

Workshop

XML-Print

Ein Werkzeug zum Satz beliebiger XML-Dokumente

Prof. Dr. Marc W. Küster Lukas Georgieff Martin Sievers

16. Dezember 2013

1. Einleitung

XML ist unbestreitbar zu *dem* Dateiformat in den „Digital Humanities“ geworden. Insbesondere der TEI-Standard¹ wird von vielen neu entwickelten Werkzeugen als Ausgabe- und Austauschformat unterstützt. Auch ein XML-basierter Werkzeugkasten wie das TextGridLab², das die verschiedenen Entwicklungsstufen geisteswissenschaftlicher Projekte unterstützt, hat zur Verankerung von XML in der Anwendergemeinschaft maßgeblich beigetragen.

Bei aller Digitalisierung ist bei den Anwendern jedoch der Wunsch erhalten geblieben, ihre Texte bzw. Ergebnisse auch in eine gedruckte (analoge) Form zu überführen. Eine Umfrage der BBAW hat jüngst bestätigt, dass TextGrid-Anwender eine Printkomponente besonderes vermissen. Diese wird im Rahmen eines eigenständigen, DFG-geförderten Projekts *XML-Print* entwickelt und steht kurz vor der Fertigstellung. Unabhängig vom TextGridLab ist die Open-Source-Software allerdings auch eigenständig nutzbar und somit für einen großen Nutzerkreis von Bedeutung.

Nach einigen Live-Präsentationen und kleineren Workshops möchten wir auf der DHD 2014 nicht nur einen Einblick in die aktuelle Vorabfassung von XML-Print 2.0 geben, sondern allen interessierten Wissenschaftlern die Chance bieten, sich ausgehend von einer XML-Datei in einzelnen Schritten systematisch einer druckfertigen PDF-Ausgabe zu nähern.

2. Funktionsweise von XML-Print

2.1 Allgemein

XML-Print ermöglicht die Überführung beliebiger XML-Dateien in eine PDF-Datei.³ Grundlage dafür ist ein erweitertes Dateiformat XSL-FO+, das auf dem etablierten Standard XSL-FO⁴ beruht und diesen um bis dato fehlende typographische Elemente wie Mehrspaltigkeit und

¹ Vgl. Burnard und Bauman 2007.

² Mehr unter *TextGrid: Digital edieren – forschen – archivieren* 2013.

³ Andere Ausgabeformate können bei Bedarf über eigene Filter hinzugefügt werden.

⁴ Siehe dazu die Definitionen der *Extensible Stylesheet Language (XSL)* 2006, Version 1.1 sowie 2012, Version 2.0.

den Satz von Apparaten ergänzt. Die im Rahmen des Projekts entwickelte neuartige *Satzengine* interpretiert die XSL-FO+-Datei und erzeugt daraus eine druckfertige Ausgabedatei.

Obwohl man diesen Prozess komplett über die Kommandozeile steuern kann, ist es ein wesentliches Ziel von XML-Print, die Zuweisung von Layoutinformationen zu XML-Elementen über eine graphische Benutzeroberfläche komfortabel durchführen zu können. Der dazu entwickelte *Stileditor* erzeugt aus XML-Quelldatei(en) und im Hintergrund generierten XSLT-Stylesheets die für die Satzengine benötigte Eingabedatei (vgl. dazu auch Abbildung 1).

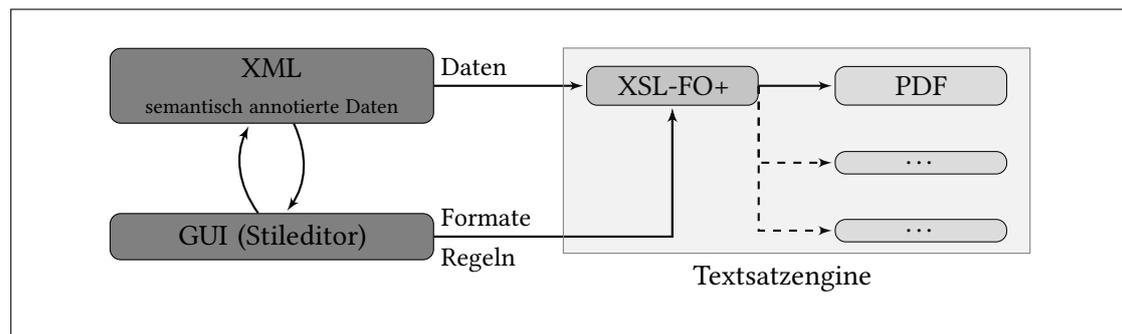


Abbildung 1: Der Ablauf der Erstellung einer Druckausgabe mit XML-Print. Mit Hilfe der graphischen Oberfläche werden XML-Elementen Layoutinformationen zugewiesen und daraus das benötigte Zwischenformat generiert. Dieses wird von der Satzengine interpretiert und in eine PDF-Ausgabe überführt.

2.2 Stileditor

Die graphische Benutzeroberfläche innerhalb von XML-Print, der *Stileditor*, verbindet die Quelldaten im XML-Format über XSLT-Templates mit Layoutinformationen und macht daraus eine Eingabedatei für die Satzengine im XSL-FO+-Format.

Der Anwender erstellt zunächst *Formate*, die das gewünschte Erscheinungsbild festlegen. Die verschiedenen Möglichkeiten sind zur besseren Übersichtlichkeit in Kategorien unterteilt (vgl. Abbildung 2). Anschließend definiert der Nutzer *Zuweisungen* zwischen einer Gruppe von XML-Elementen auf der einen und einem der Formate auf der anderen Seite. Die Auswahl der

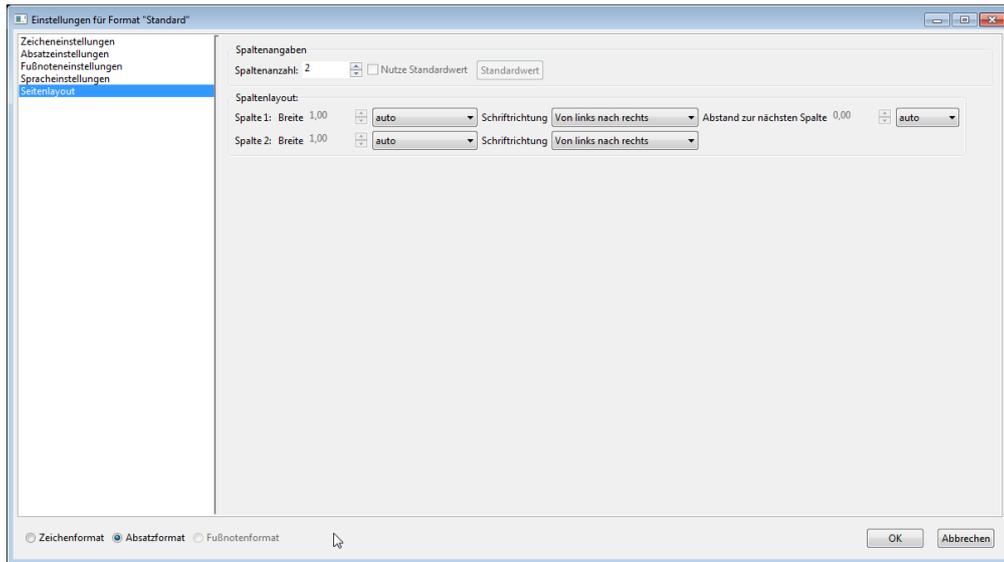


Abbildung 2: Formate beschreiben das gewünschte Aussehen bestimmter XML-Elemente. Zur besseren Übersicht sind die einzelnen Möglichkeiten in Kategorien unterteilt (linke Seite).

XML-Elemente geschieht dabei graphisch über ihre Position im XML-Baum oder alternativ über einen XPath⁵-Ausdruck (siehe Abbildung 3).

Über diesen Grundmechanismus lassen sich bereits sehr viele Layoutspezifikationen vornehmen. Für globale typographische Einstellungen wie Seitenformate oder die Definition von Feldern und Apparaten existieren zusätzlich jeweils eigene Dialoge (vgl. Abbildung 4).

2.3 Satzengine

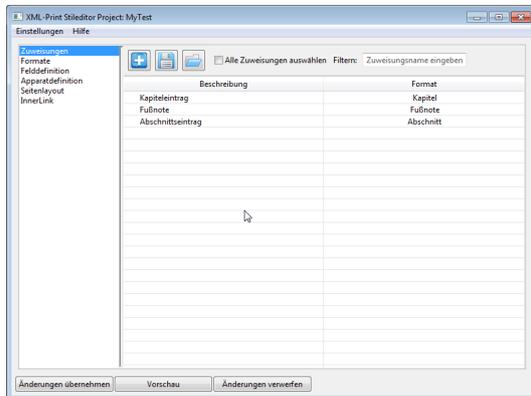
Es herrscht kein grundsätzlicher Mangel an Textsatzwerkzeugen mit hoher typographischer Ausgabequalität. Neben dem kommerziellen InDesign⁶ ermöglichen die Open-Source-Werkzeuge \TeX ⁷ sowie die Satzkomponente des TUSTEP⁸-Systems seit über dreißig Jahren hochwertigen Textsatz – allerdings erst, nachdem man sich in die Eigenheiten und -arten der Software einge- arbeitet hat. Zudem sind beide Programme sehr Texteditor-lastig und damit für viele heutige

⁵ Die XPath-Syntax ermöglicht den Zugriff auf beliebige Teile eines XML-Dokuments. Zur Definition siehe *XML Path Language (XPath)* 2010.

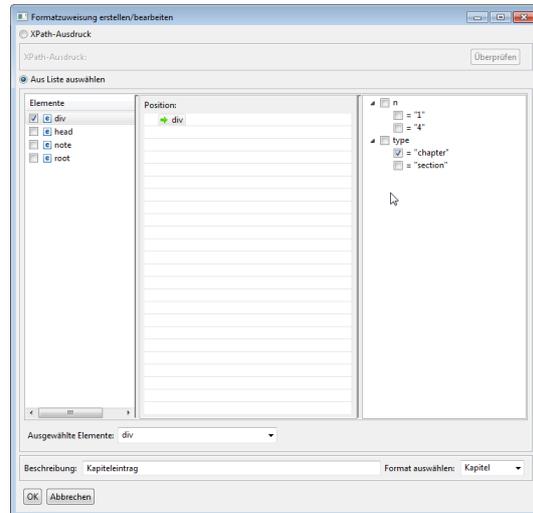
⁶ Siehe <http://www.adobe.com/products/indesign.html>.

⁷ Zur Entstehungsgeschichte von \TeX siehe z. B. <http://tug.org/whatis.html>.

⁸ Siehe <http://www.tustep.uni-tuebingen.de/index.html>.

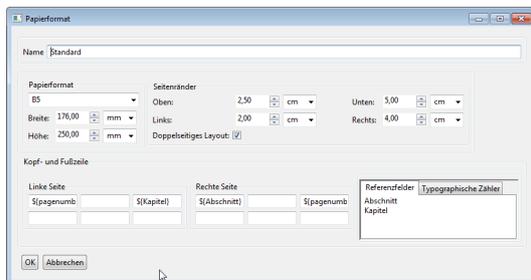


(a) Die Reihenfolge der Liste aller definierten Zuweisungen (rechte Seite) legt gleichzeitig die Priorität fest.

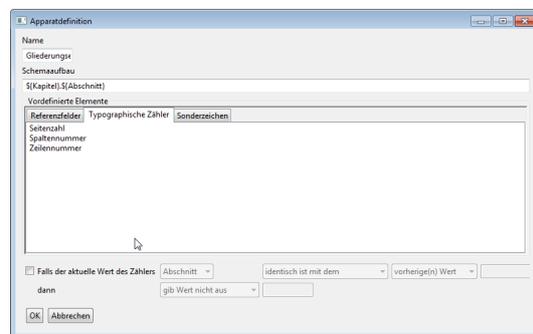


(b) Die Auswahl der XML-Elemente erfolgt wahlweise graphisch über ihre Position im XML-Baum oder über einen beliebigen XPath-Ausdruck.

Abbildung 3: Eine Zuweisung bestimmt für eine Gruppe von XML-Elementen, welches Format verwendet werden soll. Aus der Gesamtübersichts (a) gelangt man durch Auswahl einer Zuweisung zu den Einstellungen (b).



(a) Neben den Seiten- und Randmaßen werden für das Papierformat auch die Inhalte der Kopf- und Fußzeile definiert. Diese können in zwölf Bereichen aus benutzerdefinierten Feldern, typographischen Zählern und Zeichenketten beliebig gebildet werden.



(b) Für Apparateinträge können aus typographischen Zählern, benutzerdefinierten Feldern und Zeichenketten beliebige Referenzschemata gebildet werden. Zusätzliche Ausnahmen erlauben die selektive Ausgabe von Bestandteilen.

