PUBLIC.RESOURCE.ORG ~ *A Nonprofit Corporation*

Open Source America's Operating System

"It's Not Just A Good Idea—It's The Law!"

October 5, 2009

Mr. Michael L. Wash, Chief Information Officer
Government Printing Office
732 North Capitol Street, N.W.
Washington, D.C. 20401

RE: XML Conversion Tests

Dear Mike:

I would like to thank you for sharing your early drafts of your XML work on the Official Journals. I hope our input has been helpful, and I was thrilled to see you launch the Federal Register XML product as open source available for all developers.

In this memorandum, I'd like to summarize the work we've been doing for the last 6 months on a series of conversion tests on the XML source. Most of the heavy lifting on the XML transformations was done by Point.B Studio, and I'd also like to acknowledge the Sunlight Foundation, which funded our work on this project.

Overview of the Conversion Process

The Federal Register and Code of Federal Regulations have been authored for some time in SGML. We used your SGML source along with the S2 utility which is part of the SGML Parser Toolkit package to provide some initial tests of the feasibility of converting this source to the more modern XML standard. In this test, we were able to convert most, but not all in this semi-automated fashion. You may find these initial tests on the CFR product here:

http://bulk.resource.org/gpo.gov/cfr/raw/

After you approached us and indicated you were also working on SGML->XML conversion, we dropped our initial automated approach and moved over to your format. As we both discovered, a fully general process would not take into account some of the special characteristics of the relatively complex SGML used in the Official Journals.

From the SGML source, you used a program to convert most of the SGML markup into XML. Our initial testing and validation on this indicated that the conversion work you were doing made a lot of sense. We then moved on to focus on two core issues:

- Conversion of the XML to print formats such as PDF.
- Conversion of the XML to more modern on-line formats, particularly HTML.

We were very pleased with the results, some of which I have shown you over the last few months as we completed the various phases of the project.

carl@media.org 1005 GRAVENSTEIN HIGHWAY NORTH, SEBASTOPOL, CALIFORNIA 95472 • PH: (707) 827-7290 • FX: (707) 829-0104

Conversion to PDF Print Format

Our first series of test on the Federal Register XML samples you furnished us focused on the question of replicating the current look-and-feel of the print product, but doing so using style sheets instead of your existing Microcomp typesetting system. You may find all the source for these conversion tests here:

http://bulk.resource.org/gpo.gov/conversion.tests/

The work is based on both sample files, and on the XML Schema Definition you provided:

http://bulk.resource.org/gpo.gov/conversion.tests/FRtoXML.xsd

Our main goal was to replicate, as closely as possible, the current look-and-feel of the printed Federal Register, but using the converted XML instead of the original SGML. The tool we used to accomplish this is PrinceXML, a piece of software developed by some of the core developers of the original CSS standards.

PrinceXML uses CSS as a way of specifying rules for the conversion of XML to PDF. The software has a few extensions as well that are used to provide more sophisticated processing of the print product. For example, in the fedreg.css style sheet we used, the following rule is included:

```
.AGENCY {   font-weight:bold;
            display: block;
            font-family: Arial, Helvetica;
            prince-bookmark-level: 1 }
```

This rule specifies that the <AGENCY> element in the XML should be displayed in Helvetica bold and that any elements are to be used in generated bookmarks. You may find an example of the transformation process in the following files:

- 01AP_01APR1.SGM—The original SGML file.
- 01AP_01APR1.SGM.xml—The XML file.
- 01AP_01APR1.HTML.pdf—The generated PDF file.

Note that bookmarks, footnotes, and tables were all automatically generated. The crosslinks were manually inserted. In addition, Point.B Studio added an intriguing color-coded palette to provide better navigation based on which CFR sections are being affected. A tabloid-sized poster illustrating this color palette is available.

Our conclusion is that with a little bit of work, including potentially some post-processing of the XML, you could replicate the current look-and-feel of the print publication. Doing so would have added benefits such as crosslinks, bookmarks, and potentially the use of color and other navigational aids.

Conversion to HTML Format

The next set of tests we conducted were converting the XML to HTML for on-line viewing. While we focused on conversion to HTML and simple web pages, I would like to stress that the XML format will permit whole site navigations to be created and alternate formats such as EPUB for E-Books to be created.

In your initial XSLT transformation from XML to HTML that you shared with us, you adopted a very straightforward approach of a simple one-to-one mapping of XML to

 elements in HTML. While this approach is quite straightforward and is ideal for a variety of applications, our usability tests indicated that a slightly more sophisticated approach would be more useful for site designers.

In particular, we identified 3 types of mappings that would be useful:

- Placing metadata in <head> elements so the pages are "SEO-optimized." Elements placed in the <head> section include the title, a description, and the name of the agency as "author." We also included the Office of the Federal Register as publisher and a publication date.
- Converting some elements to block elements, in particular a . These are for header elements that should be displayed separately. Specifically, the following XML elements were converted to <DIV>: <FEDREG>, <RULES>, <RULE>, <PREAMB>, "<PRTPAGE P" to "<a name="PRTPAGE" id="anchor, "<AGENCY TYPE=" " to "<div class="AGENCY" id=" ", , , , , <HD SOURCE="HD1, HD2, HD3"> to , , containing elements FP, NAME, TITLE converted to div, , , <FP SOURCE="FP-1"> to , <FRDOC>, <BILCOD>, <SECTION>, <EXTRACT>, <SECTION>, <FP SOURCE="FP-2"> to , <DATED>, <SUBAGY>, <DEPDOC>, <ADD>, <NOTE>, <DATES>, <APPR>, <PRORULES>, <PRORULE>, <NOTICES>, <NOTICE>, <FTNT> (see section on FOOTNOTES), <PREAMHD>, <FP SOURCE="FP1-2">, <CORRECT>, and <EDITOR>
- Having the remaining elements converted to inline elements using the tag: <SECTNO>, <VOL> to Vol., <NO> to No., <DATE>, <INCLUDES>, <UNITNAME>, <E T=" to <span class="ET-, <E T="03"> to [wordspace], <AC T="1"/> to , and <FR>
- Several elements required special consideration: <P SOURCE="NPAR"> to <p class="NPAR">, <MATH> and <MID>, <LI> to
 (</LI> removed), and <SUP> to <sup>.

You can find a simple example of this conversion in the file at this location:

http://webchick.org/FRtoXML/FR-2000-02-01.html

This is a conversion of the following FDSyS file:

http://www.gpo.gov/fdsys/search/pagedetails.action?acCode=FR&granuleId=&packageId=FR-2000-02-01

That file was converted by you to XML, which is here:

http://bulk.resource.org/gpo.gov/conversion.tests/FR-2000-02-01.xml

Finally, we asked Point.B Studio to imagine just a few changes in the underlying XML, in particular the parsing of indented lists into XML tags and the identification of citations. With just a few simple changes in the underlying XML, the results are really quite visible. You can find the sample file at the following location:

http://webchick.org/FRtoXML/

A number of features are apparent in this transformation:

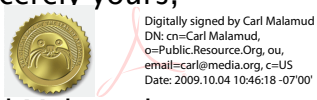
- Color-coding and custom icons are used to provide a visual cue as to which CFR sections are affected. We believe many readers of the Federal Register have a particular agency or subject focus, and the color-coding helps guide the reader. In addition, Point.B Studio has proposed a visual icon system be developed similar to the isotypes developed by Otto Neurath for the National Park System.
- As can be seen, indented lists have been parsed out, leading to a much readable layout. Identifying the list elements is fairly straightforward.
- A table of contents has been provided under the text "Table of Rules and Regulations." Note that a variety of navigational interfaces can be constructed by gathering metadata from throughout the Federal Register.
- Crosslinks are provided to the Code of Federal Regulations (the eCFR from the Archives), to the Federal Register (on FDSyS), and to the U.S. Code (on Cornell's LII system). What is significant here is that the application developer can choose different sources for different kinds of cross links. For example, we have even identified links from species such as Manduca blackburni to the relevant Encyclopedia of Life pages.
- Addresses, such as "901 Locust, Room 506, Kansas City, Missouri 64106" generate a map when the user mouses over the address.

This HTML file is meant to illustrate some of the possibilities that are now open to application developers now that the Official Journals have been moved into modern markup standards such as XML and are available in bulk.

A recommendation I have made previously which I'd like to repeat is that it would be immensely valuable if the Government Printing Office adopted the common industry practice of appointing somebody to handle developer relations, a position meant to work with people who will be repurposing and reusing the Official Journals to build new sites. While it may be seem paradoxical to appoint somebody responsible for encouraging others to develop competing sites to FDSyS, I know from our previous conversations that you are strongly supportive of the idea that good ideas come from the most unexpected places and that developer activity will only help make the GPO sites even better.

On behalf of Point.B Studio, the Sunlight Foundation, and Public.Resource.Org, I would like to thank you for this wonderful opportunity to help alpha test the XML rollout of the Official Journals. I believe your release will be considered a milestone in how government distributes information.

Sincerely yours,



Carl Malamud
President & CEO
Public.Resource.Org

cc: Mr. Raymond A. Mosley, Director, Office of the Federal Register