Submitted on: 06.07.2016

2016 Satellite meeting - *News, new roles & preservation advocacy: moving libraries into action*
10 – 11 August 2016
Lexington, Kentucky USA, USA

Archiving and Accessing HTML-Based Newspapers Using XML and CDATA Strings

Eric Weig
Digital Library Architect, at University of Kentucky Libraries Special Collections Research Center, Lexington, Kentucky, USA
eweig@uky.edu

Abstract:

This article outlines one in-house model for archiving and providing access to HTML-based news in the Kentucky Digital Newspaper Program (KDNP) at the University of Kentucky (UK). To allow for search and retrieval of HTML-based news in the KDNP which already contains news content digitized from analog sources, the encapsulation of HTML content using XML encoded CDATA strings read by a prototype open-source PHP viewer is described.

Keywords: News, newspapers, born digital, libraries, collections, HTML, web harvesting, digital preservation.

Introduction

Archiving HTML-based information, including news, is a concern for librarians, archivists and researchers who are interested in maintaining the historical record for future generations. [7] News digitized from analog sources, as well as born-digital PDF files gathered from the modern digital publishing process, center around a linear page image content model. Much time and effort has been spent over the last decade to design and implement digital library systems for archiving and providing page-turning access to these resources in digital format. Using various structural and descriptive metadata standards, good systems have been developed. Most notably, the work of the National Digital Newspaper Program (NDNP) at the Library of Congress provides a good model for how this kind of system can be successful. [7] In comparison, HTML is not constrained to a linear data model. [1] No clearly defined standards exist yet for a page-turning type of access model for HTML-based news which, in the context of the World Wide Web, is potentially more susceptible to impermanence due to its virtual and often temporary nature. [8] As Jill Lepore describes in a recent article published in The New Yorker, “The Web wasn’t built to preserve its past”. [7] This makes the work of preserving

and providing access to HTML-based news increasingly relevant, as more and more newspaper publishers shift to a purely web-based format. The content must be gathered for preservation as well as an access model that will allow for integration within the page-turning type of access currently used by digital newspaper archives. This will allow the incorporation of HTML-based news into existing page-image based archives such as the NDNP. This will facilitate meaningful search and discovery for HTML-based news alongside even older historic newspaper content derived from analog and/or born digital image-based sources such as Portable Document Format (PDF).

Harvesting HTML News

There are many existing options for website harvesting, including commercial services as well as software, both proprietary and open-source. For the purposes of my initial work in developing and testing an archiving and access model for HTML-based news, no existing harvester was used. A simple Bash script using Wget calls was constructed to gather web-based newspaper issues from *The Clay City Times* (http://www.claycity-times.com/news/), a Kentucky newspaper title the KDNP has been granted license to archive for preservation and access. All resources referenced by the HTML, including images and CSS, were also gathered.

```
#! /bin/bash

# Download the entire contents of http://www.claycity-
times.com/news/

wget -r -l 0 http://www.claycity-times.com/news/
```

Figure 1. Simple BASH script to harvest all the web files for the Clay City Times HTML based newspaper.

Preserving and Accessing HTML vs. Page Images

A digital page-image based newspaper and an HTML-based news web page are visually distinct from one another. The page-image based example is a type of digital facsimile for preservation and access. The facsimile is virtual and the original is analog. The digital facsimile attempts to reproduce the tactile experience of its print source, with content management systems assisting page-turning, reading, and searching. On the other hand, the HTML-based news page holds few if any of the same tactile correlations to a print newspaper. It was created in its original manifestation as a web page. Its digital preservation is based not on facsimile, but rather faithful replication of as much of the original data as possible in order to emulate or imitate the original web page.

Even though digital page-image based newspapers and HTML-based news web pages are very different, they do share some basic attributes which allow them to be combined in complementary ways. Here are some of the most pertinent ways the two formats can complement one another.

- Each represents news, so they can be given common descriptive newspaper metadata such as title, issue date, and publisher.
- Each is serial in nature.

- Each has a clearly demarcated beginning and end; a page image corresponding to the top and bottom of the analog page from which it was derived, and a web page by open and closing HTML tags.

