Report on

SPASE Text Normalization and Mark-up

March 11, 2009

Overview

SPASE resource metadata can contain textual content that describe the resource. While descriptive text may be brief some formatting of the text may be necessary to convey the necessary information. A principal of the SPASE data model is that an element may contain a value or other elements. Mixed content (values plus elements) is not allowed. To be consistent with this principal SPASE should adopt a set of text normalization rules which will allow some structuring of textual content. This document describes a recommended set of text normalization rules and demonstrates how those rules can be used in practice.

Text Normalization Rules

Text or values assigned to an element are constrained by the data type of the element. For example, a date/time value is constrained by SPASE to be ISO8601 compliant. This constrains the formation and content of the date/time value. Text values in SPASE are any string permitted characters. To aid in determining the layout or structural intent of the author the following rules are to be applied to text to create a normalized form:

1. All lines are to end with a newline character.
2. All text is left justified. No line has leading whitespace.

Text Interpretation Rules

After normalization of text the following rules can be used to interpret the layout intent of the author.

1. Blank lines indicate paragraph breaks.
2. Lists
    a. Must be preceded by a blank line.
    b. Items are indicated by a line beginning with a reserved character followed by a space. Three levels of lists are supported. The reserved characters are:
        * : First level list
        - : Second level list (must appear within a first level context)
        . : Third level list (must appear within a second level context)
    c. End with a blank line.
3. Tables
    a. Begin and end with a line that starts with "+--".
    b. The first "row" of a table is the field headings.
    c. Fields in a table are separated with a vertical bar ("|").
    d. Visual row separators are lines which begin with "|--".

XML Implementation

For a schema based XML parser to deliver the content of tags in a form which can be properly normalized according to the SPASE rules, the white space normalization must be "preserve". In an XML Schema this is the behavior of the data type of "string". Using other data types of white space normalization will obscure the layout intent of the author.

Examples

For the purpose of illustrating the normalization and interpretation of text consider the following XML fragment. This fragment includes all supported layout constructs.

```
<Description>Multiple paragraphs of text
      that is indented in the XML file.
      A list appears between the first two
      paragraphs. This is the first paragraph.

         * First line in list.
            - First line in level 2 list
            - Second line in level 2 list
            - Third line in level 2 list
         * Second line in list.
         * Third line in list.
         * Forth line in list.

      This is the start of the second paragraph.
      All text should appear on multiple lines.
      with no leading spaces and one line between
      paragraphs. Next is a table.

         +---------------------------------+
         | Col1   | Col2                   |
         |--------|------------------------|
         | One    | First line of table    |
         | Second | Second line of table   |
         +---------------------------------+

      This is the paragraph after the table.
      It's very short.
</Description>
```

During processing and normalization the "Description" tag is removed and all text is left justified, resulting in:

```
Multiple paragraphs of text
that is indented in the XML file.
A list appears between the first two
paragraphs. This is the first paragraph.

* First line in list.
- First line in level 2 list
- Second line in level 2 list
- Third line in level 2 list
* Second line in list.
* Third line in list.
* Forth line in list.

This is the start of the second paragraph.
All text should appear on multiple lines.
with no leading spaces and one line between
paragraphs. Next is a table.

+--------------------------------+
| Col1   | Col2                  |
|--------|-----------------------|
| One    | First line of table   |
| Second | Second line of table  |
+--------------------------------+

This is the paragraph after the table.
It's very short.
```

The normalized text can then be rendered using different methods. To render the text as plain text similar to original text in the description lists and tables can be indented. To render as HTML the paragraphs, lists and tables can be encapsulated with the appropriate tags. An HTML equivalent for the text is:

```
<p>multiple paragraphs of text
that is indented in the XML file.
A list appears between the first two paragraphs.
This is the first paragraph.
</p><p>
<ul><li>First line in list.</li>
<ul><li>First line in level 2 list</li>
<li>Second line in level 2 list</li>
<li>Third line in level 2 list</li>
</ul><li>Second line in list.</li>
<li>Third line in list.</li>
<li>Forth line in list.</li>
</ul></p><p>
This is the start of the second paragraph.
All text should appear on multiple lines.
with no leading spaces and one line between
paragraphs. Next is a table.
</p><p>
<table>
<tr><th>Col1</th><th>Col2</th></tr>
<tr><td>One</td><td>First line of table</td></tr>
<tr><td>Second</td><td>Second line of table</td></tr>
</table>
</p><p>
This is the paragraph after the table.
Its very short.
</p>
```

Which will appear in a browser as: