XCSL: XML Constraint Specification Language

Marta H. Jacinto*
Universidade do Minho, Departamento de Informática
Braga, Portugal, 4710-057
Marta.jacinto@itij.mj.pt

and

Giovani R. Librelotto†
Universidade do Minho, Departamento de Informática
Braga, Portugal, 4710-057
grl@di.uminho.pt

and

José C. Ramalho and **Pedro R. Henriques**
Universidade do Minho, Departamento de Informática
Braga, Portugal, 4710-057
{jcr, prh}@di.uminho.pt

Abstract

After being able to mark-up text and validate its structure according to a document's type specification, we may start thinking it would be natural to be able to validate some non-structural issues in the documents. This paper is to formally discuss semantic-related aspects. In that context, we introduce a domain specific language developed for such a purpose: XCSL. XCSL is not just a language, it is also a processing model. Furthermore, we discuss the general philosophy underlying the proposed approach, presenting the architecture of our semantic validation system, and we detail the respective processor. To illustrate the use of XCSL language and the subsequent processing, we present a case-study. Nowadays, we can find some other languages to restrict XML documents to those semantically valid—namely Schematron and XML-Schema. So, before concluding the paper, we compare XCSL to those approaches.

Keywords: XML, Document Semantics, XCSL, XML-Schema, Schematron, Constraint Specification.

Resumen

Tendo a possibilidade de anotar texto e validar a sua estrutura de acordo com a especificação do tipo de documento, é natural começar a pensar o quão necessário e importante seria validar aspectos não-estruturais nos documentos. O objectivo deste artigo é, precisamente, discutir formalmente aspectos relacionados com a semântica dos documentos. Nesse contexto, introduzimos uma linguagem de domínio específico desenvolvida com tal finalidade: o XCSL. XCSL não é apenas uma linguagem; é também um modelo de processamento. Discutiremos ainda a filosofia geral subjacente à abordagem proposta, apresentando a arquitectura do nosso sistema de validação semântica e detalhando o processador respectivo. Para ilustrar o uso da linguagem XCSL e o processamento subsequente, mostramos um caso de estudo. Actualmente podemos encontrar algumas outras linguagens que também podem ser usadas para restringir documentos XML — nomeadamente Schematron e XML-Schema. Antes de concluir o artigo, comparamos o XCSL com essas abordagens.

Palabras claves: XML, Semântica de Documentos, XCSL, XML-Schema, Schematron, Especificação de Restrições.

*Currently working at ITIJ - Computer Department, Ministry of Justice, Portugal
†Work is supported by a grant from CNPq - Brazil

1 Introduction

This paper has to do with the semantic specification of documents marked up in XML. To go straight to our target topic, we clearly assume that the reader is familiar with *XML and companion* for document's structuring and processing. For details in these topics we suggest the reading of [17].

XCSL, XML Constraint Specification Language, is a domain specific language with the purpose of allowing XML designers to restrict the content of XML documents. It is a simple, and small language tailored to write contextual conditions constraining the textual value of XML elements, in concrete documents (instances of some family specified by a DTD).

Basically, the language provides a set of constructors to define the actions to be taken by an XML semantic validator. These actions are triggered whenever a boolean expression, over the attributes value or the elements content, evaluates to false. These predicates, that we call *constraint* or *contextual conditions* are also written in XCSL syntax. XCSL also designates a processor that traverses the document's internal representation (the tree) and checks its correctness from a semantic point of view (remember that the structural correctness is validated by the parser that builds the tree).

Moreover, XCSL is an XML language; so it becomes possible to add restrictions to XML documents using an XML dialect. That approach offers a complete XML framework to couple with syntax and semantics to document designers. The benefits of such an approach are obvious.

Constraining the content of each document's component is a way to guarantee the preservation of its semantic value; the need for semantic specification and its implications are discussed in section 2.

To keep the approach as standard as possible, XCSL conditions are translated into the XSLT [4] pattern language; this decision enables the use of standard XSL [12] processors to implement the validator. Those topics are discussed in section 3. The details concerning the implementation of the XCSL processor are introduced in section 5.

A formal specification of the proposed language, XCSL, is provided in section 4, showing a diagram that depicts the XML-Schema [10], and listing the respective DTD; in that section, we also detail the elements introduced in the DTD. For those more familiar with grammar based language definitions, we include CFG for XCSL as well.