Considering these attributes, it is possible to build an HTML news object similarly to how digital page-image based newspapers are built. I began with the set of harvested news data from *The Clay City Times*. This included HTML news pages that comprised content for the month of September, 2009. However, the Wget process simply replicated the existing directory structures. These directory structures and filenames held no semantic information to establish which files comprised which years, months, and days. Looking at what my Bash script had downloaded, the CSS and image files were grouped in subdirectories quite nicely, although the pages themselves were in one directory with names like 'index.html?p=1234'. Upon further study, however, one important piece of descriptive metadata was found to be common across the set of HTML files. In each file, a <time> tag was present:

```
<time class="entry-time" itemprop="datePublished" datetime="2010-01-
21T16:27:21+00:00">
```

Figure 2. Example date published timestamp in HTML harvested news files.

Keying on this tag, I did a little more bash scripting utilizing grep and Perl regular expressions to rename the files based on date.

```
#!/bin/bash

name=$1
dir=/path/2/html

for f in $dir/$1.html;
do
    date=$(     grep     -oP     '(?<=<time     class\=\"entry\-time\"
    itemprop\=\"datePublished\" datetime\=").*(?=\+00\:00">)' "$f" )
    mv -i "$f" "${date//[^a-zA-Z0-9\._\- ]}".html
done
```

Figure 3. Example date published timestamp in HTML harvested news files.

After running this process, 3,578 files were soon sorted by year and month and day into subdirectories.

```
cla
|-2009
| `-01
|   `-01
|    |-index_0245.html
|    |-index_0246.html
|    |-index_0247.html
|    `-index_0248.html
```

Figure 4. Outline of directory structure for news HTML files organized by the date 2009-01-01.

An additional discovery was that the number at the end of the filenames constituted sequence order within the date groupings. Thus, the harvested files were now organized into individual directories containing the HTML news pages in sequential order for a given day.

Providing Access to News Web Pages

The next step was to fashion an online presentation model which would blend well with page-image based newspaper issues in the KDNP. Little tagged descriptive metadata was accessible within the HTML. So, the ability to enhance that metadata with additional elements would be useful in order to better match the richness of metadata already defined for the page-image based newspaper issues, specifically with the Newspaper Digitization Interest Group (NDIG) metadata specification used in the KDNP. [4] This metadata could then be used to both preserve and provide access to the HTML-based newspaper object. For this reason, a standard format for storing existing metadata harvested, and adding more metadata to the HTML-based news object, was needed. XML for storing the descriptive and structural metadata parts of the HTML-based news object made sense. A header section could contain specific newspaper metadata elements, while a structural metadata section could define the organization of the object's related parts defined as 'web pages' rather than 'page images'. Furthermore, by using CDATA in the XML to encode the harvested HTML pages, a document reader could be fashioned to allow for a web-page-turning type of access. [2]

Encapsulating HTML News with XML

The HTML chunks are wrapped in <hypertext> elements as CDATA. Each <webpage> element contains a time element which encodes the timestamp gathered from the publication date stamp in the HTML. The CDATA is not parsed during validation. This allows for preservation of the original HTML directly within the XML, while maintaining structural validity of the XML document.

```
<?xml version="1.0"?>
<htmlnews>
<record>
<pages>
<webpage></webpage>
<webpage></webpage>
</pages>
<fulltext></fulltext>
</record>
</ntmlnews>
```

Figure 5. Main XML elements for html news document.

```
<webpage>
<time>2009-09-26T11:52:29+00:00</time>
<title>Web Page 1</title>
<hypertext><![CDATA[<article class="post-107 post … ]]></hypertext>
</webpage>
```

Figure 6. Example of the webpage XML element including CDATA.

HTML News Object Page Turner

Let's take a look at how the KDNP displays page-image based newspapers.

Figure 7. Example of page-image based newspaper in the KDNP (https://kdnp.uky.edu/catalog/xt7j9k45r520).

In the above example, page-images are displayed and navigable in sequential order. HTML-based news in the KDNP must be discoverable and navigable in similar ways. In order to serve HTML-based news in the KDNP, a simple PHP-based viewer that reads the XML file and displays the CDATA strings as rendered HTML within a Web browser is used. [11] The viewer for HTML news groups the issues into parts that are not "pages", but "web pages". This is illustrated in the following screenshot which shows HTML-based news in the KDNP.