Abbildung 4: Globale Einstellungen in XML-Print. „Seitenlayout“ (a) und „Apparatdefinition“ (b) sind Kategorien der Hauptansicht (vgl. Abbildung 3a, linke Seite).

Anwender, insbesondere für Anfänger, die gewohnt sind, mit graphischen Systemen zu arbeiten, eher abschreckend.

Auch aus der Sicht der Informatik sind beide Programme – aus nachvollziehbaren Gründen – nicht auf der Höhe der Zeit, so dass neben der graphischen Benutzerschnittstelle die Konzeption und Implementierung einer neuartigen Satzengine zentrales Ziel des Projekts „XML-Print“ war. Durch den dazu gewählten Ansatz einer funktionalen Programmiersprache kann die Mehrprozessoren-Architektur heutiger Computer durch Parallelisierung sehr gut ausgenutzt werden. Zudem werden Seiteneffekte aufgrund des Aufbaus funktionaler Sprachen verhindert.

Auf Basis des Frameworks Mono⁹ wurde in F#¹⁰ eine plattformunabhängige Software entwickelt. Diese lässt sich nicht nur aus der graphischen Oberfläche von XML-Print heraus aufrufen, sondern auch eigenständig über die Kommandozeile / Shell. Somit sind auch ein Batchbetrieb oder die serverseitige Nutzung ohne weiteres möglich.¹¹

3. Inhalte des Workshops

Der Workshop zeigt Anwendern anhand einer Beispieledition die Arbeitsweise von XML-Print auf. Dazu wird ausgehend von einem XML-Dokument zunächst der grundsätzliche Mechanismus von *Formaten* und *Zuweisungen* erklärt und auf verschiedene Elemente des Quelldokuments angewendet. Sodann werden verschiedene Elemente einer Publikation wie Papiergröße, Seitenränder, Schriftart, Kopf- und Fußzeile oder Spaltenzahl angepasst. Abschließend werden spezielle Strukturen wie Fuß- und Endnoten sowie Apparate integriert.

Den Teilnehmern wird die gesamte Bandbreite von XML-Print veranschaulicht, so dass sie in die Lage versetzt werden, das Werkzeug direkt auch für ihre konkreten Publikationsprojekte einzusetzen. Neben der finalen Druckausgabe können dabei z. B. auch verschiedene Lesefassungen

⁹ Siehe <http://mono-project.com/>.

¹⁰ Siehe <http://fsharp.org/>.

¹¹ Als Beispiel für den Einsatz auf einem Server sei auf die Druckfunktion innerhalb des Deutschen Wörterbuchs des Wörterbuchnetzes verwiesen: <http://woerterbuchnetz.de/DWB/>.

eines Quelldokuments mit variierendem Informationsgehalt für unterschiedliche Zielgruppen erzeugt werden.

4. Teilnehmerkreis / Technische Ausstattung

Der Workshop richtet sich an alle interessierten Wissenschaftler, die ihre XML-Dokumente in ein druckfertiges Format überführen wollen oder auch verschiedene Lese- und Zwischenversionen benötigen. Besondere Vorkenntnisse sind nicht nötig. Da es sich um eine „Hands-on“-Sitzung handelt, sollte die Teilnehmerzahl 25 nicht übersteigen. Die Teilnehmer benötigen einen Laptop, auf dem das Programm vorab installiert wurde.¹² Für die Präsentation wird ein Beamer benötigt.

Literatur

Burnard, Lou und Syd Bauman (2007). *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative.

Extensible Stylesheet Language (XSL) (2006). *W3C Recommendation 05 December 2006*. Version 1.1. <http://www.w3.org/TR/xsl11/>.

Extensible Stylesheet Language (XSL) (2012). *W3C Working Draft 17 January 2012*. Version 2.0. <http://www.w3.org/TR/xslfo20/>.

TextGrid: Digital edieren – forschen – archivieren (2013). <http://textgrid.de/>.

XML Path Language (XPath) (2010). *W3C Recommendation 14 December 2010*. Version 2.0. <http://www.w3.org/TR/xpath20/>.

XML-Print: typesetting arbitrary XML documents in high quality (2013). <https://sourceforge.net/projects/xml-print/files/>.

¹² XML-Print ist erhältlich unter: <https://sourceforge.net/projects/xml-print/files/>.