A Semantics-Aware LaTeX-to-Office Converter

Lukas Kohlhase and Michael Kohlhase <https://latin.omdoc.org>
Jacobs University Bremen

We present a LaTeX-to-Office conversion plugin for LaTeXML that can bridge the divide between publication practices in the theoretical disciplines (LaTeX) and the applied ones (predominantly Office). The advantage of this plugin over other converters is that LaTeXML conserves enough of the document- and formula structure, that the transformed structures can be edited and processed further.

Problem

- Technical documents from the STEM fields augment the text with structured objects – images, mathematical/chemical formulae, diagrams, and tables – that carry essential parts of the information.
- There are two camps with different techniques for authoring documents.
  - theoretical disciplines (Mathematics, Physics, and CS) prefer LaTeX,
  - applied ones (e.g. Life Sciences, Chemistry, Engineering) use Office Suites almost exclusively.
- Transforming between these two document formatting approaches is non-trivial:
  - TeX/LaTeX uses in-document macros to "program" documents, empowering authors to automate document aspects (LaTeX packages).
  - Office suites rely on document styles that adapt visual parameters.

  each camp deems their approach vastly superior and the other's insufferable.
- Problems in trans-paradigm collaboration.

State of the Art & Approach

Three ways of transforming TeX/LaTeX to Office documents.
(converse direction different methods)

1. **copy/paste or import from PDF** There are two problems with this route:
   - mathematical formulae are not preserved

copy from PDF	paste (libreoffice)
$h_{\mu_\varphi}(f) + \int_X \varphi d\mu_\varphi = \sup_{\mathcal{M}(f,X)} \{h_\mu(f) + \int_X \varphi d\mu\},$	$h_{\mu_\varphi}(f)+ \square\ \phi d\mu_\phi = \sup\{h_\mu(f)+ \square\ \phi d\mu\},$

Copy & Paste in Word Processors

   - even if the result looks OK the results have lost their links (e.g. for citations/references or label/ref), or become difficult to edit, because they do not conform to the styling system of the word processor.

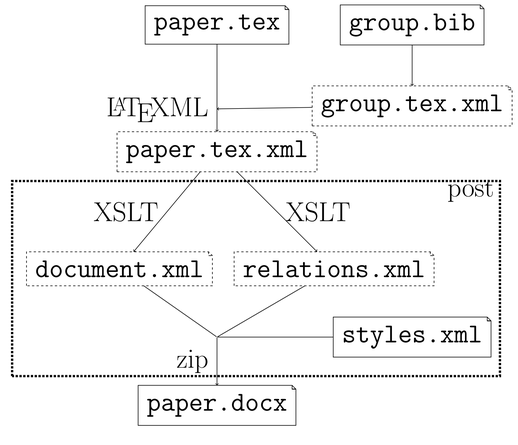
   Fundamental Problem: Process only converts the appearance of the document (meaning is lost)
2. **Parsing LaTeXsources, generating Office format**, e.g. `latex2rtf` (custom TeX parser, incomplete) or `TeX4HT` (use TeX, seed DVI, parse that, needs bindings).
3. **Our approach: Convert to XML, transform this**: LaTeXML converts to XML conserving author-supplied semantics (given LaTeXML bindings)

The Office Formats

WML (MS Office) and ODT (Open/LibreOffice) follow the same architectural paradigm: zip-packaged directories of XML files that contain document content, metadata, and styling.
Math as external objects in StarOffice/MathML in ODT and as proprietary XML format in WML, e.g. for 1.5×10^7:

```
<omml:oMath>
  <r><t>1.5  </t></r>
  <sSup>
    <e><r><t>10</t></r></e>
    <sup><r><t>7</t></r></sup>
  </sSup>
  1.5\times 10^{7}
</omml:oMath>
```

Transformation



The user does not see all these transformation, generation, and packaging steps: given a LaTeX paper, all she has to do is type

```
latexmlc paper.tex --destination=paper.docx
```

`destination=paper.odt` gives ODT

Result: Converted Formula in MS Word

$$h_{\mu_\varphi}(f) + \int_X \varphi d\mu_\varphi = \sup_{M(f,X)} \{h_\mu(f) + \int_X \varphi d\mu\},$$

Limitations

1. we cannot currently generate the text-based input format (StarMath or the WML TeX variant) (further editing difficult)
2. conversion of citations and references into "semantic" formats partial

Future Work

1. For ODF formulae, use of the TeXMaths plugin for Libreoffice, which uses LaTeX instead of StarMath for user input of formulae z
2. develop an "office package" for LaTeXand a corresponding LaTeXML binding, which allows the direct markup/transformation of higher-level structures.
3. extend the transformation to carry over even more semantics from the format into semantically extended office formats like CPoint or CWord

Availability

Public Domain on Github:
`https://github.com/KWARC/LaTeXML-Plugin-Doc`