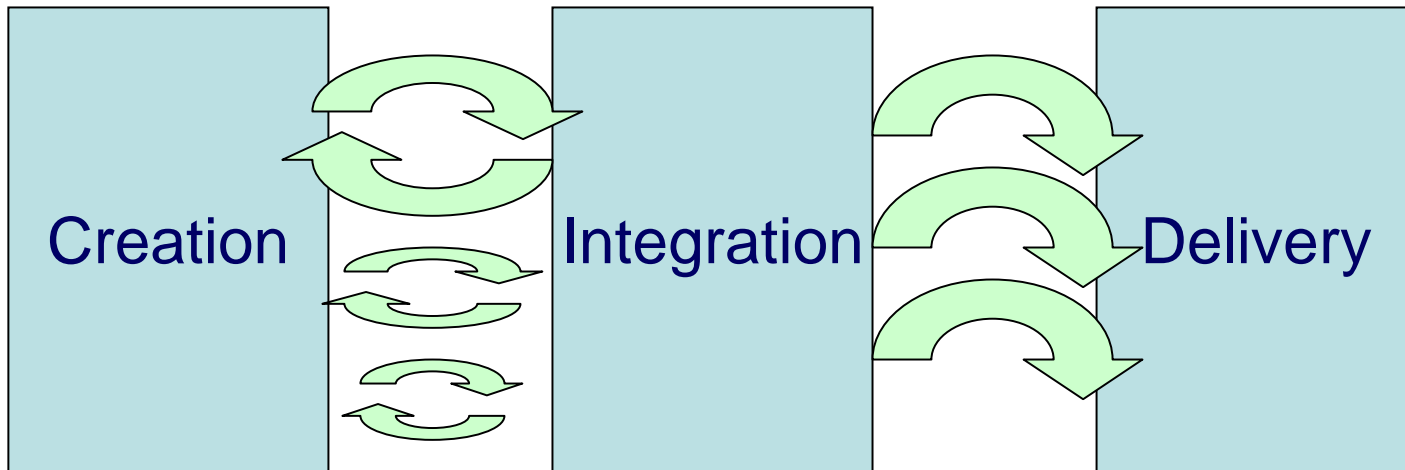


Data Integrity and Document- centric XML

Using Schematron for Managing Text Collections

Dazhi Jiao, Tamara Lopez

XML Collections in the DLP



Overview

- XML Validation
 - Defined, tools
- Schematron
 - History, features, the process, the language
- DLP implementation
 - Customizations, framework, demo
- Conclusions
 - Current/future directions, challenges

XML validation: Types

- Structure and Content types
 - Syntactic, focused on definition, fixed
- Rules for Use
 - Semantic, focused on relationships, variable

