# XML to paper publishing with manual intervention

Oleg Parashchenko

bitplant.de GmbH

XML: Aplicações e Tecnologias Associadas, 2010



# Outline

- Motivation
- Queneric Workflow
  - Where To Edit
  - Further requirements
  - Scheme
- Sample System
  - Technical Details
  - Example
  - Practical Expierence
- Further Work
- Summary



# Ideal World

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

### Motivation

Generic Workflow Sample System Further Work Summary

# Failure: Page Breaks

### 1 Lorem ipsum dolor sit amet

Lorem Ipsum dodor sit amet, conseteur sodipseing ellit, sed diam nenumy eirmot tempor inviduant ut labore et dodore magna aliquvam crat, sed diam voluptus. At vero ces et accussam et justo duo dolores et en rebum. Stet tita koed gubergern, no sea takimata sanctus est Lorem ipsum dobor sit amet. Lorem ipsum dolor sit amet, conseteur sedipseing elitir, sed diam nommy eirmod tempor inviduant ut labore et dolore magna aliquvam erat, sed diam voluptus. At vero ces et accussam et justo duo dolores et en rebum. Stet cilia kasel gubergern, no sea takimata sanctus est Lorem ipsum dobor sit amet. Lorem ipsum dobor dolores et carbon institution dolores et accussam et pisco duo dolores et dam voluptus. At vero ces et accussam et justo duo dolores et oa rebum. Stet cilia kasel gubergen, no sea takimata sanctus est Lorem ipsum dobor sit amet.

Duis autem vel cum iriure dolor in hendrerit in vulputate vulit esse molestie consequat, vel illum dolore eu feugiat mulla facilisis at vero erns et accumsan et iusto odio dignissim qui blandit prossent luptatum zzril delenit augue duis dolore te feuguit mulla facilisi. Lorem ipsum dolor sit aunet, consecteture adipiscing elit, sed diam nonumny nibh enismod tincidunt ut laoreet dolore manna aliunam erat volutust.

### 2 Ut wisi enim ad minim veniam

Ut wis einim ad minim vaniam, quis nostrud exerci tation ullumcorper suscipii bloortis niel ut aliquip ex oa commodo consequat. Duis autem vel cum iriure dolor in hendrerit in vulputate velit cese modestie consequat, vel illum delore en fengiat nulla faciliis at vero cres et accumsan et insto odio diguissim qui blandit praesent luptatum zezil delenit augue duis dolore te fengiat nulla facilisi.

Nam liber tempor cum soluta nobis ekifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dobr sit amet, consectetuer adipiscing elit, sed diam nonummy nibh enismod tineidunt ut haorest dolore magna aliquam erat volutpat. Ut wis einim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut alioniu ex en commodo consecutat.

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

At very ose at accusan ext justo due dobers et en rebunt. Ste et tils loss il gubergen, no van takinnata sauctus es to torem ipsum dober sit annet. Lorem ipsum dober sit annet, conseteur suspinging dirt, va di dam nommy eitrond tempor inviduale. Islabore et dobere magna aliquyam erat, sed diam vulqutina. At labore et dobere magna aliquyam erat, sed diam vulqutina. Ste tils lasse grave et dobere magna aliquyam erat, sed diam vulqutina. Ste tils labore et dobere magna aliquyam erat sed diam vulqutina. Ste sit annet. Lorem ipsum dobor si annet consecteur subject eitit, At accussam aliquyam diam diam dobore dobores due erimode est annet. Lorem in tils vilham diam diam dobore dobores due erimode est annet. Lorem in tils vilham diam diam dobore dobores due erimode esce erat, et nomuny sed tempor et et invibrant pisto lobores due citiz en et gulsegyen, lossi magna no rebum. sanctus sea sed takinata su tver ovdeptina. set Lorem in psim dobor si amet.

- 5

# Failure: Hieroglyphs

How to Say Hello in Different Lan-
guages
• English: Hi
• French: Salut
• German: Hallo
• Russian: Привет
• Chinese:

# 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

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



Summary

Where To Edit Further requirements Scheme

# Not in PDF

### 1 Lorem ipsum dolor sit amet

Lemm jum shile of a mark consolute subjecting dift, and dispose mark and bins subjects. At very not a stream of a subject met. and other one return. But of the help polytrpes, now a subject met. and the subject of the subject of the subject of the stream of the subject of these nearest of mark others, At very not a vector of plant for a case, and offen subject on the subject of plant subject of mark of the subject of the subject of the subject of plant subject on an offen subject of the subject of plant subject of plant subject surface on Lemm pour sides of met. Lemm perm shale at surface on Lemm pour sides of met. Lemm perm shale surface on the subject of the subject of the subject of surface of the subject of the subject of the subject of the distribution. At very not at excess of plant shall not asset to or viern. See disk shall pringers, no we sufficient search

Dis autem vel eun izine iskor in hendrerk in vulptante voll een nederle consepant, vel film dekere en fonglet zufal fecilies at vero en et enermen et inste olis dipprisien qui blandit possent huptatum zeril dekent sugue chis dekore te fenant gat mila facilie. Lecen ipsem obbe vit aust, concecteturalipiering ells, sed diem neurmeny tibb estemat directum en honest deken rangus adigume eure voltepat.

