XML Schema Checklist

Roger Costello
October 2012

Before publishing your XML Schemas, check that they contains these 4 things:

1. xml:lang
2. xml:id
3. pattern facets
4. user-defined entities

xml:lang

Add `xml:lang` on the root element. `xml:lang` is a standard XML attribute, of type `xs:language`, used to show the language of the data in an instance document. Here is an example:

```
<Document xml:lang="en">
   ...
</Document>
```

Showing the language of the data is important for:

- correctly rendering or styling the text
- applying spell-checking during content authoring
- appropriate selection of voice for text-to-speech systems

xml:id

Liberally sprinkle `xml:id` throughout the document. `xml:id` is a standard XML attribute, of type `xs:ID`, used to uniquely identify a section of an instance document. Every major section in an instance document should be uniquely identified. Here is an example:

```
<BookStore xml:lang="en" xml:id="ABC_Books">
    <Book xml:id="PM">
        ...
    </Book>
    <Book xml:id="JV">
        ...
    </Book>
</BookStore>
```

A section with an identifier can be linked to from within the instance document as well as from outside the instance document. XML APIs such as DOM are able to take advantage of them; for instance, the DOM has this method: `getElementById("PM")`, which returns the section with the identifier `PM`.

pattern facet

Use the `pattern` facet on every leaf element with a `xs:string` data type (or with a data type that derives from `xs:string`). The `pattern` facet is used to constrain the allowable characters, thereby reducing the risk of attacks from malicious code and reducing the risk of spilling sensitive information. Here is an example:

```
<xs:simpleType name="characters">
    <xs:annotation>
        <xs:documentation>
            A message is a series of characters.  A message that is
            conformant with this schema is composed of characters
            with values in the range of 1 through 127.
        </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
        <xs:pattern value="[&#1;-&#127;]*" />
    </xs:restriction>
</xs:simpleType>

<xs:simpleType name="line">
    <xs:annotation>
        <xs:documentation>
            Messages are divided into lines of characters.
            A line is a series of characters that is delimited
            with the two characters carriage-return and line-feed.
            Each line of characters MUST be no more than 998
            characters, and SHOULD be no more than 78 characters,
            excluding the CRLF.
        </xs:documentation>
    </xs:annotation>
    <xs:restriction base="characters">
        <xs:maxLength value="1000" />
        <xs:pattern value="[&#1;-&#127;-[&#13; &#10;]]*&#13;&#10;" />
    </xs:restriction>
</xs:simpleType>
```

The value of a `pattern` facet is a regular expression. To enable reuse of your regexes, use user-defined entities in your XML Schema. That's described next.

user-defined entities

User-defined entities in your XML Schema enable reuse. Here is an example:

```
<?xml version="1.1" encoding="UTF-8"?>
<!DOCTYPE xs:schema [
```

```
<!ENTITY US_ASCII_CHARACTER "&#1;-&#127;">
<!ENTITY CR "&#13;">
<!ENTITY LF "&#10;">
<!ENTITY CRLF "&CR;&LF;">
<!ENTITY VCHAR "&#33;-&#126;">      <!-- Viewable (printing) char -->
<!ENTITY COLON "&#58;">
<!ENTITY DIGIT "0-9">
<!ENTITY ALPHA "a-zA-Z">
<!ENTITY DQUOTE "&#x22;">           <!-- Double quote -->
<!ENTITY SP "&#32;">                <!-- Space -->
<!ENTITY HTAB "&#9;">               <!-- Horizontal tab -->
<!ENTITY WSP "(&HTAB;|&SP;)">       <!-- Whitespace -->
]>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

    <xs:simpleType name="characters">
        <xs:annotation>
            <xs:documentation>
                A message is a series of characters.  A message that is
                conformant with this schema is composed of characters
                with values in the range of 1 through 127.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:string">
            <xs:pattern value="[&US_ASCII_CHARACTER;]*" />
        </xs:restriction>
    </xs:simpleType>

    <xs:simpleType name="line">
        <xs:annotation>
            <xs:documentation>
                Messages are divided into lines of characters.
                A line is a series of characters that is delimited
                with the two characters carriage-return and line-feed.
                Each line of characters MUST be no more than 998
                characters, and SHOULD be no more than 78 characters,
                excluding the CRLF.
            </xs:documentation>
        </xs:annotation>
        <xs:restriction base="characters">
            <xs:maxLength value="1000" />
            <xs:pattern value="[&US_ASCII_CHARACTER;-[&CRLF;]]*&CRLF;" />
        </xs:restriction>
    </xs:simpleType>
```

Observe the entity declarations at the top (before `xs:schema`) and the use (and reuse) of the entities within the `pattern` facets.