SanjayaLiyanage My GSOC 2010 proposal

Google Summer Of Code 2010 - Project Proposal

Project	**Implementing a streamable subset for XPointer xpointer() scheme for XInclude**
Student Name	**Sanjaya Chathuranga Liyanage**
Email	sanjayacl@gmail.com
Time Zone	**UTC+5:30(Sri Lanka)**

Abstract

Xerces2 is a XML parser written in java which allows to parse,manipulate and validate XML documents.Xerces XPointer xpointer() scheme lets users to select document fragments using the XPath expressions.When this project is completed Xerces XPointer xpointer() scheme will be supported for streamable subset as well and Xerces will allow the users to select document fragments in more efficient and easy manner and the XInclude processor will be greatly improved.In other words the objective of this project is to improve the Xerces' streaming XInclude processor so that it provides support for a streamable subset of XPointer xpointer() scheme.

Description

XPointer xpointer() scheme which is based on XPath has the ability to address strings,points and ranges according to the definitions provided in Dom 2: range.It is used with XPointer Framework for providing high level of functionality when addressing portions of XML documents.XPointer framework and XPointer Element() scheme just meet very few requirements for XInclude when considering XPointer xpointer() scheme that supports to select document fragment using the XPath expressions.Although the current xpointer() scheme supports for number of XPath expressions in an advance manner, inability of the current xpointer() scheme to address streamable subsets reduces the performance of XInclude.

When this project which will implement a streamble subset of the XPointer xpointer() scheme is completed as a result of that Xerces' streaming XInclude processor will be improved by leaps and bounds and will be very useful to the users who use Xerces as their XML parser to deal with XML documents.

I will use numbering the elements in XML document together with introducing new symbols in order to access whatever the exact position user indicate using numbers and symbols.Once this project will be completed users will be able address the XML document using numbers and symbols.

The below mentioned example shows what kind of support XInclude in Xerces will provide when the project is completed.

Let the students.xml:

<?xml version="1.0"?>

<students>

- <student id="A">
    - ..
      </student>
- <student id="B">
    - ..
      </student>
- <student id="C">
    - ..
      </student>
- ..

</students>

When the Input is:

- <?xml version="1.0"?>
- *<root xmlns:xi*="http://www.w3.org/2001/XInclude" >

- <xi:include href="students.xml" xpointer="xpointer(students/student[@id='C'])"/>

- </root>

Result with current XInclude:

<?xml version="1.0"?>

<root xmlns:xi="http://www.w3.org/2001/XInclude" >

- <student id="C">
    - ..
      </student>

</root>

But when this project is completed the same result can be obtained simply as below :

```
<?xml version="1.0" ?>

<root xmlns:xi="http://www.w3.org/2001/XInclude" >

<xi:include href="students.xml" xpointer="element(/1/3)"/>

</root>
```

I am planing to identify the nodes(including the non element nodes),points and ranges of the XML document using the numbering mechanism.

The design of this project can simply be modeled as below in three phases.

Identifying the tokens and grammar in the XPath expression
Validating the tokens and grammar in the XPath expression
Parsing the exact portion of XML document as specified in the XPath expression

When I will begin to start the development of the code I will extend the currently existing XPath class and do the necessary developments in my derived class because of two major reasons.

1. Current XPath.Scanner inner class is very powerful as base scanner in XPath class supports much more than what is used for XML schema identity constraints .
2. As the time period is limited there is no need of reimplementing the valuable methods like *parseExpression()* and I can just override them to do the minor changes.

I am going to implement this support for the XPath Nodes and XPath Axis which are subsets of XPath.So I have to replace the Axis and NodeTest inner classes in the derived class.The Tokens class have to be changed in order to support new tokens and to expand the hash table.

Scanning the expression and adding the tokens in it to the token table is done by scanExpr() method of the XPath.Scanner class and therefore I will replace the XPath.Scanner in my derived class and modify other necessary methods in a way that it supports more symbols in the XPath expressions.And also I am going to create a method *createScanner()* that returns a scanner object when called from *parseExpression()* method in my extension.

For the second phase of the design which is the validating part I think the existing Validation package can be directly reused because I do the necessary changes in XPath.Scanner so that I can reuse the existing Validation package without modifications.

For the third phase of the design I have to basically deal with the packages Parsers and Xpointer.I am going to do the changes for Validation. XML11Configuration class, Xpointer.XpointerHandler class and Xpointer.ElementSchemePointer class for supporting the parsing the portions of XML document as specified in the XPath expression.

Deliverables

1. Modified Xerces source code that implements streamable subset of XPointer xpointer() scheme for XInclude
2. Some test cases
3. Necessary documentation.

Things Done So Far

1. As I was not familier that much with xpath expressions I took time to go through the xpath subset like xpath axes,node() test,predicate functions, predicate operators,predicate number and string literals.
2. I checked out the xerces trunk.
3. Setup my development environment
4. Just had a glance on source code.

Development Schedule

26th April - 24th May	Intereact with the mentor and the community to build good relationships. Get more understanding about the current XPath capabilities and weeknesses. Start the designing part to create the class hierachy and to obtain the data flow.
24th May - 12th July	Finalize the milestone and begin to code.
12th July - 16th July	Submitting the mid term evaluations
16th July - 9th August	Begin with whatever the milestone yet to be reached. Use test cases to validate the results. Begin the documentation part.
9th August - 16th August	Scrub code,Write tests and improve documentation.
20th August	Final evaluation deadline

30th August	Submitting required code samples

Community Interaction

I subscribed to the Xerces developers list and let the community get to know that I am interesting in this project and asked their knowledge on this project.I also contacted the mentor and he helped me whenever I came across a troublesome situation.While i was dealing with the Xerces developers list I subscribed to the xsl-list@lists.mulberrytech.com mailing list as well.I asked few questions there about XPath expressions.When I am done with my draft proposal I will ask the feedbacks of community.I am looking foraward to interactively engage with the community in the future and try my best to be a practive member in the list.

About Me

I am Sanjaya Chathuranga Liyanage and I am an undergraduate student of Department of Computer Science and Engineering,University of Moratuwa,Sri Lanka.One of the project I did for Open source(An extension for Anjuta IDE) in my leisure time really make me attracted towards the open source.I contacted the Anjuta group in IRC chat when I was in troublesome situations.I was actually dealing much more with Windows in past.But now the situation is different and I always keen to deal with open source.I installed Ubuntu in my computer as well.

I am willing to learn new things and technologies.For learning languages which are unknown to me I used to develop a small game kind of thing in that language.I have an intension to create my own XML parser one day and I hope this project will create a great background to reach my target. On the other hand I need to be exposed to the open source community and ultimately become a committer for open source.

Resources/References

- Project idea described at https://issues.apache.org/jira/browse/XERCESJ-1428
- To gain a decent knowledge on XPath http://www.w3schools.com/
- For XPointer Framework http://www.w3.org/TR/2003/REC-xptr-framework-20030325/
- For XPointer Element http://www.w3.org/TR/2003/REC-xptr-element-20030325/
- For XPointer xpointer http://www.w3.org/TR/2002/WD-xptr-xpointer-20021219/
- For XInclude http://www.w3.org/TR/xinclude/