: Web 3.0 & Active Documents for STEM". submitted to the Elsevier Executable Paper Challenge. 2011. URL: https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf.

[Mil] Bruce Miller. *LaTeXML: A LaTeX to XML Converter*. URL: http://dlmf.nist.gov/LaTeXML/ (visited on 03/03/2011).

[Sta+09] Heinrich Stamerjohanns et al. "MathML-aware article conversion from LaTeX, A comparison study". In: *Towards Digital Mathematics Library, DML 2009 workshop*. Ed. by Petr Sojka. Masaryk University, Brno, 2009, pp. 109–120. URL: http://kwarc.info/kohlhase/submit/dml09.pdf.

[Sta+10] Heinrich Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. Ed. by Serge Autexier, Petr Sojka, and Masakazu Suzuki, pp. 299–307. URL: http://kwarc.info/kohlhase/papers/mcs10.pdf.

[Uni] Open University. *MathTran: A TEX to Image Converter Web Service*. seen May 2010. URL: http://www.mathtran.org (visited on 03/03/2011).

[Zbl] *Zentralblatt MATH*. URL: http://www.zentralblatt-math.org (visited on 03/03/2011).