An HTML5 Conformance Checker

Henri Sivonen

What? Why?

- Checks if the input meets the machine-checkable conformance criteria for HTML5
- Quality assurance tool for authors
- Find errors you didn't intend to make

HTML5

- Features for Web applications
- Use case and interoperability driven
- Thoroughly specified processing models
- Does not pretend to be SGML-based
- Ongoing process – no complete spec yet

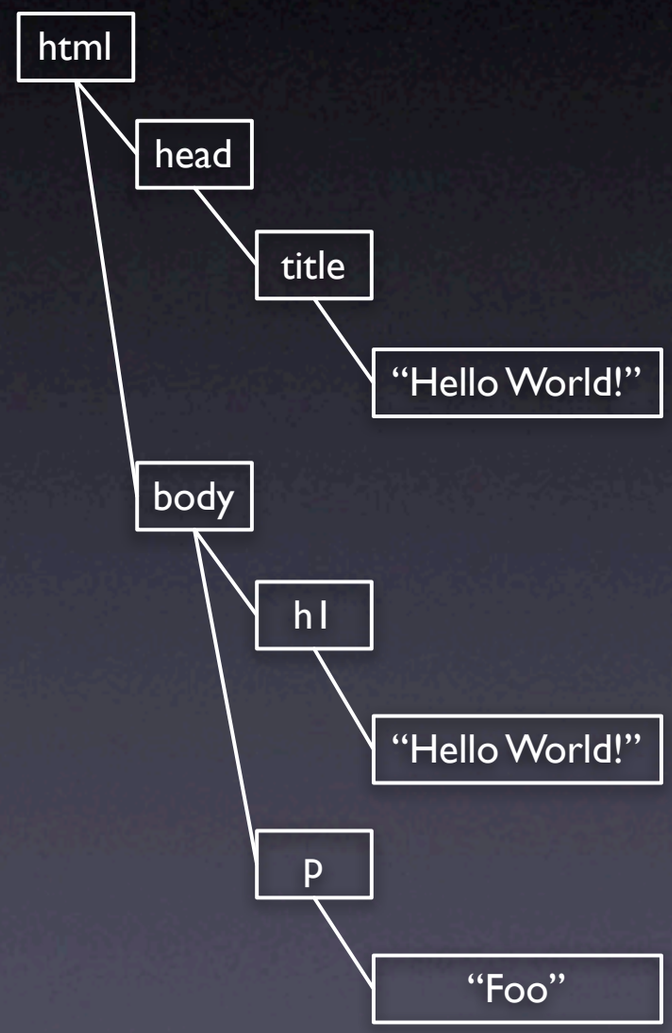
HTML5 and XHTML5

- Two serializations
- Similar document trees
- text/html ⇒ HTML5
- application/xhtml+xml ⇒ XHTML5

Looks Kinda Similar…

- ```
  <!DOCTYPE html>
  <html>
    <head>
      <title>Hello World!</title>
    </head>
    <body>
      <h1>Hello World!</h1>
      Foo
    </body>
  </html>
  ```

- ```
  <html xmlns="http://www.w3.org/1999/xhtml">
    <head>
      <title>Hello World!</title>
    </head>
    <body>
      <h1>Hello World!</h1>
      <p>Foo</p>
    </body>
  </html>
  ```



HTML 4 Validation

- SGML DTD-based
  - But browsers don't support SGML minimizations like <title/Hello/
- All theoretically machine-checkable constraints are not checked:
  - <ins datetime="foobar"> is valid but not conforming

HTML5 Conformance Checking

- No DTDs
- If a machine can check a requirement, do it!
- Schema capabilities not an excuse
- No official schema
- No endorsed schema languages

No Schemata?

- Feed Validator
- Turing-complete languages can check everything that is machine-checkable
- Lots of hand-crafted code
- Wouldn’t schemata be nice as a baseline?

Best of Both Worlds

- A RELAX NG schema as the baseline
- Refine with Schematron
- Refine even more with Java

RELAX NG

```
blockquote.elem =
  element blockquote { blockquote.inner & blockquote.attrs
                     }
blockquote.attrs =
  ( common.attrs
  & blockquote.attrs.cite?
  )
  blockquote.attrs.cite =
    attribute cite {
      common.data.uri
    }
blockquote.inner =
  ( common.inner.block )
```

Schematron

- <rule context="h:blockquote">
    <report test="ancestor::h:header">
      The blockquote element cannot appear as a
      descendant of the header element.
    </report>
  </rule>

- <rule context='h:input[@list]'>
    <assert test='id(@list)/self::h:datalist or
                  id(@list)/self::h:select'>
      The list attribute of the input element must
      refer to a datalist element or to a select element.
    </assert>
  </rule>

Java

- Table integrity checker
- Unicode normalization checking
- Format of text content of elements

Conclusions

Correct Expectations

- Mapping HTML5 to XHTML5 works
- Schemata insufficient but easy to develop
- Non-schema-based checkers needed
- The quality of error messages from RELAX NG validation are a problem

RELAX NG Surprises

- RELAX NG less applicable than expected
- Bad for exclusions
- *RELAX NG DTD Compatibility* more trouble than it is worth

Schematron Surprises

- Less applicable than expected
  - Ancestor–descendant relationships
  - Referential integrity
- Embedding Schematron inside RELAX NG is overrated
- Could be treated as a rapid prototype

Overall

- Success!

Questions?