ITS 2.0 VALIDATION AND MIGRATION TOOLS FOR HTML5 AND XHTML

Jiří Kosek

Distribution: Public

MultilingualWeb-LT (LT-Web)
Language Technology in the Web

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable title:	ITS 2.0 VALIDATION AND MIGRATION TOOLS FOR HTML5 AND XHTML
Contractual date of delivery:	December 2013
Actual date of delivery:	June 2013
Author(s):	Jiří Kosek (subcontractor)

Revision History

Revision	Date	Author	Organization	Description
1	11/09/2013	Jiří Kosek		First draft
2	19/09/2013	Jiří Kosek		Improvements based on the feedback from others

CONTENTS

1. INTRODUCTION

ITS 2.0 specification (http://www.w3.org/TR/its20/) defines set of metadata categories which can be used to augment various document types. Translation and localization processes can be much more effective after such augmentation by useful metadata. Previous version of ITS 1.0 was targeted only for XML based content including XHTML. Because of this ITS 1.0 could rely on XML namespace mechanism when adding ITS markup into a host vocabulary such as XHTML or DocBook.

In the past few years it become clear that future of Web development will be centered around HTML5 not around XHTML. Unfortunately HTML5 does not provide XML namespace syntax and different HTML5-friendly syntax for expressing ITS 2.0 metadata has to be used. However this created a lot of small but important problems. How to roundtrip ITS augment content between HTML5 and XHTML? How to validate HTML5 pages containing ITS markup?

Results of this deliverable are:

- Offline validation tool for HTML5 with ITS 2.0 markup.
- Offline conversion tools between HTML5 and XHTML with ITS 2.0 markup.
- Integrated ITS 2.0 support into a widely used HTML validation services validator.nu and validator.w3.org.

2. USAGE SCENARIOS

2.1. Validation of HTML5 content

HTML5 document uses specially prefixed attributes to store ITS 2.0 metadata, for example:

```
<p><span its-term="yes">HTML</span> is Web markup langugage.</p>
```

Such attributes are not standard part of HTML5 they are allowed only in conforming HTML5+ITS documents (see http://www.w3.org/TR/its20/#conformance-product-html5-its).

For normal HTML5 pages several validators can be used for checking their syntax correctness. Two most popular validators are W3C validator and validator.nu. As a part of this deliverable both has been recently extended with HTML5+ITS support. As a result any web author now can use ITS markup in HTML pages and markup will be automatically checked for correctness without doing any special additional action.

W3C validator automatically checks against HTML5+ITS conformance all documents labelled as HTML5 via proper !DOCTYPE (`<!DOCTYPE html>`).

In validator.nu support for HTML5+ITS validation must be explicitly selected by using “Preset” control.

Both validators in action are shown on the following screenshots.

SAMPLE VALIDATION OUTPUT FROM VALIDATOR.NU

Validation results

Validator Input

Text Field

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset=utf-8>
    <title>Validation test</title>
  </head>
  <body>
    <p>The <span translate=np>World Wide Web Consortium</span> is
      making the World Wide Web worldwide!</p>
    <p><span its-term="yves">HTML</span> is Web markup langugage.</p>
    <p its-foo="bar">Another problem.</p>
  </body>
</html>
```

Encoding: As set by the server/page

Schemas: http://s.validator.nu/html5-its.rnc http://s.validator.nu/html5/assertions.sch http://c.validator.nu/all/

Preset: HTML5 + SVG 1.1 + MathML 3.0 + ITS 2.0

Parser: Automatically from Content-Type

XMLNS Filter:

- [ ] Be lax about HTTP Content-Type
- [ ] Show Image Report
- [x] Show Source
- [ ] Show Outline

Validate

Group Messages

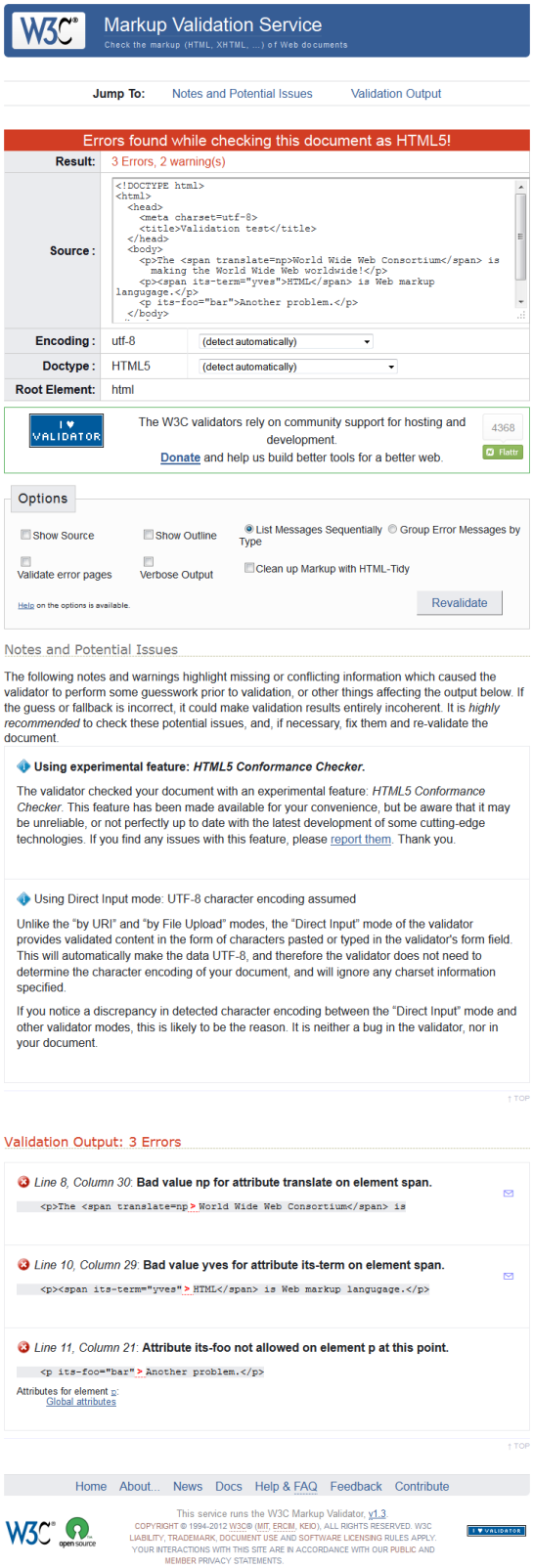
1. **Info**: The Content-Type was `text/html`. Using the HTML parser.
2. **Error**: Bad value `np` for attribute `translate` on element span.
   From line 8, column 12; to line 8, column 30
   `<p>The <span translate=np>World`
3. **Error**: Bad value `yves` for attribute `its-term` on element span.
   From line 10, column 8; to line 10, column 29
   `p>↵ <p><span its-term="yves">HTML</`
4. **Error**: Attribute `its-foo` not allowed on element p at this point.
   From line 11, column 5; to line 11, column 21
   `.</p>↵ <p its-foo="bar">Anothe`
   Attributes for element p:
   Global attributes

There were errors.

Total execution time 24 milliseconds.

About this Service • Simplified Interface

SAMPLE OUTPUT FROM W3C VALIDATOR



Markup Validation Service
Check the markup (HTML, XHTML, ...) of Web documents

Jump To: Notes and Potential Issues Validation Output

Errors found while checking this document as HTML5!

Result: 3 Errors, 2 warning(s)

Source:

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset=utf-8>
    <title>Validation test</title>
  </head>
  <body>
    <p>The <span translate=np>World Wide Web Consortium</span> is
      making the World Wide Web worldwide!</p>
    <p><span its-term="yves">HTML</span> is Web markup
language.</p>
    <p its-foo="bar">Another problem.</p>
  </body>
```

Encoding: utf-8 (detect automatically)
Doctype: HTML5 (detect automatically)
Root Element: html

I ♥ VALIDATOR
The W3C validators rely on community support for hosting and development.
Donate and help us build better tools for a better web.
4368 Flattr

Options

Show Source | Show Outline | List Messages Sequentially | Group Error Messages by Type
Validate error pages | Verbose Output | Clean up Markup with HTML-Tidy
Help on the options is available. Revalidate

Notes and Potential Issues

The following notes and warnings highlight missing or conflicting information which caused the validator to perform some guesswork prior to validation, or other things affecting the output below. If the guess or fallback is incorrect, it could make validation results entirely incoherent. It is *highly recommended* to check these potential issues, and, if necessary, fix them and re-validate the document.

