DtdAnalyzer

A tool for analyzing and manipulating DTDs

Demian Hess, Avalon Consulting, LLC

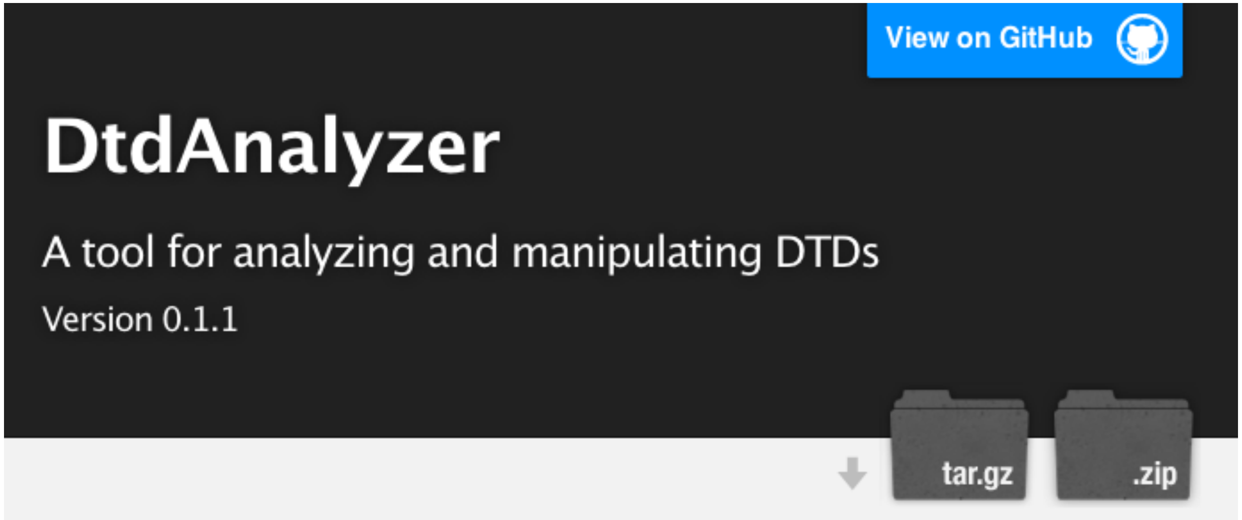
Chris Maloney, NCBI/NLM/NIH, contractor with A-Tek, Inc.

Audrey Hamelers, NCBI/NLM/NIH, contractor with IMC/KEVRIC

JATS Users Conference
October 17, 2012

<JATS-Con>

What is the DtdAnalyzer?



DtdAnalyzer

A tool for analyzing and manipulating DTDs

Version 0.1.1

- DTD goes in, XML comes out
- Open source project
  http://ncbitools.github.com/DtdAnalyzer/
- Java based

```
dtdanalyzer -s mydtd.dtd output.xml
```

DtdAnalyzer

Main Use Cases

- Analyze DTDs
- Create documentation
- Scripts and scaffolding

DtdAnalyzer

Comparing DTD Versions

Two versions of a DTD:

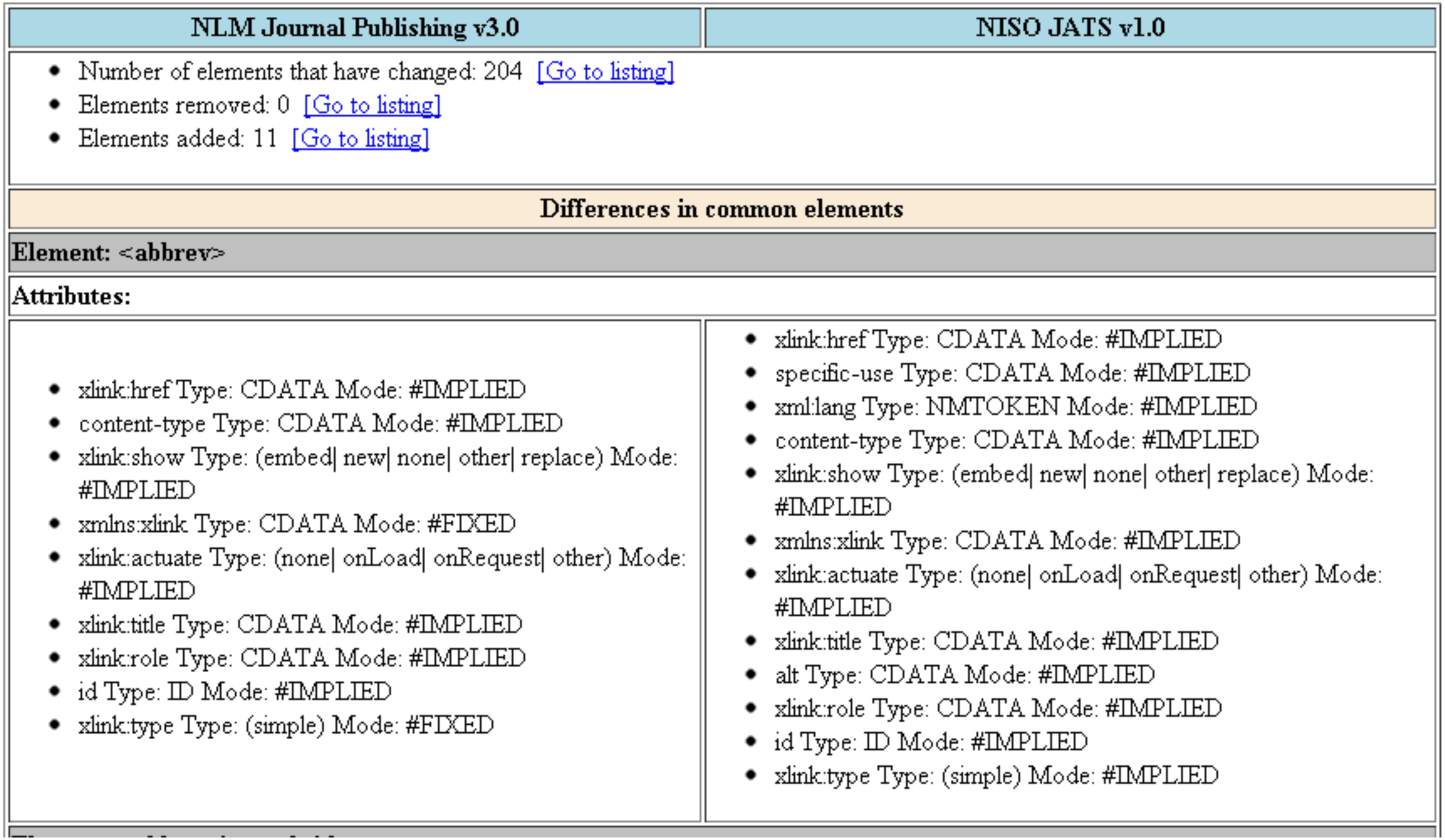
- NLM Publishing Version 3.0
- NISO JATS Version 1.0

Questions

- Which elements have changed?
- Which elements were added?
- Which elements were removed?

DtdAnalyzer

DTD Comparison Report



NLM Journal Publishing v3.0	NISO JATS v1.0
• Number of elements that have changed: 204 [Go to listing] • Elements removed: 0 [Go to listing] • Elements added: 11 [Go to listing]	
Differences in common elements	
Element: <abbrev>	
Attributes:	
• xlink:href Type: CDATA Mode: #IMPLIED • content-type Type: CDATA Mode: #IMPLIED • xlink:show Type: (embed| new| none| other| replace) Mode: #IMPLIED • xmlns:xlink Type: CDATA Mode: #FIXED • xlink:actuate Type: (none| onLoad| onRequest| other) Mode: #IMPLIED • xlink:title Type: CDATA Mode: #IMPLIED • xlink:role Type: CDATA Mode: #IMPLIED • id Type: ID Mode: #IMPLIED • xlink:type Type: (simple) Mode: #FIXED	• xlink:href Type: CDATA Mode: #IMPLIED • specific-use Type: CDATA Mode: #IMPLIED • xml:lang Type: NMTOKEN Mode: #IMPLIED • content-type Type: CDATA Mode: #IMPLIED • xlink:show Type: (embed| new| none| other| replace) Mode: #IMPLIED • xmlns:xlink Type: CDATA Mode: #IMPLIED • xlink:actuate Type: (none| onLoad| onRequest| other) Mode: #IMPLIED • xlink:title Type: CDATA Mode: #IMPLIED • alt Type: CDATA Mode: #IMPLIED • xlink:role Type: CDATA Mode: #IMPLIED • id Type: ID Mode: #IMPLIED • xlink:type Type: (simple) Mode: #IMPLIED

DtdAnalyzer

XML Representation of any DTD

```
<declarations>
    <elements>...</elements>
    <attributes>...</attributes>
   <parameterEntities>...</parameterEntities>
   <generalEntities>...</generalEntities>
</declarations>
```

DtdAnalyzer

Use XML Tools for Analysis

How many elements in a DTD?

```
count(//element)
```

How many attributes have multiple declarations?

```
for $attribute in $dtd//attribute
  let $declarations := distinct-values(
    for $d in $attribute/attributeDeclaration
      return concat($d/@type, "|", $d/@mode, "|", $d/@defaultValue)
  )
order by $attribute/@name
return
  if (count($declarations) gt 1) then concat("Attribute ",
        string($attribute/@name), " has multiple declarations" )
  else ()
```

DtdAnalyzer

Create Structured Comments

Delimited Comment

```
<!--~~ <split>
Specifies the main ingredients of a
banana split:
* One banana,
* Two banana
~~-->
<!ELEMENT split (banana)*>
```

Resulting XML

```
<element name="split" dtdOrder="61">
    <annotations>
       <annotation type='notes'>
        <p>Specifies the main ingredients
        of a banana split:</p>
        <ul>
          <li>One banana</li>
          <li>Two banana</li>
        </ul>
      </annotation>
   </annotations>
```

DtdAnalyzer

Different Annotation Sections ...

```
<!--~~ <split>

~~ model
Four bananas make a bunch and so do many more.

~~ tags
tagA tagB

~~ examples
    <split>
      <banana instrument='guitar'>Foo</banana>
      <banana instrument='drums'>Bar</banana>
    </split>
~~-->
```

DtdAnalyzer

... Produce Different Results

```
<annotation type='model'>
  <p>Four bananas make a bunch and so do many more.</p>
</annotation>

<annotation type='tags'>
  <tag>tagA</tag>
  <tag>tagB</tag>
</annotation>

<annotation type='examples'>
  <pre><code>&lt;split&gt;
   &lt;banana instrument='guitar'&gt;Foo&lt;/banana&gt;
   &lt;banana instrument='drums'&gt;Bar&lt;/banana&gt;
  &lt;/split&gt;</code></pre>
<annotation>
```

DtdAnalyzer

Comments in Markdown

- Easy to read as text
- Easy to learn and use
- Widely used
- Can be mixed with HTML

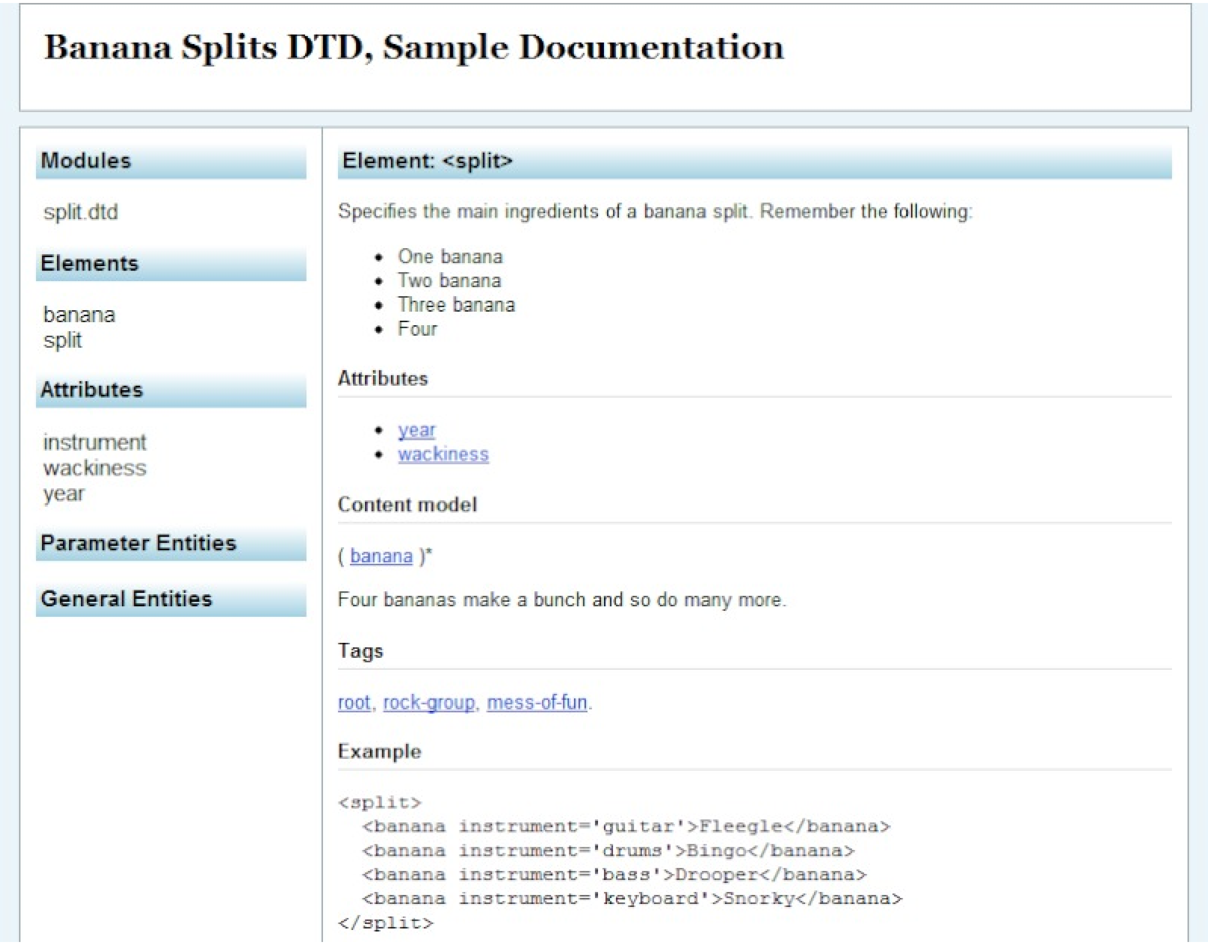
```
## Similar tools
* DTDInst
* DTDDoc
* LiveDTD
```

```
<h2>Similar tools</h2>
<ul>
  <li>DTDInst</li>
  <li>DTDDoc</li>
  <li>LiveDTD</li>
</ul>
```

DtdAnalyzer

Human Readable Documentation

```
dtddocumentor -dir output -s my-dtd.dtd
```



Banana Splits DTD, Sample Documentation

Modules

split.dtd

Elements

banana
split

Attributes

instrument
wackiness
year

Parameter Entities

General Entities

Element: <split>

Specifies the main ingredients of a banana split. Remember the following:

- One banana
- Two banana
- Three banana
- Four

Attributes

- year
- wackiness

Content model

(banana)*

Four bananas make a bunch and so do many more.

Tags

root, rock-group, mess-of-fun.

Example

```
<split>
  <banana instrument='guitar'>Fleegle</banana>
  <banana instrument='drums'>Bingo</banana>
  <banana instrument='bass'>Drooper</banana>
  <banana instrument='keyboard'>Snorky</banana>
</split>
```

DtdAnalyzer

Documentation Command Options

```
dtddocumentor -dir output -s my-dtd.dtd
```

- --roots foo bar baz
  Specify a set of root elements.
- --exclude mml:
  Exclude any element beginning with "mml:" ...
- --exclude-except mml:math
  ... except "mml:math"
- --css mycss.css
- --js myjs.js
  Skin with custom CSS and JS

DtdAnalyzer

When Documentation Isn't Enough

Documentation specifies proper nesting

```
<!--~~ <section>
The level of each section must be nested properly: level 2 must be inside level 1,
level 3 must be inside level 2, etc.
~~ examples
    <section level="1">
      <section level="2"><p>Hello world</p></section>
    </section>
~~-->
<!ELEMENT section (p | section)+>
<!ATTLIST section level CDATA #REQUIRED>
<!ELEMENT p (#PCDATA)>
```

DtdAnalyzer

Rule Enforced with Schematron

```
<sch:schema xmlns:sch="http://purl.oclc.org/dsdl/schematron">
<!-- ... -->
  <sch:pattern id="elements">
    <sch:title>Element Checks</sch:title>
    <sch:rule context="//p"/>
    <sch:rule context="//section">
      <sch:report test="parent::section and
        (number(parent::section/@level) ne (number(@level) - 1))">
        Child section level must be one greater than its parent
       </sch:report>
    </sch:rule>
  </sch:pattern>
<!-- ... -->
</sch:schema>
```

DtdAnalyzer

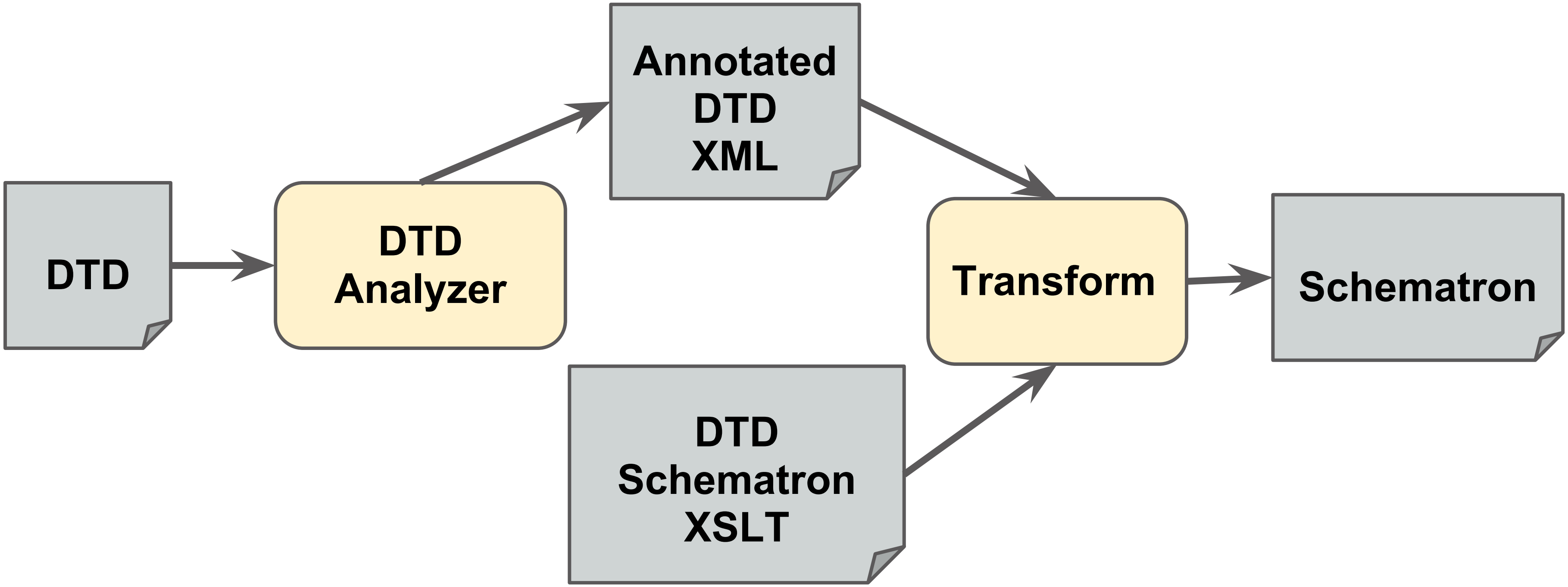
Annotations Descriptive *and* Prescriptive

```
<!--~~ <section>
~~ examples
  <section level="1">
   <section level="2">
    <p>Hello world</p>
   </section>
  </section>
~~ schematron
 <report test="parent::section and (number(parent::section/@level) ne
                                   number(@level) - 1)">
  Child section level must be one greater than its parent's level
 </report>
~~-->
<!ELEMENT section (p | section)+>
<!ATTLIST section level CDATA #REQUIRED>
```

DtdAnalyzer

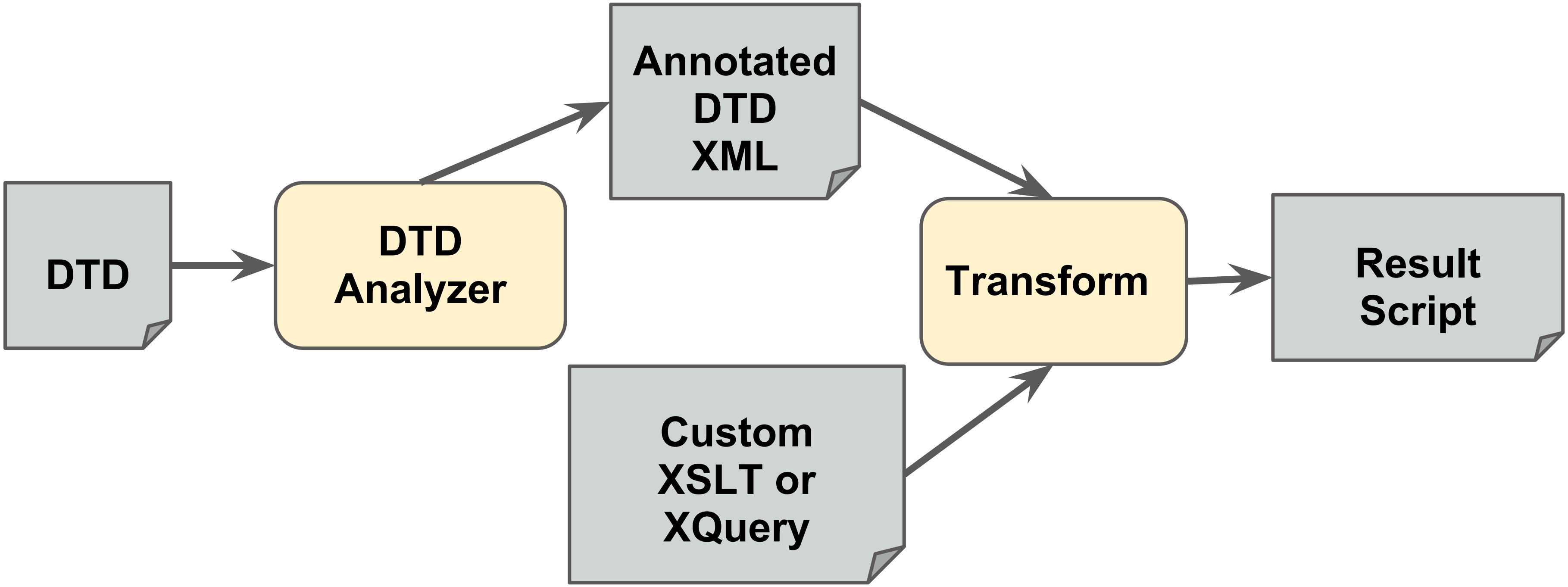
Rule Checking Process

```
dtdanalyzer -s my-dtd.dtd -x dtdschematron.xsl schematron.sch
```



General Pattern for Scripting

```
dtdanalyzer -s my-dtd.dtd -x scaffold.xsl result.xsl
```



DtdAnalyzer

Automatic Scaffolding

dtdanalyzer -s journalpublishing3.dtd -x scaffold.xsl result.xsl

```
<!-- ************************************************************************** -->
<!-- Template for: article -->
<!-- ========================================================================== -->
<!-- Content model:
  ( front, body?, back?, floats-group?, ( sub-article* | response* ) )

  Attributes:
  -article-type (Type: CDATA; Mode: #IMPLIED)
  -dtd-version (Type: CDATA; Default Value: 3.0; Mode: #FIXED)
  -xml:lang (Type: NMTOKEN; Default Value: en)
  -xmlns:mml (Type: CDATA; Default Value: http://www.w3.org/1998/Math/MathML; Mode:
  #FIXED)
  -xmlns:xlink (Type: CDATA; Default Value: http://www.w3.org/1999/xlink; Mode:
  #FIXED)
  -xmlns:xsi (Type: CDATA; Default Value: http://www.w3.org/2001/XMLSchema-instance;
  Mode: #FIXED)
-->
<!-- ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ -->
<xsl:template match="article">
  <xsl:copy>
    <xsl:copy-of select="@*"/>
    <xsl:apply-templates/>
  </xsl:copy>
</xsl:template>
```

DtdAnalyzer

Advantages of Scaffolding

- Reduces typing
- Enforces coding conventions
- Auto-generation of code can be driven by annotations

Example: Creating Minified XML

Problem

- Need XSL that "minifies" element names
- Second XSL to restore elements to original names

Solution

- Copy scaffold.xsl
- Add function to create minified name
- Repeat for "unminify" script

DtdAnalyzer

From Minified XML ...

```xml
<el> <h>Comparing Forward-Error Correction and Redundancy</h></el>
<eo>
  <ep xmlns:xlink="http://www.w3.org/1999/xlink" contrib-type="author"  xlink:type="simple">
    <ce name-style="western">
      <cf>Mouse</cf>
      <cg>Anon E.</cg>
    </ce>
    <w>Master of Ceremonies</w>
    <i>Institute of Scientific Research in Redundancy</i>
    <bp>
      <br>USA</br>
      <bv>301.555.1234</bv>
    </bp>
    <bs xlink:type="simple">anon.e.mouse@nowhere.net</bs>
    <ct xlink:type="simple">
      <hc>Anon E. Mouse is Master of Ceremonies at the Institute of Scientific Research in Redundancy. Her research in the field has been recognized by publications known thoughout the world as leading publications in her research field. Anon has made significant contributions to the standards most commonly used in the fields of error reporting and correction and continues to participate in the working groups for said standards.</hc>
    </ct>
  </ep>
</eo>
```

