

XML to paper publishing with manual intervention

Oleg Parashchenko

bitplant.de GmbH

XML: Aplicações e Tecnologias Associadas, 2010

Outline

- 1 Motivation
- 2 Generic Workflow
 - Where To Edit
 - Further requirements
 - Scheme
- 3 Sample System
 - Technical Details
 - Example
 - Practical Experience
- 4 Further Work
- 5 Summary

Ideal World

- Unlimited space on HDD
- No errors in software
- ...
- XML→PDF. No problems, just use XSL-FO.

Failure: Page Breaks

1 Lorem ipsum dolor sit amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

2 Ut wisi enim ad minim veniam

1

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, At accusam aliquyam diam diam dolore dolores duo eirmod eos erat, et nonummy sed tempor et et invidunt justo labore Stet clita ea et gubergren, kasd magna no rebum. sanctus sea sed takimata ut vero voluptua. est Lorem ipsum dolor sit amet.

2

Failure: Hieroglyphs

How to Say Hello in Different Languages

- English: Hi
- French: Salut
- German: Hallo
- Russian: Привет
- Chinese: □ □

IV

Main Contribution

If you can't fix it, feature it.

Layout correction
is a step in a workflow
can not be eliminated

Related Work

Oops...?

Outline

- 1 Motivation
- 2 **Generic Workflow**
 - **Where To Edit**
 - Further requirements
 - Scheme
- 3 Sample System
 - Technical Details
 - Example
 - Practical Experience
- 4 Further Work
- 5 Summary

Not in PDF

<p>1 Lorem ipsum dolor sit amet</p> <p>Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.</p> <p>Duis autem vel eum irure dolor in hendrerit in vulpate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis ut vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.</p> <p>2 Ut wisi enim ad minim veniam</p>	<p>Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum irure dolor in hendrerit in vulpate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis ut vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.</p> <p>Non hinc tempor erat sedita sedis idkifed optio congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.</p> <p>Duis autem vel eum irure dolor in hendrerit in vulpate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.</p> <p>At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy nibh euismod tincidunt ut laoreet dolore magna aliquam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, At necrum aliquyam etiam diam dolores etiam ipsum etiam et, ut wazirny sed tempus et ei hendrerit justo labore Stet clita ea et gubergren, kasd magna no rebum. sanctus sea sed takimata ut vero voluptua. est Lorem ipsum dolor sit amet.</p>
<p>1</p>	<p>2</p>

Required: high-level control

Not in XML

How to Say Hello in Different Languages

- English: Hi
- French: Salut
- German: Hallo
- Russian: Привет
- Chinese: 

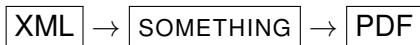
IV

- XML is “read only”
- Not enough control sequences
- Invalid XML*

*Example: an image instead of the text: not allowed

Required: low-level control

Between XML and PDF



SOMETHING:

- high-level control
- low-level control

Examples of SOMETHING:

- FrameMaker
- Microsoft Word
- T_EX
- XSL-FO

Outline

- 1 Motivation
- 2 **Generic Workflow**
 - Where To Edit
 - **Further requirements**
 - Scheme
- 3 Sample System
 - Technical Details
 - Example
 - Practical Experience
- 4 Further Work
- 5 Summary

User-friendly SOMETHING

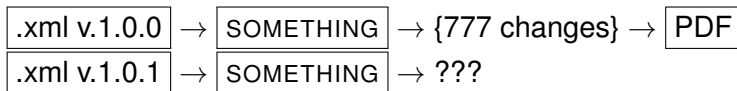
How to edit SOMETHING?

XSL-FO is not SOMETHING anymore.

```
...
<fo:list-block id="id2243582" space-before.optimum="1em" space-before.mi
nimum="0.8em" space-before.maximum="1.2em" space-after.optimum="1em" spa
ce-after.minimum="0.8em" space-after.maximum="1.2em" provisional-label-s
eparation="0.2em" provisional-distance-between-starts="1.0em"><fo:list-i
tem id="id2243483" space-before.optimum="1em" space-before.minimum="0.8e
m" space-before.maximum="1.2em"><fo:list-item-label end-indent="label-en
d()" "><fo:block>●</fo:block></fo:list-item-label><fo:list-item-body start
-indent="body-start()" "><fo:block>Duis autem vel eum iriure dolor</fo:blo
ck></fo:list-item-body></fo:list-item><fo:list-item id="id2243459" space
-before.optimum="1em" space-before.minimum="0.8em" space-before.maximum=
"1.2em"><fo:list-item-label end-indent="label-end()" "><fo:block>●</fo:blo
ck></fo:list-item-label><fo:list-item-body start-indent="body-start()" "><
fo:block>in hendrerit in vulputate velit esse molestie consequat</fo:blo
ck></fo:list-item-body></fo:list-item><fo:list-item id="id2243426" space
-before.optimum="1em" space-before.minimum="0.8em" space-before.maximum=
"1.2em"><fo:list-item-label end-indent="label-end()" "><fo:block>●</fo:blo
ck></fo:list-item-label><fo:list-item-body start-indent="body-start()" "><
fo:block>vel illum dolore eu feugiat nulla facilisis</fo:block></fo:list
...

