

Literate Documentation for XML – METS-ODD Treatment

Kevin Reiss 6/12/2007

The Metadata Encoding and Transmission Standard (METS) provides a means to encode metadata about any type of digital object. Serving a similar role for digital objects that the Text Encoding Initiative (TEI) does for any form of text, METS is a loosely-defined vocabulary meant to encode metadata for any type of digital object. The current METS W3C XML Schema is an open schema divided into seven different sections that are roughly analogous to the role taken by the current group of modules defined for the TEI.

The basic METS standard also has a companion W3C XML Schema, called the METS Profile, which provides a means to restrict the content models of the generic METS schema and specify the format that a particular class of METS documents must take. The profile expresses both syntactic and semantic constraints for the subset of METS instances conforming to the profile. However, there is currently no way to produce a validating schema from a given METS profile to support automatic processing of METS documents at the profile level. The profiles currently consist of human readable semi-structured instructions stating what an author must do to create a document that conforms to the profile. In order for the METS Profile to be a truly effective as a mechanism to support the creation and processing of conforming documents there needs to be a means to produce both concise, clear prose documentation and a validating schema that can check for both syntactic and semantic constraints that differ from the generic METS model. The need to produce a validating XML schema and effective, well-formatted prose documentation points to literate programming based upon the TEI P5 One Document Does it All (ODD) as a means to solve these problems.

METS / Extended ODD

This project will use the current TEI P5 ODD vocabulary and processing system as a base for experimenting with the production of a validating schema and well-formatted prose documentation for a METS Profile. The ODD format itself will be extended to allow a profile to express the semantic issues discussed in the next section. The TEI ODD processing system will also need to be customized to process the ODD extensions that will allow METS profile authors to express semantic constraints such as the ID/IDREF mechanism. This system will also be useful in helping profile authors clearly express syntactic constraints as well. It will enable them to easily restrict, constrain, or extend the basic METS content models using the well-designed mechanisms built into the ODD.

Syntactic v. Semantic Issues

A particularly problematic issue for METS documents that seek to conform to a given profile is the resolution and restriction of XML ID/IDREF references to appropriate element types that can be associated with ID values in a METS document. There is no current way to computationally restrict the usage and application of these references in the current standard beyond the generic matching a syntactically valid XML IDREF

value to a corresponding XML ID value within the same METS document. Most of the publicly available METS profiles restrict the use of ID/IDREF elements to specific file types. I will also attempt to identify other relevant semantic issues that appear in METS profile that could also be potentially clarified using a literate programming system. I will not seek to systematically explore every semantic issue that could possibly be identified in METS documents, but will focus on those that seem most pressing to document authors after taking a survey of the publicly available METS profiles.

Though the investigation the expression of semantics will be the primary focus of the project, the syntactic constraints that most current XML schemas cannot will also be touched upon by this project. All of the available METS Profiles proscribe attribute value and co-occurrence constraints upon the content models that seek to conform to the profile. The project report delivered will take care to differentiate between questions of syntax and semantics.

Deliverable: METS-ODD Demonstration Application

The project will include an expression of the current W3C METS XML Schema in what will be called METS-ODD format. These extensions will take the form of XML element and attributes that clarify these issues. A validating schema will be produced through the use of the existing TEI P5 ODD system. This schema will facilitate the authorship of METS Profiles using the METS-ODD format. Extensions to the ODD system will be created and incorporated into this schema using the existing ODD customization mechanism that enable the author to express the syntactic and identified semantic constraints that are necessary to create a useful METS Profile. Figure one displays an example of what the extension mechanism that enables an author to clarify of ID/IDREF semantics for a profile might look like within a METS-ODD document instance.

Figure One:

```
<idconstraints>
  <elname>fptr</elname>
  <atname>FILED</elname>
  <must-reference>file/@format="image/jpeg"</must-reference>
  <profile-constraint>The FILEID of a <fptr> element must reference a <file> element
    that has an attribute value of "image/jpeg".</profile-constraint>
</idconstraints>
```

This declaration, and others like it, will be processed by the extended ODD processing system. XPATH expressions like that supplied in the hypothetical <must-reference> element above will be embedded into XSLT or Schematron statements that can be used to validate METS Profile document instances. Similar statements that use XPATH and or Schematron will be provided to allow an author to express any other problematic semantic constructs identified as commonly occurring within METS profiles. These constructs may include such problems such as the propagation of assertions made by the presence of a particular attribute value to the child elements of the element that has the attribute value attached to it and the resolution of the semantics expressed by the

parent/child relationship between XML elements within a METS Profile. This extended system will also provide another important benefit to the authors of METS Profiles. The METS-ODD format will enable authors to express syntactic constraints via XPATH and Schematron expressions using the extended ODD's mechanism for declaring XML content models.

The project report will include the following sections:

1. The history of literate programming for SGML/XML
2. A discussion of the problematic nature of the representation of semantics within XML schema.
3. The problems of semantic expression, the lack of a validating schema, and inadequate natural language in the context of METS and the METS Profile.
4. The details of the exploration of these issues in the current, publicly available METS Profiles
5. The results of how successful the experimental METS-ODD format and processing are at allowing a METS author at expressing the semantic and syntactic constraints that make up a profile.
6. The relation of these problems to general purpose XML schema design and construction.
7. A consideration of how effective the extended ODD system created for the project would be as a general purpose literate programming tool for XML.