DtdAnalyzer

... to Restored XML

```
<title-group>
   <article-title>Comparing Forward-Error Correction and Redundancy</article-title>
</title-group>
<contrib-group>
   <contrib xmlns:xlink="http://www.w3.org/1999/xlink" contrib-type="author"   xlink:type="simple">
      <name name-style="western">
         <surname>Mouse</surname>
         <given-names>Anon E.</given-names>
      </name>
      <role>Master of Ceremonies</role>
      <aff>Institute of Scientific Research in Redundancy</aff>
      <address>
         <country>USA</country>
         <phone>301.555.1234</phone>
      </address>
      <email xlink:type="simple">anon.e.mouse@nowhere.net</email>
      <bio xlink:type="simple">
         <p>Anon E. Mouse is Master of Ceremonies at the Institute of Scientific Research in Redundancy.
Her research in the field has been recognized by publications known thoughout the world as leading
publications in her research field. Anon has made significant contributions to the standards most commonly
used in the fields of error reporting and correction and continues to participate in the working groups for said
standards.</p>
```

DtdAnalyzer

Auto-Generated Minified DTD

dtdanalyzer -x make-minified-dtd.xsl -s articleauthoring3.dtd minified.dtd

```
<!ELEMENT hj (#PCDATA|cl)*>
<!ATTLIST hj xlink:href CDATA #IMPLIED>
<!ATTLIST hj content-type CDATA #IMPLIED>
<!ATTLIST hj xlink:show (embed|new|none|other|replace) #IMPLIED>
<!ATTLIST hj xmlns:xlink CDATA #FIXED "http://www.w3.org/1999/xlink">
<!ATTLIST hj xlink:actuate (none|onLoad|onRequest|other) #IMPLIED>
<!ATTLIST hj xlink:title CDATA #IMPLIED>
<!ATTLIST hj xlink:role CDATA #IMPLIED>
<!ATTLIST hj id ID #IMPLIED>
<!ATTLIST hj xlink:type (simple) #FIXED "simple">

<!ELEMENT dp (co?,hc*,ij*)>
<!ATTLIST dp specific-use CDATA #IMPLIED>
<!ATTLIST dp xml:lang NMTOKEN #IMPLIED>
<!ATTLIST dp abstract-type CDATA #IMPLIED>
<!ATTLIST dp id ID #IMPLIED>

<!ELEMENT hr (#PCDATA)*>
<!ATTLIST hr content-type CDATA #IMPLIED>
<!ATTLIST cs id ID #IMPLIED>
```

DtdAnalyzer

Could Use for Multilingual Tags

```xml
<titres>
   <titre>Comparing Forward-Error Correction and Redundancy</titre>
</titres>
<auteurs>
   <auteur xmlns:xlink="http://www.w3.org/1999/xlink" contrib-type="author"   xlink:type="simple">
      <noms name-style="western">
         <surname>Mouse</surname>
         <pre-noms>Anon E.</pre-noms>
      </noms>
      <role>Master of Ceremonies</role>
      <affiliation>Institute of Scientific Research in Redundancy</affiliation>
      <addresse>
         <pays>USA</pays>
         <telephone>301.555.1234</telephone>
      </addresse>
      <e-mail xlink:type="simple">anon.e.mouse@nowhere.net</e-mail>
      <biographe xlink:type="simple">
         <paragraphe>Anon E. Mouse is Master of Ceremonies at the Institute of Scientific Research in
Redundancy. Her research in the field has been recognized by publications known thoughout the world as
leading publications in her research field. Anon has made significant contributions to the standards most
commonly used in the fields of error reporting and correction and continues to participate in the working groups
for said standards.</paragraphe>
```

DtdAnalyzer

Future Possibilities

- Tighter integration with JATS public documentation
    - Official deployed JATS could be annotated with structured comments
    - JATS customizations that import those core modules would get integrated documentation
    - Customized documentation could transclude core JATS documentation

DtdAnalyzer

Other Possibilities

- Use structured content model for DTD comparisons
- Leverage annotations for scripting and scaffolding
- Promote community use and involvement

DtdAnalyzer

Feedback and Collaboration

- GitHub project page:
  http://ncbitools.github.com/DtdAnalyzer/
- Mailing list:
  https://groups.google.com/d/forum/dtdanalyzer
- Bug reports:
  https://github.com/NCBITools/DtdAnalyzer/issues