### 2 Ut wisi enim ad minim veniam

veniam

Ut wiel enim ad minim veniem, quis metruel exceel tation ullamcorper suscipit lebertie niel ut allepip ex na commodo consequat. Dais autem vel ema iriure dobre in hembrerit in velprinte velit nese mobatte consequat, vel libra dobre en facigite atali facilies sis nt vero cross et accumsan et insto celio dignissim qui blandit

facilid. Non liber tempor cum soluta nobie ektični option congue nikil imperdict doming id quod maxim placenat facer possim seems. Lorem ipoms dobe sit aned, consecteior selijeiong ell., soil diam normmy milh culmond timelmat ut lacoset defoce magna aliquam cent volutpot. Ut wis culm ad minim ression, upis nortent enervi tation utimerosper solegibi koberti nicil ut

aliquip et es commodo consequat.

Dais autem vel cum iriure doke in hendrerit in vulputate
velit esse molestic consegunt, vel illum dokee en fenzist mills

All was one at assume as figured on where et on return. See et this body plottegers, no wes hishinan assume not Lower inpure doler wit mark, Lower inpure doler wit mark, consectour market and the second of the control of the second of

# Required: high-level control

# Not in XML

How to Say Hello in Different Languages • English: Hi · French: Salut • German: Hallo • Russian: Привет Chinese: TN/

- 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

$$oxed{\mathsf{XML}} o oxed{\mathsf{SOMETHING}} o oxed{\mathsf{PDF}}$$

### SOMETHING:

- high-level control
- low-level control

# Examples of SOMETHING:

- FrameMaker
- Microsoft Word
- TFX
- XSL-FO

# Outline

- Motivation
- Generic Workflow
  - Where To Edit
  - Further requirements
  - Scheme
- Sample System
  - Technical Details
  - Example
  - Practical Expierence
- 4 Further Work
- 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="lem" 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="lem" 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="lem" space-before.minimum="0.8em" space-before.maximum= "1.2em"><fo:list-item-label end-indent="label-end()"><fo:block></fo:block> 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

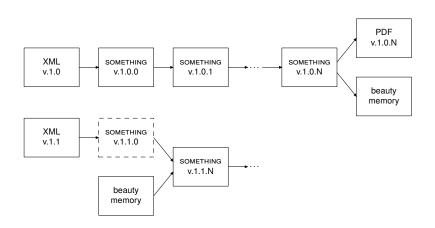
- Motivation
- Generic Workflow
  - Where To Edit
  - Further requirements
  - Scheme
- Sample System
  - Technical Details
  - Example
  - Practical Expierence
- 4 Further Work
- Summary



# 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

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



Technical Details Example Practical Expierence

# From Generic To Specific

SOMETHING TEX

beauty memory diff and patch

support tools consodoc

# TEX As SOMETHING

$$\boxed{.\mathsf{xml}} + .\mathsf{xsl} \rightarrow \boxed{.\mathsf{tex}} + .\mathsf{sty} \rightarrow \boxed{.\mathsf{pdf}}$$

- .xsl: XML logical elements to TFX logical elements
- .sty: formatting
  - ok: high-level control
  - ok: low-level control
  - user-friendly?

# From XML to TEX using TEXML

# XML to T<sub>E</sub>X:

$$\boxed{.\mathsf{xml}} + .\mathsf{xsl} \to \boxed{.\mathsf{texml}} \to \boxed{.\mathsf{tex}}$$

### The role of TeXML:

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

Technical Details Example Practical Expierence

# 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

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

# **Project**

### Files:

# SConstruct:

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

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

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

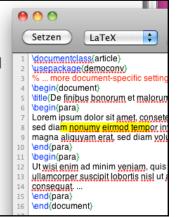
# **Building the Project**

```
$ 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

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. ...



# 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-...
                        2010-05-09 04:38...
+++ tmp/article.tex
@@ -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

# Outline

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

# Writing Documentation

$$\boxed{\mathsf{OpenOffice}} \to \boxed{\mathsf{XML}} \to \boxed{\mathsf{PDF}}$$

Small incremental changes

Ideal

# Industrial Technical Documentation — 1

Book:  $\approx$  450 pages

 $\approx$  60 changes:

 $\approx$  40 bad stylesheets

pprox 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:

- $\bullet \approx$  20 not applied
- 2 false applied

Everything due to bad stylesheets

But: essential changes are applied



# Outline

- Motivation
- Generic Workflow
  - Where To Edit
  - Further requirements
  - Scheme
- Sample System
  - Technical Details
  - Example
  - Practical Expierence
- 4 Further Work
- 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

$$oxed{XML}$$
 + XSLT1  $ightarrow$   $oxed{XML2}$  + XSLT2  $ightarrow$   $oxed{FO}$   $ightarrow$   $oxed{PDF}$ 

XML2: in presentation-oriented DTD

XSLT2: formatting

### XML2 as SOMETHING

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

# **Further Work**

- Improved T<sub>E</sub>X packages
- Sample projects
- Consodoc

# Summary

- Main contribution: manual intervention is a part of workflow
- Supporting contribution: sample workflow on top of TEX.
- Works in practice.
- Further work: the T<sub>E</sub>X part of the system.
- Possible work: Microsoft Word and XSL-FO.