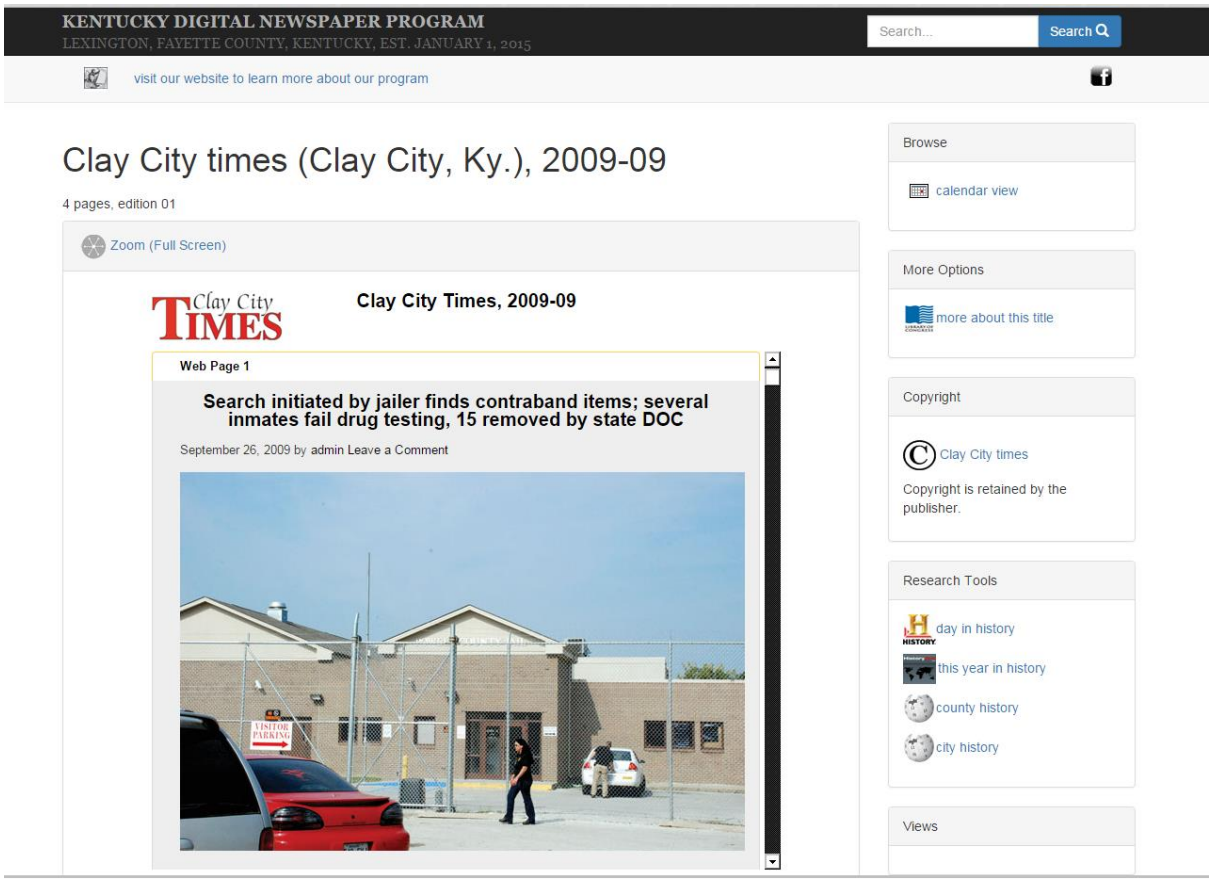

Figure 8. Example of a HTML based newspaper in the KDNP (https://kdnp.uky.edu/catalog/kd9d50ft8m8b).

Indexing HTML News in the KDNP

The KDNP site runs on the Blacklight discovery platform framework, which sits in front of a Solr index. [12] This allows for descriptive metadata searching, limiting by defined facets, and full-text searching based on Optical Character Recognition (OCR) generated text. Solr indexes JSON formatted metadata. In order to view a page-image based newspaper in the KDNP, an external page-image viewer is linked to from the metadata record for an issue.