```

Reuse Of Changes



Generic `diff` and `patch`.

“Beauty memory”

Support Tools

Anything where layout correction is a part of workflow:

- Edit SOMETHING
- Remember changes
- Build with changes

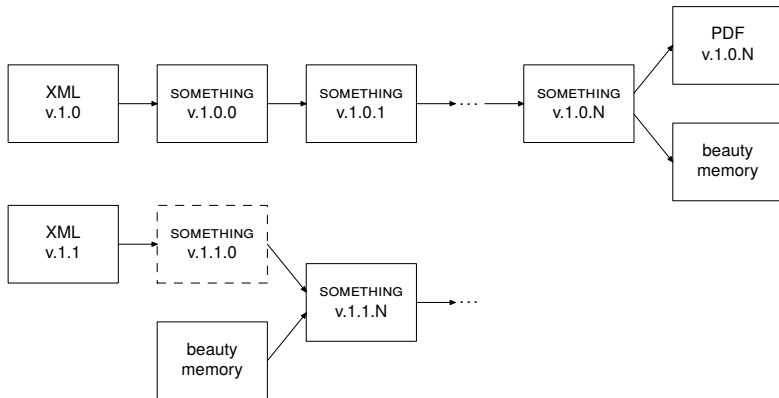
Outline

- 1 Motivation
- 2 **Generic Workflow**
 - Where To Edit
 - Further requirements
 - **Scheme**
- 3 Sample System
 - Technical Details
 - Example
 - Practical Experience
- 4 Further Work
- 5 Summary

Summary Of Requirements

- Layout correction is a part of workflow
- XML → **SOMETHING** → PDF
 - high-level control
 - low-level control
 - user-friendly
- **Beauty memory**
- Support tools

Scheme



Outline

- 1 Motivation
- 2 Generic Workflow
 - Where To Edit
 - Further requirements
 - Scheme
- 3 Sample System**
 - Technical Details**
 - Example
 - Practical Experience
- 4 Further Work
- 5 Summary

From Generic To Specific

SOMETHING

TEX

beauty memory

diff and patch

support tools

consodoc

T_EX As SOMETHING

`.xml` + `.xsl` → `.tex` + `.sty` → `.pdf`

`.xsl`: XML logical elements to T_EX logical elements

`.sty`: formatting

- ok: high-level control
- ok: low-level control
- user-friendly?

From XML to T_EX using T_EXML

XML to T_EX:

`.xml` + `.xsl` → `.texml` → `.tex`

The role of TeXML:

- generating correct `.tex` is hard
- automatic generation of human-friendly code layout

Beauty memory

`diff` and `patch`

User-friendly code means `diff/patch`-friendly code

Support tools

consodoc: **constructor of documentation**

<http://getfo.org/consodoc/>

On top of SCons build system

<http://scons.org/>

Complete rework required.

Outline

- 1 Motivation
- 2 Generic Workflow
 - Where To Edit
 - Further requirements
 - Scheme
- 3 **Sample System**
 - Technical Details
 - **Example**
 - Practical Experience
- 4 Further Work
- 5 Summary

Project

Files:

```
SConstruct  
in/  
  article.xml  
support/  
  democonv.xsl  
  democonv.sty
```

SConstruct:

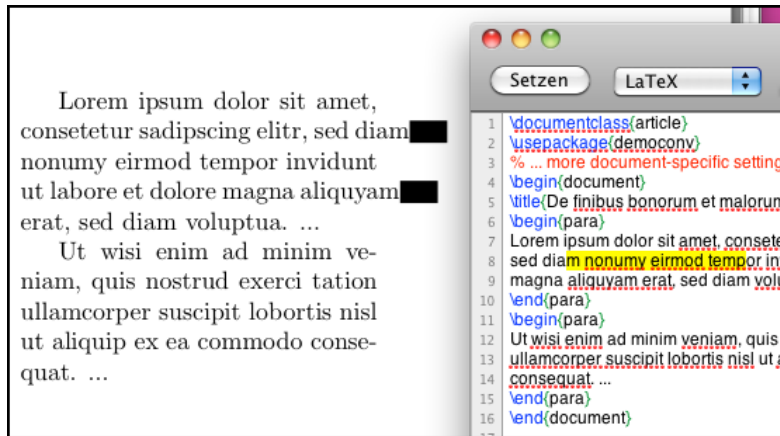
```
import Consodoc  
Consodoc.default_process(  
    in_file = 'in/article.xml',  
    in_xslt = 'support/democonv.xsl'  
)
```

article.xml, democonv.xsl, democonv.sty:
see the paper

Building the Project

```
$ cdoc
scons: Reading SConscript files ...
scons: done reading SConscript files.
scons: Building targets ...
xmllint --noent --xinclude --nonet -o tmp/article.xml in/article.xml
xsltproc -o tmp/article.texml support/democonv.xsl tmp/article.xml
texml tmp/article.texml tmp/article.tex.orig
no_patch(["tmp/article.tex"], ["tmp/article.tex.orig"])
run_pdflatex(["tmp/article.pdf"], ["tmp/article.tex"])
Overfull \hbox (12.68092pt too wide) in paragraph at lines 7--10
Overfull \hbox (8.90321pt too wide) in paragraph at lines 7--10
copy_file(["out/article.pdf"], ["tmp/article.pdf"])
scons: done building targets.
$
```

From PDF to .tex



The image shows a side-by-side comparison of a PDF document and its corresponding LaTeX source code. On the left, a PDF page contains two paragraphs of Latin placeholder text. On the right, a window titled 'LaTeX' displays the source code for this page. The code includes document class, package loading, title setting, and paragraph formatting commands. The PDF text is rendered in a serif font, and the source code uses color-coding for different command types.

