

## Because Librarians Should Have Fun Too

### ***Introduction***

Meta Data (XML sets) are to Librarians and Formulas are to Mathematicians. But just because you admire the presentation does not mean you understand the fine points like an expert. Unfortunately, Computer Science can not help you there.

Nonetheless, Mathematicians have their presentation aid document type, MathML, and their publishing aid document type, XHTML + MathML 2.0, a combination of web presentation (HTML – XHTML ver. 1.1) with embedded formula presentation (MathML ver. 2.0).

**Why, then, should not Librarians have a web plus meta data presentation format?**

### ***Methods***

Both MathML and XHTML ver. 1.1 are defined by Document Type Definitions (DTD). Metadata Object Description Schema (MODS) from the U.S. Library of Congress is defined with World Wide Web Consortium (W3C) XML Schema (XSD). To ease integration, an XSD form of XHTML ver. 1.0 was used. Unfortunately, the modularization of XHTML ver. 1.1 which made possible the integration of MathML (ver. 2.0), made an easy generation of the XSD Schema format impossible.

The Standardized Generalized Mark-Up Language (SGML) PUBLIC Identifier for the DTD with the addition of MathML is:  
"-//W3C//DTD XHTML 1.1 + MathML 2.0//EN"

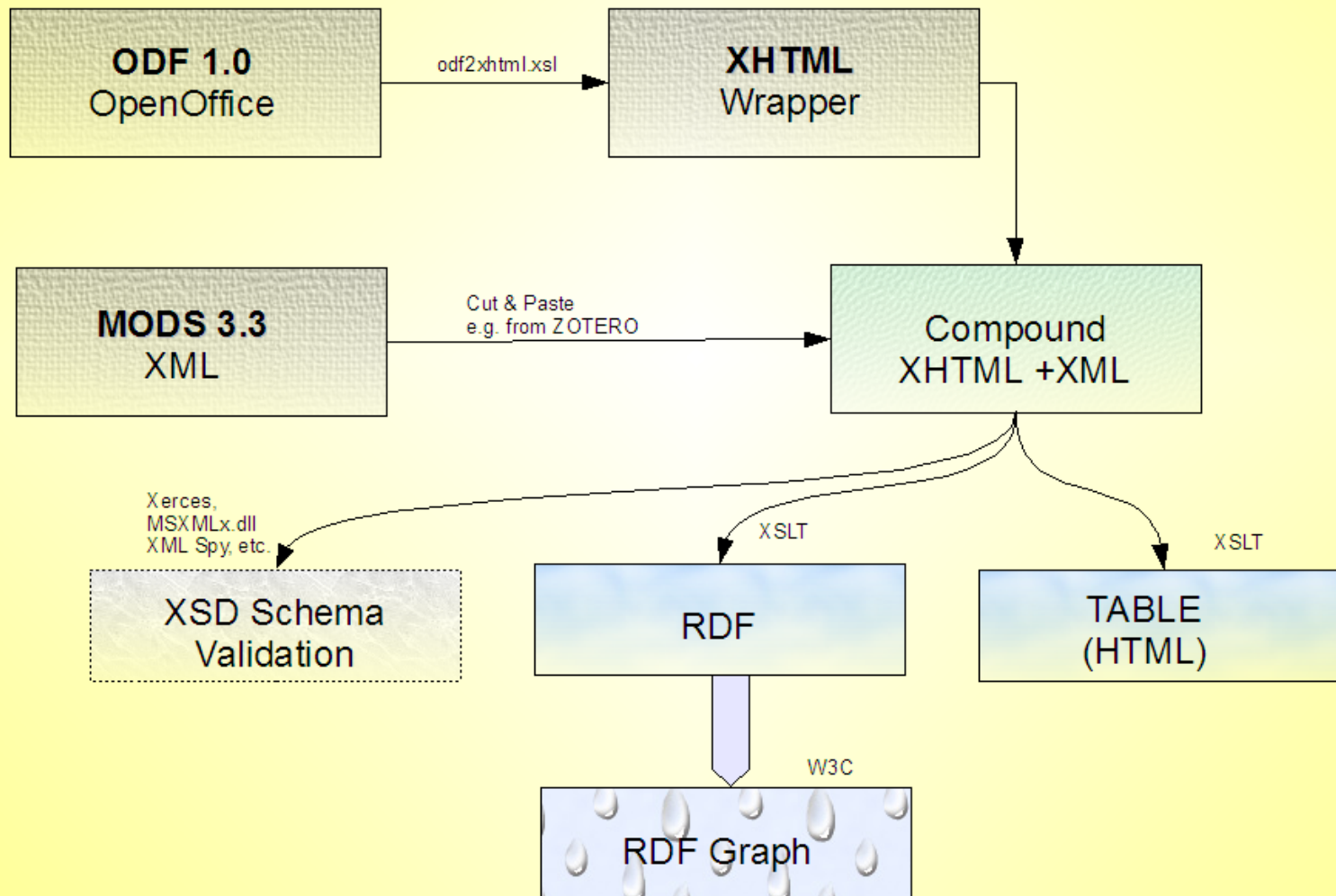
By analogy, a public identifier for our combined schema would be something like:

"-//OOo//XSD XHTML 1.0 + MODS 3.3//EN", although XSD schema do not require the declaration of a <!DOCTYPE ... > header, or a public identifier.

The XHTML "wrapper" for <modsCollection ...> or <mods ... > XML is generated by creating a blank "section" in an OpenOfficeText (ODT) document, then exporting the document with the odf2xhtml.xsl. This creates a user-styled XHTML document with a blank division (<div>) tag in to which a <modsCollection /> or a <mods /> tag can be pasted. The MODS XML can be written by hand, or exported in MODS format from a program such as ZOTERO.

The resulting compound document can be validated, transformed to Resource Description Framework (RDF) file or MODS data transformed to a tabular format.

<!DOCTYPE html PUBLIC "-//OOo//XSD XHTML 1.0 + MODS 3.3//EN" >



The XHTML 1.0 + Math 2.0 file operations are identical. The [W3C RDF Validator](#) will list RDF Tuples, and draw a graph.

## ***Results***

## **Compound Document Format**

### ***The Wrapper (XHTML)***

The information wrapper of choice for the World Wide Web is the Hypertext Mark-Up Language Family (HTML Family). Virtually all documents made for display in web browsers are served as static HTML – even though the document content may have been produced to-order, that is, dynamically. Considerable effort has been expended on software to manipulate the style and layout of HTML pages to make them pleasing to the eye. Our tool of choice for wrapper development is OpenOffice. This software suite is open source, an ISO Standard [ISO] and will export (write) any file in XHTML 1.0 format.

In addition, and of equal importance to Librarians, these wrappers expose their own meta data according to Dublin Core Metadata Initiative (DC) definitions.

All the features of OpenOffice layout and style are available, including document template features where copies of the same wrapper are used to make uniform the “look and feel” of a collection of pages.

### ***MODS XML***

A MODS format XML file is inserted between division tags of the wrapper and the result, a multi-namespace or Compound Document saved.

```
<div class="Sect1">  
  <mods ... >  
</div>
```

The MODS XML need not be a complete reference for validation, but it must be valid XML. While the Compound Document can be validated, and will display in a web browser, the dynamic Cascading Style Sheet (CSS) methods used for the MODS file display are somewhat crude, as raw data often is. No one will miss the point that it is valid, reusable raw data, usable verbatim in other wrappers without loss of validity.

The [Zotero] add-on for the FireFox web browser will enable one to accumulate “libraries” of web reference sources which later can be written out in MODS 3.2 format. These are, in effect, electronic bibliographic information index cards.

Zotero will also export plain text citations in several formats. When possible, MODS data should be listed in plain text with the authority for the text layout, in addition to the rest of the MODS elements. It is recommended that in-situ references to “authority” be used, as in the following example:

```
<extension xmlns:dcterms="http://purl.org/dc/terms/">
  <dcterms:Standard>
    <dcterms:conformsTo>Chicago Manual of Style</dcterms:conformsTo>
    <dcterms:isReferencedBy>http://dublincore.org/usage/terms/history/#Standard-001</dcterms:isReferencedBy>
    <dcterms:DCMIType>Text</dcterms:DCMIType>
  </dcterms:Standard>
  <dcterms:bibliographicCitation xml:lang="en">POCS</dcterms:bibliographicCitation>
  <dcterms:bibliographicCitation xml:lang="English">Plain Old Chicago Style</dcterms:bibliographicCitation>
</extension>
```

*The extension will not be validated<sup>1</sup>* except for well-formed XML syntax. The reference for the standard must be a Uniform Resource Identifier (URI), in the case above, the default definition of a standard (citation). The DCMI Type (genre) refers to the content of the standard – not the content of the reference – so, for example, <mods:TypeOfResource> and <mods:genre> are unrelated to the citation.

## Output Formats

### **Validation**

Validation is a necessary option, but as with other tools (Spill Chalkier, The Sorries, etc.), not in-line all the time. Really, why bother? Well ... you bother because like mathematicians, people who deal with meta data do not always catch well buried semantic mistakes. Computers are dumb, after all, which explains why experts knew what they meant, but you don't.

---

1 The citation is encoded as a plain text string, validation would add little, for substantial effort – linking the DCTERMS schema to MODS. DC meta data terms in XHTML are not validated either, however, the modification of encoding syntax should tell you that *DC meta data in XHTML (or HTML) is not XML*. An acceptable work-around is GRDDL, for those who actually work. There is no acceptable talk-around. Unless you link in the extension name space (like PII), a MODS <extension/> is unparsed character data (CDATA), no matter how much it looks like XML.

Go away, kid, you are getting on my nerves.

### ***Resource Definition Framework***

The Compound Document has multiple meta data sets. The wrapper document has meta data, and each MODS file also represents a (bibliographic) resource. Just as MODS divisions must adopt a new name space to be written in an XHTML table, a MODS division must be segregated from meta data in the wrapper. At some future date, an XML citation may include a complex over-loaded ontology, and rather than confuse that issue with fragmentation, we use the <extension>, plain text citation, in our RDF.

### ***Tabular***

The tabular listing contains most of the pertinent information in the citation list. However it points out the deficiencies in the non-ontological “tag soup” approach to meta data representation.

### ***Discussion***

The companion zip file includes the XSD schema to validate compound documents of XHTML 1.0 (strict) plus either MODS 3.3 or MathML 2.0, along with finished examples of each. The Wrapper (XHTML) should be made to order depending upon the MathML or MODS subject. OpenOffice makes the assembly of these wrappers routine – just insert an empty Section where appropriate – pictures, commentary and references can be inserted outside the empty section.

Open Office is available from [OpenOffice.org](http://OpenOffice.org)

The companion zip file is available for download [here](#).

### **MODS Specific**

The MODS 3.3 (XSD) schema was borrowed from the Library of Congress. One change was necessary: 1) a consistent reference to the XLINK name space was made. This was done to alleviate the need for web connectivity during validation, but also to eliminate a requirement for cross-domain connectivity which many view as a possible security risk.

Zotero is available from [Zotero.org](http://Zotero.org) As a FireFox add-on, this program provides a quick way to generate MODS format (and plain text) references.

References to Personally Identifiable Information [PII] are also included as a [MODS extension](#). It is the author's hope that this addition to MODS will enable more comprehensive references and collation, of document collections which would otherwise not be indexed for reasons of personal privacy. The PII XSD schema is included in the zip file.

## **MathML Specific**

The MathML (XSD) schema was borrowed from the NCBI, and originated with the W3C. Three changes were necessary: 1) a consistent reference to one XLINK name space file was made, and 2) where necessary, the attributes were set to "qualified", and 3) the name space prefix was set to "math:" instead of "mml:". The changes were made so that OpenOffice Formula could be used as a MathML editor – in much the same way as Zotero is used as a "MODS Editor". Still, several changes must be made to the MML v1.01 (modified by SUN Microsystems) output, primarily in the presentation. For example, the schema name for "math:fontstyle" is really "math:mathvariant", etc.. Nonetheless, the OpenOffice Formula Editor provides a good head start for the novice programmer.

## ***References***

[ISO] – Information technology — Open Document Format for Office Applications (OpenDocument) v1.0, ISO/IEC 26300:2006 First Edition 2006-12-01

[Zotero] – The Center for History and New Media <<http://chnm.gmu.edu/>>, George Mason University, Fairfax, VA, USA, v1.0.4, see also: <http://www.zotero.org/>

[PII] – The Personally Identifiable Information Name Space <<http://purl.org/pii/terms/>>, Gannon J. Dick, Azle, TX, USA, v1.0.3, see also: <http://www.rustprivacy.org/>