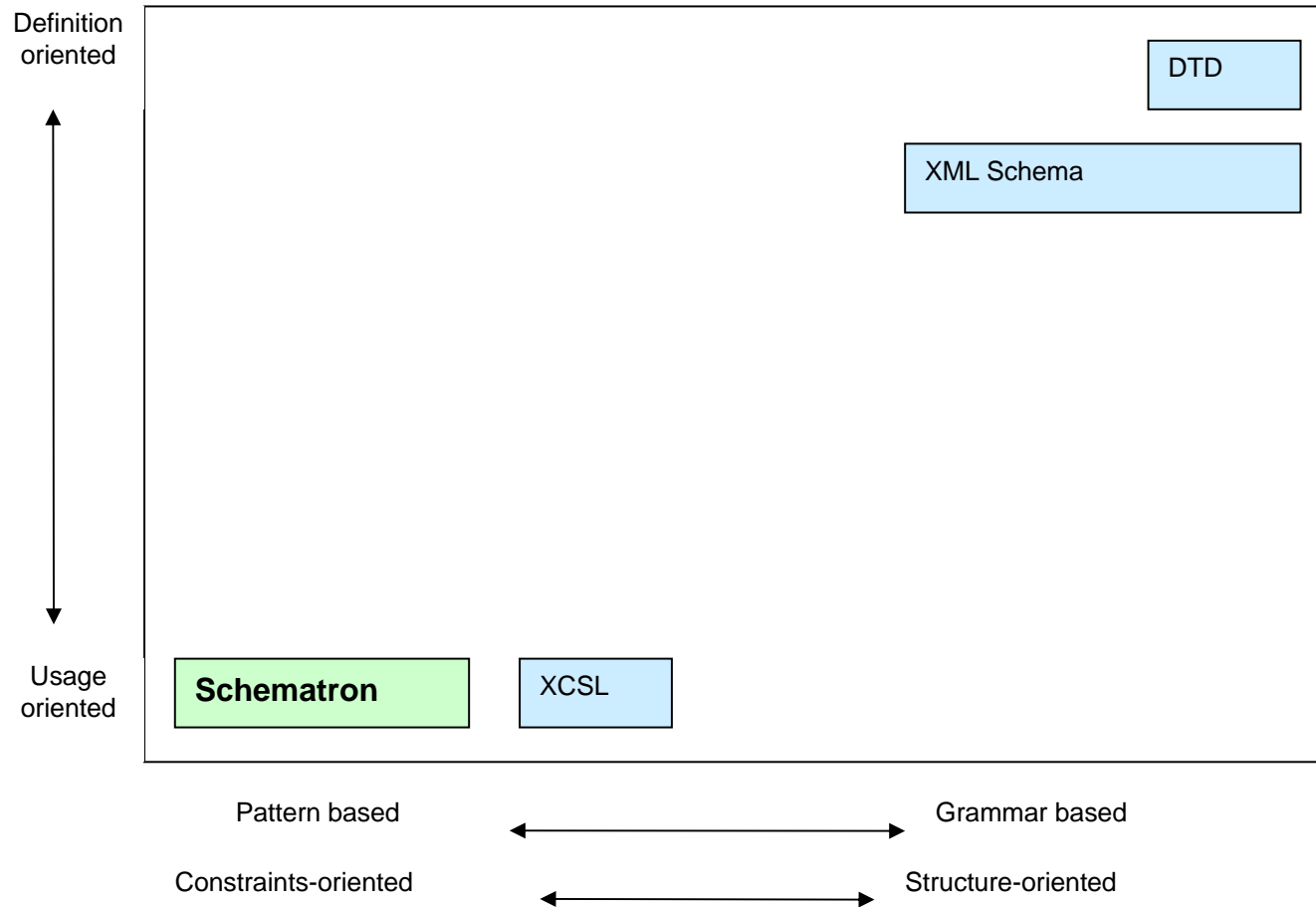
XML validation: Tools

- XML Parsers/Validators
 - Use schema definitions: XSD, RelaxNG, DTD
- XSLT
 - Use templates and XPath
- Rule or Constraint Validators
 - Use XML on top of (usually) XSLT

Constraint Validators

- Several in development
 - IBM BICS, XCSL, EAD Reportcard, Schematron
- Relationships, not definitions
 - Sibling content
 - Attribute values
 - Attributes vs. children elements
 - Datatypes
- Pattern - based
- User and usage Oriented

XML validation: Overview





Schematron: History

- Born in 1999 (or thereabouts)
- ISO Standard (ISO/IEC 19757-3:2006) May 2006
 - Part 3 of Document Schema Definition Languages (DSDL)
- Strong commercial adoption
 - Sun, Topologi, Oxygen, informal endorsement by Microsoft.

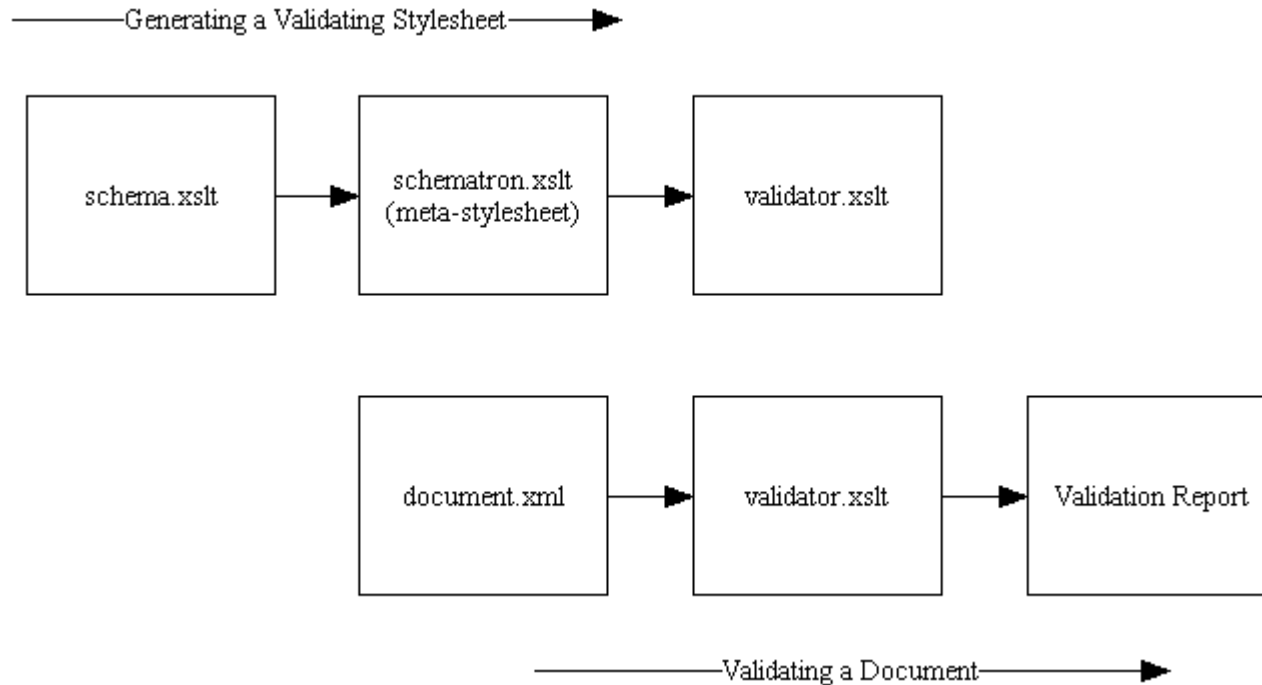


Schematron: Features

- Flexible
 - Loose or strict, partial or comprehensive
- Usable
 - XSLT is hidden, error messages are friendly, contains provisions for interface elements like icons.
- Scalable
 - Promotes generic tool development
- Expressive
 - Can leverage all of the tree axis defined in XPath



Schematron: Overview



Source: Dodds, 2001



Schematron: Schema

- Start with a guideline

Within the publication statement, the publisher must be the "Indiana University Digital Library Program."



Schematron: Schema

- Locate the context of the guideline in a rule
 - EAD:
 - <ead>
 - <eadheader>
 - <filedesc>
 - <publicationstmt>
 - XPath:
 - /ead/eadheader/filedesc/publicationstmt
 - Schematron
 - <sch:rule
 - context="/ead/eadheader/filedesc/publicationstmt">



Schematron: Schema

- Assert (test) something in the context

```
<sch:rule context="/ead/eadheader/filedesc/publicationstmt">
```

```
<sch:assert test="publisher/text()='Indiana University  
Digital Library Program'">
```

The publisher element must have the value
"Indiana University Digital Library Program".

```
</sch:assert>
```

```
</sch:rule>
```



Schematron: Schema

- Advanced features
 - Group rules into Patterns
 - Workflow, Editorial, Technical
 - Can be abstracted – for sharing across schema types
 - Diagnostic messages
 - Links to external documentation
 - Variables

DLP Schematron

- XSLT
 - reference implementation 1.5
- XML reporting language
 - customization
- Integrated into Xubmit
- Accessible via Oxygen Plugins

Demonstration



Future Directions

- Uses in the DLP
 - Content creation by partners (Lilly, Archives, Newton)
 - Content creation by vendors
 - Computer or human generated files that must conform to a conceptual model
 - METS Profile
 - Fedora Content models

Future Directions

- Reporting mechanism
 - Different views
 - summary vs. full
 - html vs. text
 - Different kinds of users
 - Content creators
 - Release engineering
 - Managers
 - Metadata librarians/Project editors

Challenges

- Readability
 - Xpath is difficult to write well.
 - Schemas are written once, used often
 - Schemas need their own guidelines
- Usability
 - Reports
 - Schematron schema creation workflow

References

Resources

- Schematron: <http://www.schematron.com/>
- Oxygen XML Editor: <http://www.oxygenxml.com>

References

Dodds, L. (2001). *Schematron: Validating xml using xslt*. Paper presented at the XSLT UK Conference. Retrieved April, November 2006 from: http://xml.coverpages.org/Dodds-schematron_xsltuk.html.

Jacinto, M., Librelotto, G. R., Ramalho, J.C.L & Henriques, P.R. (2002). Constraint specification languages: comparing XCSL, Schematron and XML-Schemas. XML Europe, 2002. Retrieved April, 2006 from: http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/03-03-02/03-03-02.pdf

Jelliffe, R. (2000). Getting Information into markup: the data model behind the schematron assertion language. A technical whitepaper for GeoTempo, Inc. October 19, 2000. Retrieved April, 2006 from: <http://www.sinica.edu.tw/~ricko/schematron.PDF>

Lee, D. & Chu, W. (2000). Comparative Analysis of Six XML Schema Languages¹. ACM SIGMOD Record 29(3), September 2000. Retrieved April, 2006 from: <http://cobase-www.cs.ucla.edu/tech-docs/dongwon/sigmod-record-00.html>