PDF text:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. ...

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. ...

LaTeX source code:

```
1 \documentclass{article}
2 \usepackage{democonvy}
3 % ... more document-specific setting
4 \begin{document}
5 \title{De finibus bonorum et malorum}
6 \begin{para}
7 Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. ...
8 \end{para}
9 \begin{para}
10 Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. ...
11 \end{para}
12 \end{document}
```

Changing .tex

tmp/article.tex:

```
sed\break diam nonumy
```

```
$ cdoc
```

```
scons: Reading SConscript files ...
```

```
*** TeX code in 'tmp/article.tex' was changed.
```

```
    Consider generating a patch.
```

```
scons: done reading SConscript files.
```

```
scons: Building targets ...
```

```
run_pdflatex(["tmp/article.pdf"], ["tmp/article.tex"])
```

```
copy_file(["out/article.pdf"], ["tmp/article.pdf"])
```

```
scons: done building targets.
```

```
$
```

Remembering the Changes

```
$ cdoc patch
scons: Reading SConscript files ...
scons: done reading SConscript files.
scons: Building targets ...
generate_patch(["in/article.patch"], [])
scons: done building targets.
$
```

Beauty Memory

in/article.patch

```
--- tmp/article.tex.orig      2010-05-...
+++ tmp/article.tex          2010-05-09 04:38...
@@ -5,7 +5,7 @@
 \title{De finibus bonorum et malorum}
 \begin{para}
 Lorem ipsum dolor sit amet, consetetur ...
-sed diam nonumy eirmod tempor invidunt ...
+sed\break diam nonumy eirmod tempor inv...
 magna aliquyam erat, sed diam voluptua....
 \end{para}
 \begin{para}
```

Version 1.1

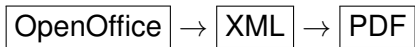
in/article.xml: a paragraph added

```
$ cdoc
scons: Reading SConscript files ...
scons: done reading SConscript files.
scons: Building targets ...
xmllint --noent --xinclude --nonet -o tmp/article.xml in/article.xml
xsltproc -o tmp/article.texml support/democonv.xsl tmp/article.xml
texml tmp/article.texml tmp/article.tex.orig
apply_patch(["tmp/article.tex"], ["tmp/article.tex.orig",
                                "in/article.patch"])
patching file tmp/article.tex
run_pdflatex(["tmp/article.pdf"], ["tmp/article.tex"])
copy_file(["out/article.pdf"], ["tmp/article.pdf"])
scons: done building targets.
$
```


Outline

- 1 Motivation
- 2 Generic Workflow
 - Where To Edit
 - Further requirements
 - Scheme
- 3 Sample System**
 - Technical Details
 - Example
 - Practical Experience**
- 4 Further Work
- 5 Summary

Writing Documentation



Small incremental changes

Ideal

Industrial Technical Documentation — 1

Book: \approx 450 pages

\approx 60 changes:

\approx 40 bad stylesheets

\approx 15 formatting instructions in XML for other application

\approx 5 essential

Note: tuning stylesheets would not help for \approx 20 places.

Industrial Technical Documentation — 2

Two month later

- content updated
- translation updated

Changes:

- ≈ 20 not applied
- 2 false applied

Everything due to bad stylesheets

But: essential changes are applied

Outline

- 1 Motivation
- 2 Generic Workflow
 - Where To Edit
 - Further requirements
 - Scheme
- 3 Sample System
 - Technical Details
 - Example
 - Practical Expierence
- 4 **Further Work**
- 5 Summary

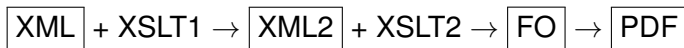
Microsoft Word As SOMETHING

ok	high-level control
ok	low-level control
ok	user-friendly
?	diff and patch

XSL-FO As SOMETHING

ok	high-level control
ok	low-level control
no	user-friendly
ok	diff and patch

Friendly XSL-FO



XML2: in presentation-oriented DTD

XSLT2: formatting

XML2 as SOMETHING

by design	high-level control
fo-attributes	low-level control
by design	user-friendly
xml diff	diff and patch

Further Work

- Improved T_EX packages
- Sample projects
- Consodoc

Summary

- Main contribution: manual intervention is a part of workflow.
- Supporting contribution: sample workflow on top of T_EX.
- Works in practice.
- Further work: the T_EX part of the system.
- Possible work: Microsoft Word and XSL-FO.