XML Data Validation

An open framework

Topics

- XML Schema Validation
- Limitations of Schema Validation
- Schematron and XSLT
- Data Validation Process
- Implementation and Tools
- Conclusion

XML Schema Validation

- Documents if the instance is a well formed XML document
- Schema Validates Data Types
- Schema Validates Data Structures (Child and Sibling relationships)

Limitations of Schema Validation

- Schema Validation cannot:
  - Attribute Constrain: If attribute X has a value, attribute Y is required.
  - Validate Logic Relations: If the parent of element A is element B, it must have an attribute Y, otherwise an attribute Z.
  - Validate Dependency: If element X has value M then Y must exist.

Limitations of Schema Validation

- Formatted string: A date must have a format of mm-dd-yyyy.
- Length Constrain: A value length must be between 9 to10.
- Multiple Ranges: Data must be in 45-50 and 100-200.
- Custom Simple Types: I.e, FacilityID

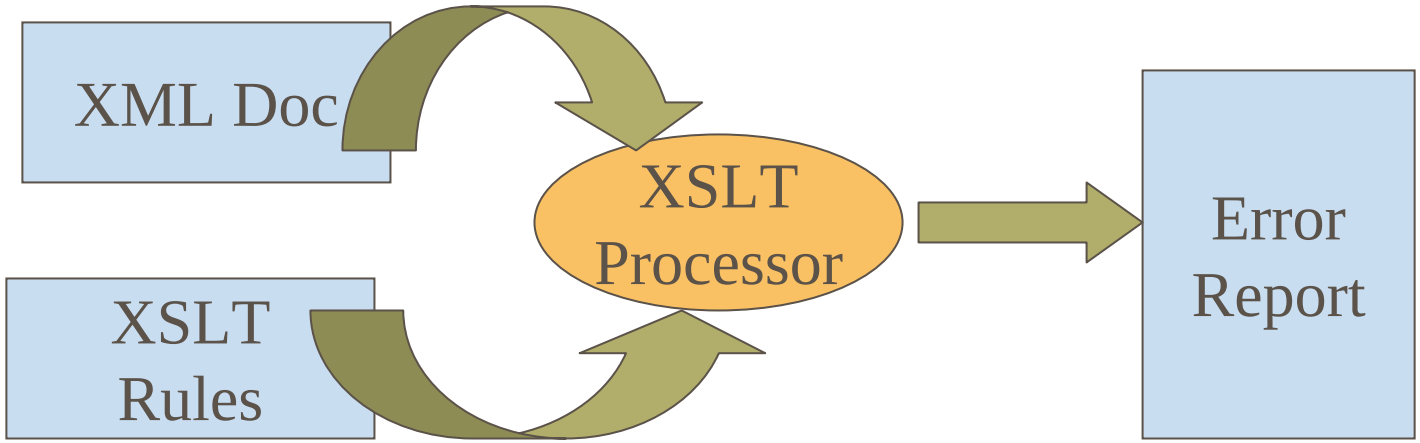
NEI Data Example

```
<TransmittalSubmissionGroup schemaVersion="3.0">
  <TransmittalRecordTypeCode>OO</TransmittalRecordTypeCode>
  <CountyStateFIPSCode>Strin</CountyStateFIPSCode>
  <OrganizationFormalName>String</OrganizationFormalName>
  <TransactionTypeCode>St</TransactionTypeCode>
  <InventoryYear>1000</InventoryYear>
  <InventoryTypeCode>String</InventoryTypeCode>
  <TransactionCreationDate>10000000</TransactionCreationDate>
  <SubmissionNumber>0</SubmissionNumber>
  <ReliabilityIndicator>0</ReliabilityIndicator>
  <TransactionComment>String</TransactionComment>
  <IndividualFullName>String</IndividualFullName>
  <TelephoneNumber>String</TelephoneNumber>
  <TelephoneNumberTypeName>String</TelephoneNumberTypeName>
  <ElectronicAddressText>String</ElectronicAddressText>
  <ElectronicAddressTypeName>String</ElectronicAddressTypeName>
  <SourceTypeCode>String</SourceTypeCode>
  <AffiliationTypeText>String</AffiliationTypeText>
  <FormatVersionNumber>0</FormatVersionNumber>
  <TribalCode>Str</TribalCode>
</TransmittalSubmissionGroup>
```

The XML segment is valid according to NEI schema. But almost all values in the record are fake and invalid.

- You really can't assure data quality using schema validation alone.

Schematron and XLST



Transform an XML document into an error report using XSLT. Rules are coded in style sheet.

Schematron

- An XML schema language
- Combine powerful validation capability with simple syntax
- Based on XSLT and XPath
- Open Source Implementation
- Currently undergoing ISO standardization

  ISO/IEC 19757 - DSDL Document Schema Definition Language

Schematron Rules

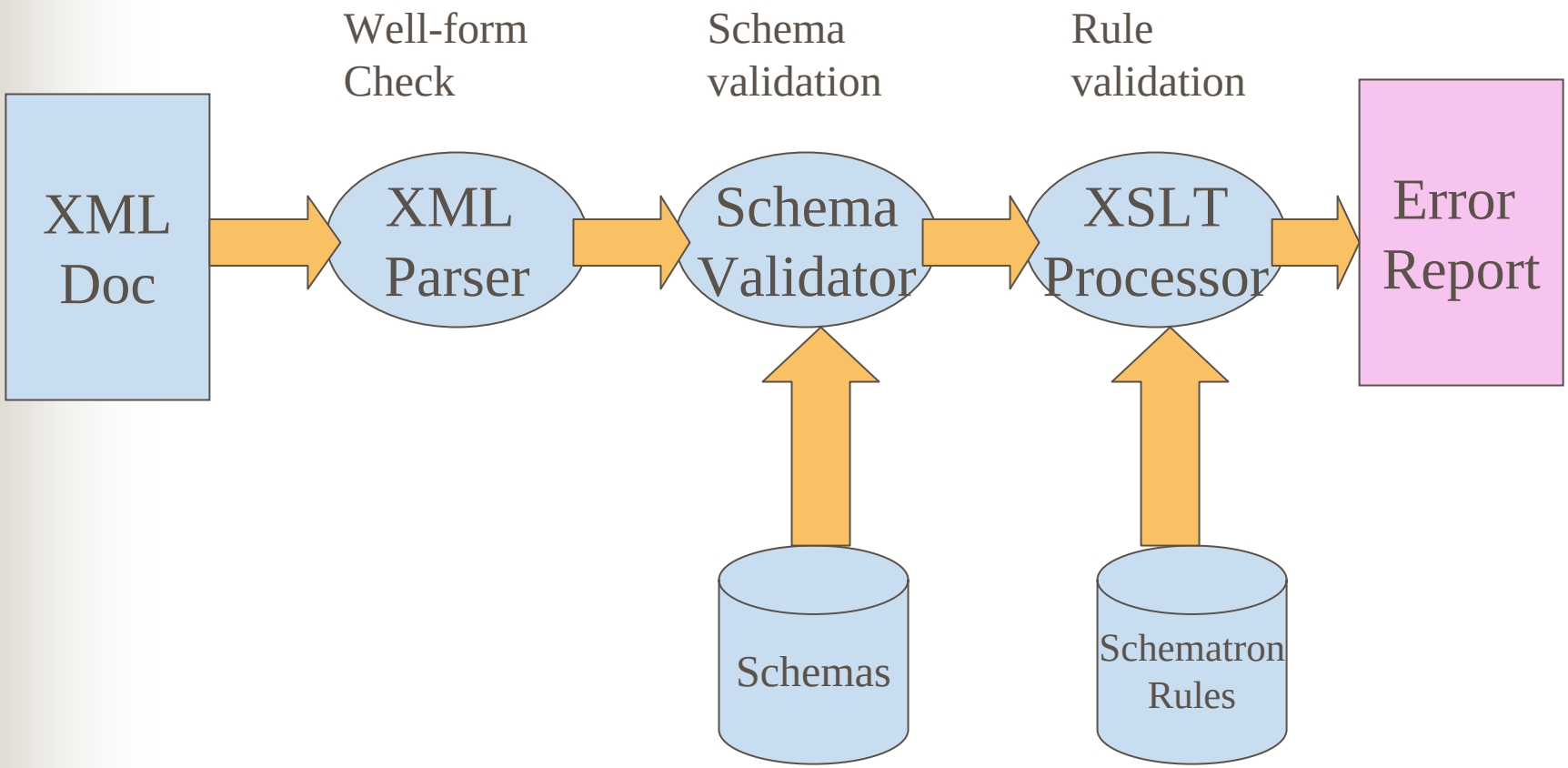
A Schematron rule has three major parts:

- The context: The element a rule applies to.
- An assertion: A statement about an element, usually an Xpath expression.
- A result: A statement to be reported if an assertion fails (or succeeds).

Schematon Rule Example

```
<sch:pattern name="Final Checks"
    id="completed">
<sch:rule context="house">
    <sch:assert test="count(wall) = 4">A
    house should have 4
    walls.</sch:assert>
</sch:rule>
</sch:pattern>
```

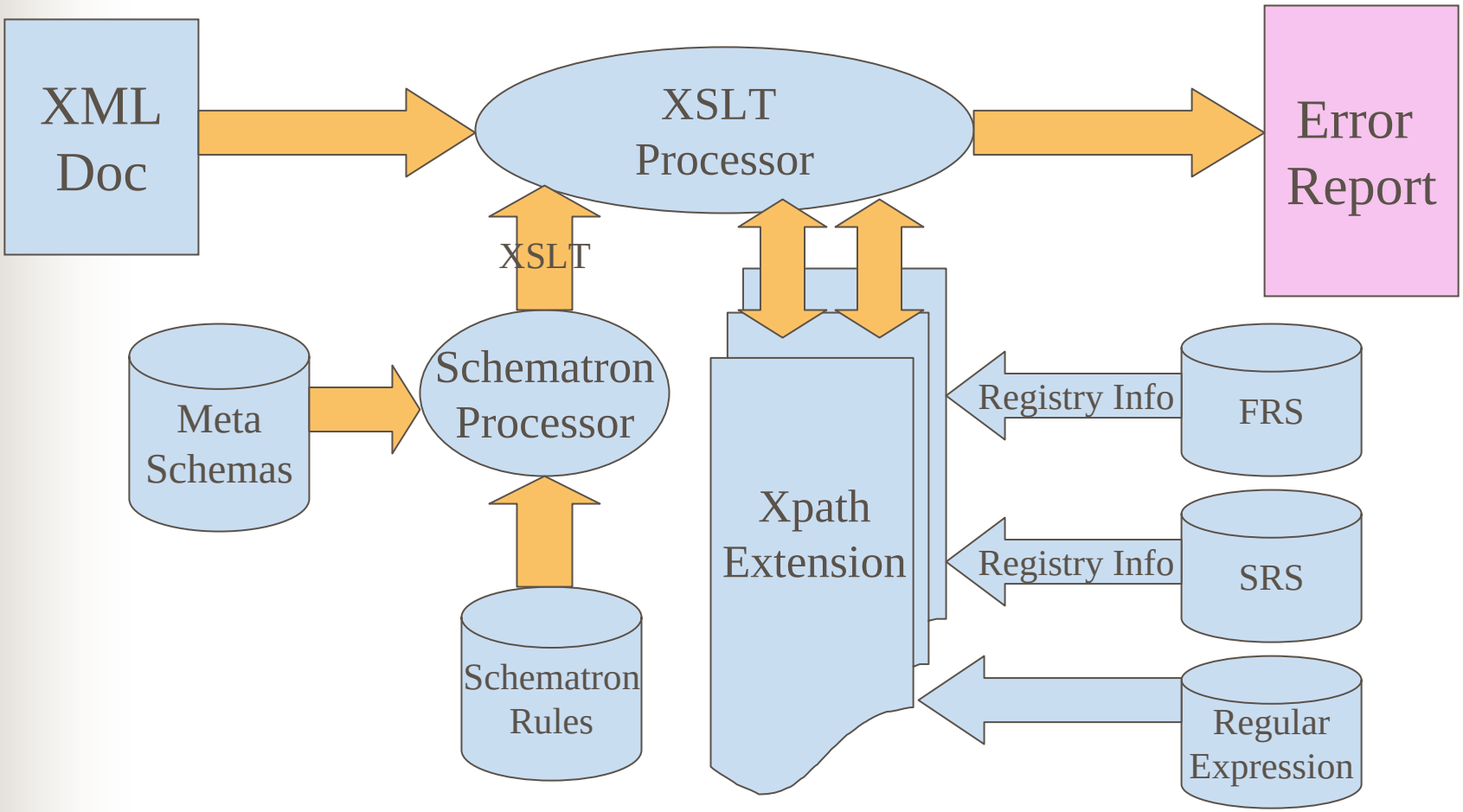
Flow Data Validation Process



Pros and Cons

- Simple rule-based XML validation framework
- Promote natural language description of errors
- Based on open standards (XSLT and Xpath)
- Open Source Schematron implementation
- Lack of regular expression support
- Custom validations against existing registries/dictionaries not available

Schematron with Extensions



Sample Schematron Rules

- Transmittal Record Type must be TR.

```
<rule context="nei:TransmittalSubmissionGroup">
     <assert test="TransmittalRecordType='TR‘”>
Transmittal must have a record type 'TR' </assert>

</rule>
```

- SourceTypeCode must be one of the values in the SourceType table.

```
<assert test="neien:CheckExist('Validate', 'select
Source_Type_Code from SourceType','SourceTypeCode',
string(nei:SourceTypeCode))">
 SourceTypeCode has a wrong value
</assert>
```

Current Implementation

- A set of web methods.
- Provides both schema validation and schematron validation.
- Has synchronous and asynchronous modes.
- Supports table lookups to any database tables.
- Can process compressed or uncompressed xml document.
- Accessible to any nodes, applications or users.

Outstanding Issues

- Schematron Development Policies:
  - Who should build and maintain Schematron rule sets for a flow?
