Web browser based CSS typesetting engine

How browser based typesetting systems can be made to work also for printed media

Shinyu Murakami

Vivliostyle Inc.
<murakami@vivliostyle.com>

Johannes Wilm

Vivliostyle Inc.
<johanneswilm@vivliostyle.com>

Abstract

All currently available typesetting systems and formats are rather limited, and the integration between workflows related to print are quite different than those related to web publishing and ebooks.

In this article we argue that the best way forward to unite the workflows is to focus on an HTML-centric workflow, using CSS for styling, and leveraging the power of browsers through the usage of Javascript for print-based layouts.

The Vivliostyle project is working on a new typesetting engine for the next phase of the digital publishing era in which web, ebook and print publishing are unified. We seek to demonstrate here that such a project is needed to bring the three publishing workflows together.

Table of Contents

Introduction

Publishing of long format text in 2015 usually takes three different forms: print as a book, a version to be used on the internet and possibly an ebook.

Ebooks are in most cases EPUB files. The textual content of EPUBs is provided by files containing a restricted version of Hyper Text Markup Language (HTML), the same format used for web pages. The styling of both web pages and EPUBs is defined through Cascading Style Sheets (CSS). Converting content between EPUBs and web pages is therefore not that difficult.

In contrast, most print typesetting systems are using quite different formats and standards than those for ebooks and the web. The workflows from document creation, through editing to final publication differ considerably with different tools and different file formats used. Publishing the same document for print, web and ebooks is therefore difficult, especially for documents that require updating after initial publication as a change in oen fo the files needs to be propagated to all other versions.

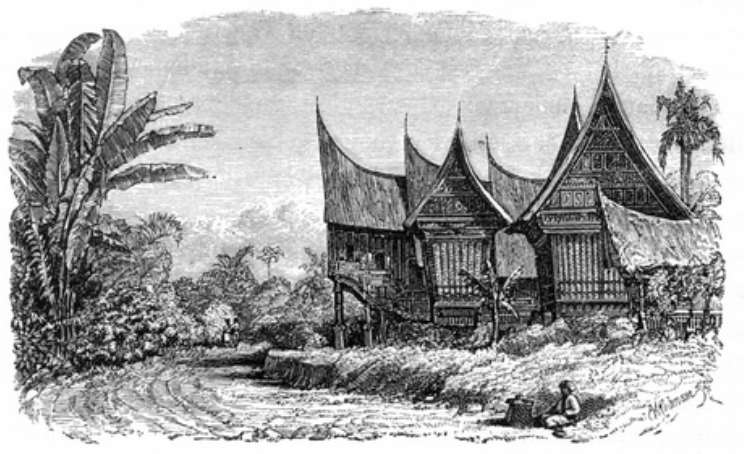
The simplest way to unify the publication processes is to use HTML and CSS also in the print publishing process, so that the same content and style information can be used for all three output formats and a change in one of these will affect all output formats equally. In order to be able to do this, some extra functionality needs to be added to CSS that is specific to the print publishing process.

Using CSS/HTML for print is not a new idea. Other projects that provide print processing functionality using HTML/CSS already exist. Among these are PrinceXML [1], the Antenna House Formatter [2], PDFreactor [3] or Pagination.js [4] and SimplePagination.js [5] .

So far, none of these solutions have been able to establish themselves as the industry standard. In the following, we will argue that all the existing solutions have fundamental shortcomings and that the Vivliostyle project is needed to effectuate a change to web technologies in the print publishing industry.

Figure 1. Vivliostyle.js pages

The Malay Archipelago



Malay Archipelago Chief's House and Rice-shed in a Sumatran Village

The Malay Archipelago is a book by the British naturalist *Alfred Russel Wallace* that chronicles his scientific exploration, during the eight-year period 1854 to 1862, of the southern portion of the *Malay Archipelago* including *Malaysia*, *Singapore*, the islands of *Indonesia*, then known as the *Dutch East Indies*, and the island of *New Guinea*. It was published in two volumes in 1869, delayed by Wallace's ill health and the work needed to describe the many specimens he brought home. The book went through ten editions in the nineteenth century; it has been reprinted many times since, and has been translated into at least eight languages.

The book described each island that he visited in turn, giving a detailed account of its *physical* and *human geography*, its volcanoes, and the variety of animals of plants that he found and collected. At the same time, he describes his experiences, the difficulties of travel, and the help he received from the different peoples that he met. The preface notes that he travelled over 14,000 miles and collected 125,660 *natural history* specimens, mostly of *insects* though also with thousands of *molluscs*, *birds*, *mammals* and *reptiles*.

The Malay Archipelago attracted many reviews, with interest from scientific, geographic, church and general periodicals. Reviewers noted and sometimes disagreed with various of his theories, especially the division of *fauna* and *flora* along what soon became

CSS Paged Media and the limitations of current implementations.

To style elements of pages (electronic or physical) at the most basic level there is a CSS module called "CSS Paged Media" [6] . It has been the main CSS module defining styling elements needed to make exact specifications in CSS for printed and paged output. Several typesetting engines for print output support CSS Paged Media already: The Antenna House Formatter supports CSS as an alternative to XSL-FO, and also PrinceXML supports it.

Unfortunately, these solutions are all entirely closed-source which probably was a deciding factor in determining that they were never able to establish CSS Paged Media as a standard neither for the web nor for their industry. Web browsers have not implemented much of it, even though they have provided features to print web pages and convert to the Portable Document Format (PDF), the file format most commonly used to ensure consistency in print outputs. Even ebook systems, which show individual pages on the screen, have not been very concerned with implementing CSS Paged Media.

An additional factor is that these rendering engines have been built from the ground up, which means that they have to recreate everything browsers have with much less development resources. While they are better than browsers at print, they generally lack behind when it comes to things not directly related to print -- such as JavaScript or other CSS specs. Even though these are not directly print-related, content producers who include design features for the web version will likely generally also expect for these to work in the print version.

What is more, the formatters that do support CSS Paged Media each have their own proprietary vendor extensions for more complex issues that are not compatible with web browsers or even each other for some more complex issues related to paged media.

Even though CSS Paged Media formatters are getting acknowledgment in the XML publishing world, they are therefore still far away from becoming mainstream tools that are widely used.

Figure 2. Vivliostyle.js pages with headers and footnotes

known as the *Wallace line*, *natural selection* and *uniformitarianism*. Nearly all agreed that he had provided an interesting and comprehensive account of the *geography*, natural history, and peoples of the archipelago, which was little known to their readers at the time, and that he had collected an astonishing number of specimens. The book is much cited, and is Wallace's most successful, both commercially and as a piece of literature.

Context

In 1847, Wallace and his friend *Henry Walter Bates*, both in their early twenties,[1] agreed that they would jointly make a collecting trip to the Amazon "towards solving the problem of origin of species";[2] *Charles Darwin*'s book on the *Origin of Species* was not published until 11 years later, in 1859, itself precipitated by a famous letter from Wallace which described the theory in outline.[3] They had been inspired by reading the American *entomologist William Henry Edwards*'s pioneering 1847 book *A Voyage Up the River Amazon, with a residency at Pará*.[4] Bates stayed in the Amazons for 11 years, going on to write *The Naturalist on the River Amazons* (1863); Wallace, ill with fever, went home in 1852 with thousands of specimens, some for science and some for sale. The ship and his collection were destroyed by fire at sea near the Guianas. Rather than giving up, Wallace wrote about the Amazon in both prose and poetry, and then set sail again, this time for the Malay Archipelago.

[1] Bates was 22, Wallace was 24.

[2] Mallet, Jim. "Henry Walter Bates". University College London. Retrieved December 11, 2012.

[3] Shoumatoff, Alex (22 August 1988). "A Critic at Large, Henry Walter Bates". New Yorker.

[4] Edwards, 1847.

Publication

The Malay Archipelago was first published in 1869 in two volumes by Macmillan (London), and the same year in one volume by Harper & Brothers (New York). Wallace returned to England in 1862, but explains in the Preface that given the large quantity of specimens and his poor health after his stay in the tropics, it took a long time. He noted that he could at once have printed his notes and journals, but felt that doing that would have been disappointing and unhelpful. Instead, therefore, he waited until he had published papers on his discoveries, and other scientists had described and named as new species some 2,000 of his beetles (*Coleoptera*), and over 900 *Hymenoptera* including 200 new species of *ant*.[1] The book went through 10 editions, with the last published in 1890.

[1] Wallace, 1869. pp. vii–ix.

Enhancing web browser's page layout with JavaScript

An approach to try to bring page layout to web browsers are Pagination.js and simplePagination.js, which use Javascript in combination with CSS to draw pages which the browsers then can turn into PDF files or send to a printer directly. They provide a lot of the features used for book printing such as table of contents, running headers, page floats, footnotes, word indexes, and margin notes which all have been programmed on top of primitives available in current browsers that are not print specific.

However, both solutions are limited the features at of books. They do not interpret CSS but take configuration options only through Javascript function arguments. And they use methods to achieve their results in current browsers that just happen to work in current implementations, but are not necessarily required to do so according to existing web specifications. For these reasons, they may be usable for certain cases of print currently, but will not be able to replace broader printing solutions.

Standardizing and implementing next generation CSS standards

The Vivliostyle projects seeks to combine and enhance both approaches: Use CSS standards and Javascript for browser based implementations.

Vivliostyle seeks to work with the World Wide Web Consortium (W3C) to enhance and promote specifications such as CSS Paged Media and other related specifications such as "CSS Page Floats" [7] and the "CSS Generated Content for Paged Media Module" [8], working with web browsers to work towards implementation of these specifications in browsers.

Figure 3. Vivliostyle.js pages with non-Latin text

ごん狐

新美南吉

これは、私が小さいときに、村の茂平というおじいさんからきいたお話です。

むかしは、私たちの村のちかくの、中山というところに小さなお城があって、中山さまというおとのさまが、おられたそうです。

その中山から、少しはなれた山の中に、「ごん狐」という狐がいました。ごんは、一人ぼっちの小狐で、しだの一ぱいしげった森の中に穴をほって住んでいました。そして、夜でも昼でも、あたりの村へ出てきて、いたずらばかりしました。はたけへ入って芋をほりちらしたり、菜種がらの、ほしてあるのへ火をつけたり、百姓家の裏手につるしてあるとんがらしをむしりとって、いったり、いろんなことをしました。

或秋のことでした。二、三日雨がふりつづいたその間、ごんは、外へも出られなくて穴の中にしゃがんでいました。

雨があがると、ごんは、ほっとして穴からはい出ました。空はからっと晴れていて、百舌鳥の声がきんきん、ひびいて

いました。

ごんは、村の小川の堤まで出て来ました。あたりの、すすきの穂には、まだ雨のしずくが光っていました。川は、いつもは水が少いのですが、三日もの雨で、水が、どっとましていました。ただのときは水につかることのない、川べりのすすきや、萩の株が、黄いろくにごった水に横だおしになって、もまれています。ごんは川下の方へと、ぬかるみみちを歩いていきました。

東大寺に隣接する稲荷神社の狐像

ふと見ると、川の中に人がいて、何かやっています。ごんは、見つからないように、そうっと草の深いところへ歩きよって、そこからじっとのぞいてみました。

底本：「新美南吉童話集」岩波文庫、岩波書店　平成9年7月15日発行第2刷

Until such support is fully implemented in browsers, Vivliostyle develops Vivliostyle.js [9], a polyfill which will use Javascript to layout pages inside browsers, similar to simplePagination.js and Pagination.js, but it will do so by reading and interpreting the CSS that accompanies the source files and it will provide for a broader usage field, so that styling options can be defined through CSS and will work for a broader usage field than just books.

Additionally, the Vivliostyle Formatter, a Command-Line Interface (CLI) application, and the Vivliostyle Browser, a Graphic User Interface (GUI) application, will embed Vivliostyle.js to allow for PDF output of HTML/XHTML and CSS source files to fit professional publishing needs.

---

[1] http://www.princexml.com

[2] http://www.antennahouse.com

[3] http://www.pdfreactor.com

[4] https://github.com/fiduswriter/pagination.js (requires CSS Regions, previously known as BookJS)

[5] https://github.com/fiduswriter/simplePagination.js (does not require CSS Regions, but has less features than Pagination.js)

[6] http://www.w3.org/TR/css3-page/

[7] http://dev.w3.org/csswg/css-page-floats/

[8] http://dev.w3.org/csswg/css-gcpm/

[9] http://vivliostyle.github.io/vivliostyle.js/ is developed by Toru Kawakubo and is based on Peter Sorotokin's Adaptive Layout implementation.