XML-Qualitätssicherung mit [parsX] Konventionsprüfungen im Rowohlt Verlag

rowohlt

rowohlt

Checkliste alt und neu

Prüfungsrichtlinien für XML-Dateien

Dokuchecker-Protokoll prüfen

- Siehe Dok-6
- Sind alle nötigen Joker ausreichend dokumentiert. Wurde für entfallene Joker die Dokumentation gelöscht?

Oxygen

- Sind Seitenzahl-PI vorhanden? Richtigkeit prüfen: 3 Stichproben im letzten Drittel
- Sind die Metadaten ausgefüllt und inhaltlich korrekt? Insbesondere <meta-isbn> darf nicht leer sein. Die ISBN muss der ursprünglichen ISBN (ISBN, unter der die XML-Instanz erstellt wurde) entsprechen.
- Sind harte Trennungen stehen geblieben?
- Stichprobenhaft prüfen, ob <versal>, <kapitaelchen> und <inline> konventionsgetreu getagged wurden. Interpunktionen dürfen nicht mitgetagged werden.
- *Mikrotypographie (Dok-12)*

– Dr.	Festwerte aus Punkt 6, Dok-12
– Mrs.	z. B.:
– Mr.	– «...
– Mr	– «–
– Mrs	– –,
– Maßeinheiten (cm, kg, ...)	– –»
– /	– ...»
– 000 bzw. ziffer	

- Das Tagging der Joker (einschub, infokasten, u-zwischen, inline) überprüfen. (Z. B. das Vorkommen von kursiv/fett hinterfragen.)
- stichprobenhaft: Wurden Sonderzeichen korrekt codiert?
- Sind die Anführungen vor «Initiale» korrekt erhalten geblieben?
- Nach www und http suchen, um <verweis>-Tagging zu überprüfen

epub / ADE

- Im Toc die Struktur überprüfen. Ist die Kapitelhierarchie korrekt?
- Sind die Toctitel korrekt (Abgleich mit XML-Auftrag)?
- Ist das Printinhaltsverzeichnis entfallen?
- Joker kontrollieren

Prüfungsrichtlinien für XML-Dateien

Dokuchecker-Protokoll prüfen

- Siehe Dokument 01_DokCheck Erlaeuterung
- Sind alle nötigen Joker ausreichend dokumentiert. Wurden entfallene Joker aus der Dokumentation gelöscht?

Oxygen

- Sind Seitenzahl-PI vorhanden? Richtigkeit prüfen: 3 Stichproben im letzten Drittel. Seitenzahl-PI darf nur in texttragenden Elementen stehen!
  - Suchlauf nach PI innerhalb von Wörtern, um Trennung ohne Divis im ePub zu vermeiden[1]
- Sind die Metadaten ausgefüllt und inhaltlich korrekt? Insbesondere <meta-isbn> darf nicht leer sein. Die ISBN muss der ursprünglichen ISBN (ISBN, unter der die XML-Instanz erstellt wurde) entsprechen. Die ISBN muss richtig gegliedert und mit Divisen versehen sein.
  - Sie alle <titelei_abs> richtig typisiert?
- Sind harte Trennungen stehen geblieben?
- Stichprobenhaft prüfen, ob <versal>, <kapitaelchen> und <inline> konventionsgetreu getagged wurden. Interpunktionen dürfen nicht mitgetagged werden.
- *Mikrotypographie (Dok-12)*

– Dr.	Festwerte aus Punkt 6, Dok-12
– Mrs.	z. B.:
– Mr.	– «...
– Mr	– «–
– Mrs	– –,
– Maßeinheiten (cm, kg, ...)	– –»
– /	– ...»
– 000 bzw. ziffer	

- Das Tagging der Joker (einschub, infokasten, u-zwischen, inline) überprüfen. (Z. B. das Vorkommen von kursiv/fett hinterfragen.)
- stichprobenhaft: Wurden Sonderzeichen korrekt codiert?
- Sind die Anführungen vor «Initiale» korrekt erhalten geblieben?
- Nach www und http suchen, um <verweis>-Tagging zu überprüfen

epub / ADE

- Im Toc die Struktur überprüfen. Ist die Kapitelhierarchie korrekt?
- Sind die Toctitel korrekt (Abgleich mit XML-Auftrag)?
- Ist das Printinhaltsverzeichnis entfallen?
- Joker kontrollieren

1 Suche:
(\p{L}+)(<\?(gwhisparsx) snr="\d+"\?>)(?=\p{L}+)
Ersetze durch: $2$1
Wichtig: Regulärer Ausdruck anwählen.

= Schematronprüfung vorhanden

Stand: 16.12.2016

rowohlt

Schematron-Prüfung und QuickFix

Kategorie:	Nr.	Regel/Prüfung	QuickFix
A	5	<meta_isbn> darf nicht leer sein.	
A	8	Suche leere Absätze.	
A	9	Suche <leerzeile> vor und nach <einschub>.	Lösche <leerzeile>.
A	10	Suche <leerzeile> vor und nach <einschub_vor>.	Lösche <leerzeile>.
A	11	Suche <leerzeile> vor und nach <u-zwischen>.	Lösche <leerzeile>.
A	12	Suche <leerzeile> vor und nach <einschub_innen>.	Lösche <leerzeile>.
A	13	Suche <leerzeile> vor und nach <infokasten>.	Lösche <leerzeile>.
A	17	Suche innerhalb des Attributs verweis-extern nach Fällen, wo nicht [http://]/ [https://] vorkommt.	Ergänze http://
A	32	Suche in Bildquellen nach png/gif als Bildformat	
A	34	Suche nach 	a) Entferne
			b) Ersetze durch WA
			c) erhalten
A	37	Folgt <meta_id></meta_id> auf <meta_projekt>?	Ergänze <meta_id></meta_id>
A	38	Wenn das erste Kindelement von <kapitel> kein Überschriftenelement ist, wird ein toctitel angemahnt	
A	39	Prüfen: Tragen alle <titelei_abs> das @typ-Attribut?	

Schematron-Prüfung ohne QuickFix

Kategorie:	Nr.	Regel/Prüfung	QuickFix
B	1	Kommen innerhalb von <versal> Interpunktionen vor?	Die Interpunktion von dem <versal>-Tag ausschließen.
B	2	Kommen innerhalb von <kapitaelchen> Interpunktionen vor?	Die Interpunktion von dem <kapitaelchen>-Tag ausschließen.
B	28	Suche nach </ziffer>[Achtelgeviert]<ziffer>.	Ersetze durch [Achtelgeviert].
B	31	Suche nach </versal>[Wortabstand]<versal>.	Ersetze durch [Wortabstand]
B	33	Suche nach <?parsx snr=""?> außerhalb von texttragenden Elementen	<?parsx snr=""?> verschieben
B	6	Sind alle vorkommenden Joker dokumentiert?	QF <jokereinsatz> ergänzen (Vorlage)
B	7	Sind alle dokumentierten Joker vorhanden?	QS Element in <jokereinsatz> löschen

B	14	Suche nach Fällen, in denen innerhalb eines Blockjokers (einschub/einschub_innen/ infokasten/ u-zwischen/einschub_vor) der erste Absatz mit <kursiv>/<bold> beginnt UND der letzte Absatz mit <kursiv>/<bold> endet.	
B	30	Suche nach	a) so lassen
		<kursiv><versal>	b) <ziffer>/<versal>/ <hoch>/<tief> entfernen
		<bold><versal>	
		<kursiv><ziffer>	
		<bold><ziffer>	
		<kursiv><hoch>	
		<bold><hoch>	
		<kursiv><tief>	
		<bold><tief>	
		innherhalb der Blockjoker (einschub/einschub_innen/ infokasten/u-zwischen/einschub_vor)	
B	15	Wenn in den Zeichenketten nach <initial> zuerst eine Abführung (») vor einer Anführung («) vorkommt, zeige die Fälle, in denen keine Anführung («) vor <initial> steht.	Ergänze die Anführung («) vor <initial>.
B	36	Suche nach <initial>[WA]	Ersetze durch <initial>
B	16	Attribute ausgenommen: Stehen [www] und [http] nicht innerhalb eines verweis-Tags?	

Keine Schematron-Prüfung

Kategorie:	Nr.	Regel/Prüfung	Quickfix
C	35	Suche nach nicht hexadezimalen Entities	Ersetze durch entsprechende hexadezimale Entity
C	3	Sind alle Zahlen von <ziffer> umgeben?	Zahlenfolge mit <ziffer>-Tag versehen
C	4	Suche nach Vorkommen von [Divis][Leerzeichen]. Ausgenommen sind Fälle, in denen auf [Divis][Leerzeichen] folgende Zeichenketten folgen: und/oder/bzw./beziehungsweise	Tilge das Leerzeichen.
C	18	Suche nach [Mr.][Leerzeichen].	Ersetze mit [Mr.][Sechstelgeviert].
C	19	Suche nach [Mrs.][Leerzeichen].	Ersetze mit [Mrs.][Sechstelgeviert].
C	20	Suche nach [Mr]Leerzeichen].	Ersetze mit [Mr][Sechstelgeviert].
C	21	Suche nach [Mrs][Leerzeichen].	Ersetze mit [Mrs][Sechstelgeviert].
C	22	Suche nach [Dr.][Leerzeichen].	Ersetze mit [Dr.][Sechstelgeviert].