In the KDNP, HTML-based news works in a similar way. Raw text for the HTML web pages is extracted from the XML document for indexing. This text is stored at the very bottom of the XML structure. A <fulltext> element is used to store the raw text, stripped of tagging, of the HTML pages. This text and descriptive metadata from the XML is used to build a JSON file for Solr indexing. Each newspaper issue is represented by a separate JSON file. The JSON files for the HTML-based news have the same fields as those for the page-image based newspapers. This allows indexing and searching by common fields and the full-text.

Conclusion

Archiving and providing access to HTML-based news in a news archive that also offers access to page-image based newspapers is possible. CDATA within XML is an important component of this approach, and allows for gathering and preserving the HTML without modifying its structure or content, while also allowing for additional descriptive metadata to be added. The XML structure then provides a foundation for preservation and access, allowing for indexing in a content management system and presentation as a navigable news object using a custom viewer written in PHP.

References

1. The World Wide Web Consortium (W3C) . What Is Hypertext? [Internet]. The World Wide Web Consortium (W3C); [cited 2016 Feb 10] . Available from: http://www.w3.org/WhatIs.html
2. Herborth C. 2010 Jan 12. Dealing with Data in XML. [Internet]. IBM DeveloperWorks; [cited 2016 Feb 10]. Available from: http://www.ibm.com/developerworks/library/x-cdata/
3. Nielsen J. 1995 Feb 01. History of Hypertext [Internet]. Nielsen Norman Group; [cited 2016 Feb 10]. Available from: https://www.nngroup.com/articles/hypertext-history/
4. Newspaper Digitization Interest Group (NDIG). 2014. Metadata Application Profile - Digital Newspapers [Internet]. Newspaper Digitization Interest Group (NDIG); [cited 2016 Feb 10]. Available from: https://sites.google.com/site/digitalnewspaperspractices/technical-specifications/metadata-specfication
5. Geiger B. 20 Jan 2016. Fate of Your Archives Is ... Uncertain [Internet]. California Newspaper Publishers Association; [cited 2016 Feb 10]. Available from: http://www.cnpa.com/california_publisher/features/fate-of-your-archives-is-uncertain/article_8f3e2cda-bfd4-11e5-86f8-9797680b24ed.html
6. Pierce V. 2012 Feb 09. Finding That Needle in the Haystack: The Power of Full Text Searching in Chronicling America. [Internet]. South Carolina Digital Newspaper Program; [cited 2016 Feb 10]. Available from: http://library.sc.edu/blogs/newspaper/2012/02/09/finding-that-needle-in-the-haystack-the-power-of-full-text-searching-in-chronicling-america/
7. Lepore J. 2015. What the Web Said Yesterday. The New Yorker [Internet]. [cited 2016 Feb 10] Available from: http://www.newyorker.com/magazine/2015/01/26/cobweb
8. Grainger S. 2000. Emulation as a Digital Preservation Strategy. D-lib Magazine [Internet]. [cited 2016 Feb 10]. Available from: http://www.dlib.org/dlib/october00/granger/10granger.html
9. Johnston L. 2014 Feb 11. Considering Emulation for Digital Preservation [Internet]. The Signal Digital Preservation: Library of Congress; [cited 2016 Feb 10]. Available from: https://blogs.loc.gov/digitalpreservation/2014/02/considering-emulation-for-digital-preservation/

10. Sawers P. 2015 Oct 22. The Internet Archive Is Rebuilding the Wayback Machine to Make Web History Easier to search [Internet]. VentureBeat; [cited 2016 Feb 10]. Available from: http://venturebeat.com/2015/10/22/the-internet-archive-is-rebuilding-the-wayback-machine-to-make-the-webs-history-easier-to-search/
11. University of Kentucky Libraries. 2016. Newz Viewer [Internet]. GitHub Code Repository; [cited 2016 Feb 10]. Available from: https://github.com/uklibraries/newz-viewer
12. Project Blacklight, 2016. Blacklight Discovery Platform Framework [Internet]; [cited 2016 Mar 31]. Availabel from: http://projectblacklight.org/