Using experimental feature: *HTML5 Conformance Checker*.

The validator checked your document with an experimental feature: *HTML5 Conformance Checker*. This feature has been made available for your convenience, but be aware that it may be unreliable, or not perfectly up to date with the latest development of some cutting-edge technologies. If you find any issues with this feature, please report them. Thank you.

Using Direct Input mode: UTF-8 character encoding assumed

Unlike the "by URI" and "by File Upload" modes, the "Direct Input" mode of the validator provides validated content in the form of characters pasted or typed in the validator's form field. This will automatically make the data UTF-8, and therefore the validator does not need to determine the character encoding of your document, and will ignore any charset information specified.

If you notice a discrepancy in detected character encoding between the "Direct Input" mode and other validator modes, this is likely to be the reason. It is neither a bug in the validator, nor in your document.

↑ TOP

Validation Output: 3 Errors

Line 8, Column 30: **Bad value np for attribute translate on element span.**

```
<p>The <span translate=np> World Wide Web Consortium</span> is
```

Line 10, Column 29: **Bad value yves for attribute its-term on element span.**

```
<p><span its-term="yves"> HTML</span> is Web markup language.</p>
```

Line 11, Column 21: **Attribute its-foo not allowed on element p at this point.**

```
<p its-foo="bar"> Another problem.</p>
```

Attributes for element p:
Global attributes

↑ TOP

Home About... News Docs Help & FAQ Feedback Contribute

This service runs the W3C Markup Validator, v1.3

Apart from making validator ITS 2.0 aware the test suite input files were incorporated into general conformance checks for Web Platform (https://github.com/w3c/web-platform-tests/tree/master/conformance-checkers). This will assure that the outcome of this work will be sustainable for a long time and for many users outside the ITS community as well.

In scenarios where online validation services could not be used, one can use offline validator. For example to validate document named test.html the following command can be used:

```
$ java -cp schema;lib/jing.jar;lib/html5-
datatypes.jar;lib/iri.jar;lib/js.jar;lib/htmlparser.jar;lib/icu4j-4_4_2.jar
com.thaiopensource.relaxng.util.Driver -c schema\html5-its-lang.rnc test/test.html
e:\src\html5-its-tools\test\test.html:8:30: error: Bad value “np” for attribute
“translate” on element “span” from namespace “http://www.w3.org/1999/xhtml”.
e:\src\html5-its-tools\test\test.html:10:29: error: Bad value “yves” for attribute
“its-term” on element “span” from namespace “http://www.w3.org/1999/xhtml”.
e:\src\html5-its-tools\test\test.html:11:21: error: Attribute “its-foo” not
allowed on element “p” from namespace “http://www.w3.org/1999/xhtml” in this
context.
```

The error messages are same as in the web interface. As you can see validation can easily catch common mistakes like malformed attribute values or names.

2.2. Conversion between HTML5 and XHTML markup

The details of syntax for writing down ITS metadata differ between HTML5 and XHTML. In HTML5 its-* prefixed attributes are used:

```
<p><span its-term="yes">HTML</span> is Web markup langugage.</p>
```

Whereas in XHTML namespaces are used:

```
<html xmlns=http://www.w3.org/1999/xhtml
      xmlns:its=”http://www.w3.org/2005/11/its”>
  …
  <p><span its:term="yes">HTML</span> is Web markup langugage.</p>
```

There are many workflows which are XML/XHTML based, but the result has to published in HTML5. There are also opposite scenarios where HTML5 content has to be processed by an existing XML toolchain. In such scenarios conversion between XHTML and HTML5 is required and ITS metadata has to be retained even if different syntax is used in an underlying formats.

As a part of this deliverable two XSLT transformations (xhtml2html.xsl and html2xhtml.xsl) were created which can be used for roundtripping ITS metadata between XHTML and HTML5 syntaxes.

3. Download and source code

All deliverables are published as an open-source. Source code of offline validation tool and conversion transformations can be downloaded from https://github.com/kosek/html5-its-tools.

Patches for validator.nu and W3C validator are available at https://bitbucket.org/kosek/syntax and https://bitbucket.org/kosek/validator and are regularly merged into the main development branch.