To illustrate the claimed positive contributions of XCSL, we introduce (in section 6) a real life case study taken from a set of official documents produced in the context of portuguese laws. We show this short example once it is expressive enough and complements the little examples that are shown along the whole document. Please notice, however, that the size of a document does not have to do with its semantic complexity — a huge document may, actually, have no semantic problems. Nevertheless, should you like to analyse more examples, they can be found in [7].

A synthesis of the paper and hints on future work are presented in the last part, section 8. Before that concluding section, we compare (in section 7) our proposal with related work; namely, Schematron [9] and XML-Schema are briefly introduced and then, through an example, we show similarities and main differences.

2 Document Semantics and Constraints

XML is aimed at document structure definition. At the present, XML, with the available DTDs, transformation languages and tools, constitute a strong and powerful specification environment. However, within this framework, it is not possible to express constraints or invariants over the elements content, as it is necessary to deal with documents static semantics.

Until now and concerning content constraining, we have just felt the need to restrict atomic element values, to check relations between elements or perform a lookup operation of some value in some database. This probably happens because other validations at higher levels are enforced by the XML parser according to a specific XML type definition (DTD or XML-Schema).

Therefore, restrictions can be classified as belonging to one of the four categories defined below:

- **Domain range checking**: This is the most common constraint. We need this type of constraint whenever we want a certain content/value to be between a pair of values (inside a certain domain). Normally, this is used when data is of date or numerical type. Example: *the price of a CD should be greater than X and less that Y.*

- **Dependencies between two elements or attributes**: We have cases in which an attribute value depends on the value of another element or attribute located in a different branch of the document tree. Example: *in portuguese language, nouns and verbs must agree in person and number.*

- **Pattern matching against a Regular Expression**: Sometimes we need to guarantee that the content follows a certain format which pattern can be defined by a regular expression. Example: *there are more than 100 formats to write dates, but in practice we must enforce one.*

- **Complex constraints**: Some constraints require a "nested loops" behavior to be specified and processed. Example: *the content of a certain element, in some context, should be unique.*

We can distinguish two completely different steps when going towards a semantic validation model:

- the definition – the syntactic part of the constraining model; the statements that express the constraints.

- the processing – the semantic part of the constraining model; the interpretation.

These two steps have different goals and correspond to different levels of difficulty at implementation time.

The definition step involves the specification of a new language or the adoption of an existent one; so we could just add extra syntax or we design a new language that could be embedded in XML or coexist outside.

The XCSL language was conceived with all the four categories enumerated in mind. Its current version, which we are focusing on in this paper, enables us to specify constraints belonging to any of these categories.

The first task in order to define a constraint specification language is to elaborate a detailed description of what we want to accomplish with that language: what kind of constraints do we want to be able to specify.

For the processing step, we need to create an engine with the capability to interpret the statements written in the above language, this is, to analyze them and evaluate the logical expression.

Static Semantics can be defined as a set of preconditions or contextual conditions over content. These contextual conditions could be checked by the XML parser during the abstract document tree building process or, if the user is using a structured editor, the built-in parser could check them during edition. This is not the approach followed in this work, because that solution requires that we remake a parser or create a new one, and our aim here is to make use of existing technology to implement our ideas, as much as possible. Our purpose is explained in the following section.

3 The XCSL approach

In [1], Ramalho wrote about his quest for an XML Constraint Specification Language (XCSL). The first questions raised were: Do we really need a new language? This new specification should be linked to elements or should be put inside the DTD?

The first idea was to specify the constraints together with the elements and attributes in the DTD. That would be a good solution if we intended to associate a constraint with an element. But we soon realized that we should associate constraints with context and not with elements: an element can appear in different contexts in a document tree and we may wish to enforce different constraints for each context. We needed a context selector like the one in query languages or style languages so the next step became the study of those languages.

Many languages were studied: XML query languages like XSLT [4], XQL [5], Element Sets [13], Lore [15], XML-GL [19], DSSSL [16], scripting languages like Perl[1] (with XML::Parser, XML-DT) and Omnimark[2].

Scripting languages were discarded because we wanted a more user-friendly language, declarative and maintaining the good characteristics of XML, like hardware and software independence.

From text review, we easily concluded that XSLT was a common subset to most of the existing query languages. Besides that, XSLT has a feature that proved very useful to specify constraints: predicates.

Choosing XSLT as one of the core components of XCSL was a good strategic decision: this enabled us to use a standard XSL processor as the constraint validator engine.

So, we have defined a set of XML elements to wrap XSLT expressions. That way we built in the language that we called XCSL.

To finish this summary we exhibit an overview of the XCSL architecture. