Use cases and examination of XML technologies to process MS Word documents in a corporate environment

Colin Mackenzie

colin@mackenziesolutions.co.uk

XML Consultant



Why develop this solution?

• Learn in a hands-on way



- Many XML developers not using XSLT3 (some not properly utilising XSLT2)
- Used Xproc 1.0 +CX but limited subset of features and non-complex requirements
- Needed a project to increase my skills that I can then transfer to customers

Why choose Word documents?



- XML's popularity now focussed on documents
- MS Word used by most major corporate users of documents
- MS Word uses OOXML
- Processing Word often requires complex development (unpacking, no nested structure etc.)

Use case – quality and consistency of styles

Content

- Are all required sections present and correctly named?
- Styling
 - Latest branding
 - Professional looking result
 - Consistent through-out document
 - Numbering and referencing affects legal meaning

1. DEFINITIONS

In this Agreement the following definitions apply:-

'The Consultant' – **Colin Mackenzie** (or such other person as the Contractor sl Company)

'The Client' - means the person, firm or corporate body requiring the services of the 'The Assignment' - means the period during which the Consultant is engaged by Company to provide services to the Client as the same may be notified by the Co References to the singular include the plural and references to the masculine incl

2. THE CONTRACT

- (a) These Terms and Conditions constitute a contract between the Compa Consultant and they govern each and every Assignment undertaken by the
- (b) For the avoidance of doubt these Terms and Conditions shall not give rise Consultant.
- (c) For the avoidance of doubt these Terms and Conditions shall not give a principle relationship between the Company and the Contractor.
- (d) No variation or alteration of these Terms and Conditions shall be valid unleare the only Terms and Conditions upon which the Company will contract w
- (e) Save for the work specification of the Client including any industry m regulations, the Consultant shall have autonomy in respect of the technical

3. THE ASSIGNMENT

(i)

- (a) The Company will endeavour to obtain suitable assignments for the Consul
- (b) The Contractor acknowledges that it is in the nature of contract work that the Consultant and accordingly the Contractor and the Consultant agree that:
 - suitability of work for the Consultant shall be determined so
 - (ii) the Company shall incur no liability towards the Contr opportunities to work in the category specified in clause 3 a

Why do things g	o wron	g with Style	es?	Current List
 Lack of training 		aBbCcDc AaBbCcDc AaBbCcDc INormal INo Spac Heading 1	AaBbCcE	List Library
Define new Multilevel list Click le <u>v</u> el to modify:	? ×		Clear All	None 1) 1. a) 1.1. 1.1. i) 1.1.1. 1.1.1.
Article I. Heading 1 Section 1.01 Heading 2 (a) Heading 3 (i) Heading 4 1) Heading 5 a) Heading 6 i) Heading 7 a. Heading 8 i. Heading 9 Number format Enter formatting for number:	Current paragraph Image: Current paragraph Link level to style: Heading 1 Level to show in gallery: Level to show in gallery: Level 1 ListNum field list name:	File Home Image: Second state Image: Second state Image: Second state I	ember-vie hashtag-a hashtag-a Normal visually-hic No Spacin Heading 1	Image: Article I. Headi 1 Heading 1— Image: Article I. Headi 1.1 Heading 1— Image: Article I. Headi 1.1 Heading 2— Image: Article I. Headi 1.1 Heading 2— Image: Image: Article I. Headi 1.1 Heading 2— Image: Image: Image: Article I. Heading 1 Image: I
Article Eont Number style for this level: Include level number from: I, II, III, Include level number from: Position Aligned at: 0 cm Number alignment: Left Aligned at: 0 cm Text indent at: 0 cm Set for All Levels	Restart list after: Legal style numbering Follow number with: Tab character Add tab stop at: Or	A - V A -	Heading 2 Headin Headin Title Subtitle Subtitle Emphasis	1. 1.1 1.1.1 1.1.1 ① Efine List Level Define New Multilevel List Define New List Style

Typical solutions

- Custom templates
- Macros, VB, ribbons
- Training
- Commercial add-ins and products

• Tend to fail over time



So what about a standards-based solution?

- Allow knowledge workers to manage styles in the template
- Leave Word UI as out of the box
- Provide suggestions and feedback to users in a language they can understand
- Define the rules for style and content clearly

Word XML workflow



Bad styling leads to invalid or semantically incorrect XML Lack of validation leads to missing content Some XML content

- Word Tables -> HTML/CALS tables
- Footnotes, links, graphic references etc. -> XML mark-up
- Flat Headings (style) -> nested XML structure
- Flat Lists (style) -> nested lists
- Other paras -> Semantic XML elements

Semantic conversion

- Para in this style -> that semantic element
 - Group content between this para and another para in some other style
- Para in this style -> that metadata element
- Para with this auto numbering scheme -> that semantic nested element
- Drop fixed text
- Use text as clue to conversion etc.

Word XML + Schematron workflow



See 2017 "The application of Schematron schemas to word-processing documents" by Andrew Sales (Andrew Sales Digital Publishing Limited)

DEMO – Validation

1 2 3 ▼ testout.docx - Word		Colin Mackenzie 🎴 🖻 — 🗇 🗙
File Home Insert Design Layout References Mailings Review View Developer Help Q Tell me what you want to 40		A Sh _{žŠ}
A Cut Calibri (Body \not 10 \not A^* A^*) Aa + & E + E + \frac{1}{2} +	aBbCcD AaBbCcDc AaBbCcDc Heading 3 Heading 4 Heading 5	AaB AaBbCcc Title Subtitle Subtitle Subtitle Select → Sensitivity
Clipboard IS Font IS Paragraph IS Styles	1	🔂 Editing Sensitivity
	Error	Heading 2 must be immediatel
This is Head2		- · ·
This is head1	Heading 3 must b Heading 2 (para b	onds ago e immediately preceded by efore actually has style 'Heading1')
This is Head 3		C Reply C Resolve
This is a para with inline bold <i>italic</i> bolditalic underline strike e ^{sup} e _{sub} and inline style then back	Colin Mackenz	ie is bold ok?
New para style	Warning	Inline italic formatting is not 🔍
A para with a comment on last word!	Warning	Old character style 'Style2'
This is now a list	Error	Item in list must end in ','
Bullet1	Error	Second last item in list must end
Bullet2	Error	Last item in list must end in a for
• Bullet3		
Then a numeric list		
1. One numeric list item	Error	Second last item in list must end
2. Two numeric list item	Error	Last item in list must end in a ful
Normal sentence is ok		
1. This is clause level 1.		
1.1 This is clause level 2.		
The above were good examples of styles using numbering from multilevel lists.		
1. This is bad clause level 1.	Warning	Para seems to have manual 🛛 🔻
Page 1 of 2 1 word DB English (United States)		↓ 1 609

A sample rule

<sch:pattern> <sch:rule context="w:p"> <sch:let name="vNumStr" value="ms:getManualNumber(.)"/> <sch:let name="vSuggestedStyleNames" value="ms:getSuggestedStyleNames(.,\$vNumStr)"/> <sch:report test="\$vNumStr" id="ManualNumber1" role="warning">Para seems to have manual number '<sch:value-of select="\$vNumStr"/>': consider replacing using styles <sch:valueof select="\$vSuggestedStyleNames"/></sch:report> </sch:rule> </sch:pattern>

XProc Implementation



DEMO – Fixes following validation

	testout.docx - Word	ilii / ilii	Colin Mackenzie 🎦 🖻 — 🗇 🗙
File Home Insert Design Layout References Mailings Review View Developer Help Q Tellme v	what you want todo		A sh _{zs}
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	AaBbc 1.1a) AaB AaBbccbc AaBbCc	E AaBbCcD <i>AaBbCcDt</i> AaBbCcDt 2 Heading 3 Heading 4 Heading 5	ABB AaBbCcC Title Subtitle Subtitle Sensitivity Select Sensitivity
Clipboard IS Font IS Paragraph IS	Styles	Error	Item in list must end in ; Sensitivity A
One numeric list item Two numeric list item		Fix	Fix AddText-::: Add semi-colon
Normal sentence is ok		Error	Second last item in list must end
1. This is clause level 1.		FIX	Fix Add Text-, and. Add senii-
		Error	Last item in list must end in a full
		Fix	Fix AddText: Add full stop
The above were good examples of styles using numbering from	n multilevel lists.	Error	Second last item in list must end
1. This is bad clause level 1.		Fix	Fix AddText-; and: Add semi- 🔻
1.1 This is bad clause level 2.		Error	Last item in list must end in a full
The two paras above have been styled and therefore numbere	d (in this case using a numbered list)	Fix	Fix AddText: Add full stop
incorrectly. The clause 2 ever ended up with inline numbered t	text when the author tried to correct it.	Warning	Para seems to have manual 🔍
The two para below are correct.	Comments 🔹	× Fix	Fix RemoveManualNumber: 🔍
2. This is clause level 1.	Fix Friday	Fix	Fix ChangeStyle-Clause1: Change
2.1 This is clause level 2.	Fix RemoveManualNumber: Remove manual number 1.	Warning	Para seems to have manual 🔍
2.1a) Experimental Clause level 3 or 1.1a	💭 Reply 🦿 Resolve	Fix	Fix RemoveManualNumber: 🔻
MORE		Fix	Fix ChangeStyle-Clause2: Change
1. This is bad clause level 1.		Warning	Para seems to have manual 🛛 🔻
		Fix	Fix RemoveManualNumber:
		Fix	Fix ChangeStyle-Clause1: Change
Page 1 of 2 1 word 🛛 🛱 English (United Kingdom) 📸			III IIII III IIII IIII IIII IIII IIII IIII IIII IIIII IIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

A sample rule with a fix

```
<xsl:for-each select="$vSuggestedStyleNames">
  <cmqf:fix id="ChangeStyle-{.}">
     <cmqf:description>
         <cmqf:title>Change style to <xsl:value-of
select="."/></cmqf:title>
    </cmqf:description>
  </cmqf:fix>
</xsl:for-each>
    </xsl:for-each>
```

ng

</cmqf:fixes>

</sch:report> </sch:rule> </sch:pattern>

Lessons learnt

- Saxon PE/EE needed for xsl:evaluate
- Functions can only be successfully called dynamically if marked as visibility="public" (although function-available returns true no matter how it is set
- Named parameters (as opposed to XPaths) to dynamically called functions need to be specified

<xsl:evaluate xpath="my:function(., xpath , \$varname)" context= "." >

<xsl:with-param name= "varname" select="\$varname" />

</xsl:evaluate>

- Ensure that injection attacks are avoided
- XProc annoyances and debugging XProc3 please!
- OOXML annoyances

Conclusion

- There is no technical reason why tools like XProc, XSLT3 and Schematron cannot be used to solve corporate Word use cases
- Limitations will be around willingness of technical staff to learn and use the technologies
- While toolsets and libraries of functions like the ones I prototyped can reduce the technical burden, this may still be a challenge

Old dogs CAN do new tricks!



<xsl:sequence

select="let \$f := function-lookup(xs:QName(\$vFinalFunctionName),\$vCountArgs)
return if (exists(\$f)) then \$f(\$pOriginalElement, ., \$vFunctionArgs)
else ()"/>

Thanks!

Any questions?

Colin Mackenzie

colin@mackenziesolutions.co.uk

XML Consultant