  - Should a schema developer be required to supply a Schematron rule set before final approval of the schema?
  - Should validations such as character length go in the Schema or Schematron?
- Schematron Use Policies:
  - Should Schematron be made available only as web service, as something that runs locally, or both?
  - Should Schematron validation be required before submittal? Should CDX run Schematron on any submittals it receives? Where in the Flow does Schematron belong?
  - How do version Schematron rule sets?

Conclusion

- Streamlined data validation is crucial to successful data exchange
- Data validation should happen as early as possible
- Technologies and tools are available for boasting data quality
- Schematron is a recommended direction

Next Call Dr Node

- Wednesday, September 29th, 2:00pm EDT
- Topic: Using your Node for RCRA

Node Mentoring Contacts

- **VB.Net/MS SQL Server 2000**

Delaware Department of Natural Resources and Environmental Control

Dennis Murphy, (302) 739-3490,
dennis.murphy@state.de.us

- **Oracle 9iAS/Oracle 9i**

Maine Department of Environmental Protection
David Ellis, (207) 624-9484,
David.H.Ellis@maine.gov

- **Microsoft .NET/Oracle 8I (TEMPO)**

Mississippi Department of Environmental Quality
Melanie Morris, (601) 961-5044,
melanie_morris@deq.state.ms.us

- **Xaware/IBM DB2**

Nebraska Department of Environmental Quality
Dennis Burling, (402) 471-4214,
Dennis.Burling@NDEQ.state.ne.us

- **Microsoft BizTalk/Oracle 8i**

New Hampshire Department of Environmental Services
Chris Simmers, (603) 271-2961,
csimmers@des.state.nh.us

- **IBM WebSphere/Oracle 8i (TEMPO)**

New Mexico Environment Department
Tom McMichael, (505) 827-0260,
tom_mcmichael@nmenv.state.nm.us

- **Sybase EAServer/Oracle 9i**

Utah Department of Environmental Quality
Mark Wensel, (801) 536-4191,
mwensel@utah.gov