Keine Schematron-Prüfung

		…›/ … ‹	…›/ … ‹
		…» /…«	…» / … «
		…!	…!
		…?	…?
		…,	…,
		…;	…;
		…:	…:
		…)	…)
		…]	…]
C	26	Suche nach Fällen, in denen nach diesen Zeichenketten kein gWA steht.	Ersetze jeweils mit
		(…	(…
		[…]	[…]
		‹…/ ›…	‹… / ›…
		«… / »…	«…/»… und gWA
C	27	Suche nach Fällen, in denen zwischen einer Zahl (bzw. <ziffer>) und der nachfolgenden abgekürzten Maßeinheit kein Sechstelgeviert steht.	Ersetze das Zeichen zwischen Zahl (bzw. <ziffer>) und der abgekürzten Maßeinheit mit einem Sechstelgeviert.
		Maßeinheiten:	
		€, $, %, ‰, mg, g, kg, t, mm, mm^2, mm^3, cm, cm^2, cm^3, m, m^2, m^3, km, km^2, ml, l, s, Std., EL, TL, Msp.	
C	29	Suche nach Fällen, in denen vor und/oder nach [/] ein Leerzeichen oder ein Zwölftel-/Achtel-/Sechstel- oder Viertelgeviert steht.	
C	38	Suche nach doppelten Leerzeichen	Tilge ein Leerzeichen
C	39	Suche nach: Viertel-/Sechstel-/Achtelgeviert/geschütztem Wortabstand + Leerzeichen **und** Wortabstand + Viertel-/Sechstel-/Achtelgeviert/geschütztem Wortabstand	Tilge Leerzeichen

XML-Qualitätssicherung – Konventionsprüfungen

Schematron – Möglichkeiten und Grenzen

Funktionsweise

- Schematron und DTD
- Aufbau einer Schematron-Regel mit SQF
- Einsatzmöglichkeiten

XML als String und als Baum

- Beispiel: "Wortabstand zwischen Versal": 2 Sichtweisen
- Umformulierung einer Anforderung

***Fazit: Was geht** (gut), **was geht nicht** (mit vertretbarem Aufwand)*

Schematron und DTD

- DTD: alles, was nicht erlaubt ist, ist verboten
- Schematron umgekehrt: was nicht explizit verboten ist (also geprüft wird), ist erlaubt (wird nicht gemeldet).
  Man muss also wissen, was geprüft wird und was nicht!

Aufbau einer Schematron-Regel mit SQF

```
<sch:rule context="br">
        <sch:report test="true()" role="warn" sqf:fix="br_löschen …">
                ▌GvH▌ [SQF] Zeilenumbruch [br/] gefunden!
        </sch:report>

        <sqf:fix id="br_löschen">
                <sqf:description>
                        <sqf:title>Element [<sch:name/>] löschen?</sqf:title>
                </sqf:description>
                <sqf:delete/>
                …
        </sqf:fix>

</sch:rule>
```

Komponenten einer Regel mit SQF

- rule context="br" Bezug auf einen XML-Knoten (Element, Text, PI)
- report test="true()" Bedingung, die dieser Knoten (nicht) erfüllt
- **Meldung** (frei formulierbar, auch mit Link zur Online-Doku)
- sqf:description Überschrift und Erläuterung frei formulierbar
- sqf:delete Aktion über XML-Knoten (auch außerhalb des Kontexts)

Einsatzmöglichkeiten

- Meldung und FIX im Prinzip für jede einzelne Stelle
- Focus auf XML-Knoten (vor allem Elemente, deren Inhalt und Attribute), NICHT Textstellen.
- Anzeige eines Fehlers nur beim ersten Vorkommen und generelle Behebung per SQF ist möglich, aber aufwändig

XML als String und als Baum

Beispiel als String:

<abs>Die <versal>QUALITÄTEN</versal> <versal>EINER</versal> <versal>FÜHRUNGSKRAFT</versal> befinden…</abs>

Beispiel korrigiert als String:

<abs>Die <versal>QUALITÄTEN EINER FÜHRUNGSKRAFT</versal> befinden…</abs>

Anforderung: Suche nach "</versal>[Wortabstand]<versal>". Ersetze durch "[Wortabstand]".

Beispiel als Baum:

- **abs**
  - "Die anzustrebenden "
  - **versal**
    - "QUALITÄTEN"
  - " "
  - **versal**
    - "EINER"
  - " "
  - **versal**
    - "FÜHRUNGSKRAFT"
  - " befinden" ...

Umformulierung einer Anforderung

Anforderung: " Suche nach Fällen, in denen innerhalb eines Blockjokers (einschub/einschub_innen/ infokasten/u-zwischen/einschub_vor) der erste Absatz mit <kursiv>/<fett> beginnt UND der letzte Absatz mit <kursiv>/<fett> endet."

Beispiel:

<einschub><abs><kursiv>Andererseits</kursiv> haben … zur <kursiv>Kunst der indirekten Kommunikation</kursiv>.</abs></einschub>

- Wird nicht gefunden, da Punkt nicht kursiv!

Umsetzung:

▌GvH▌ [einschub] zu über 90% in [kursiv]!

Fazit:

Stärken von Schematron

- Prüfen von Elementen, Attributen und Inhalten mit frei formulierbaren Meldungen und Anweisungen
- Finden und Korrektur einzelner (oder Gruppen von) Stellen nach Sichtprüfung

Problematisch

- Operationen innerhalb ungegliederter Textabschnitte (mixed content)
  Typisch: Mikrotypo

- Änderungen, die global an vielen Stellen ausgeführt werden sollen.

Alternativen

- Suche/Ersetze mit Regular-Expressions
  - ACHTUNG Markup!
  - Tool in Entwicklung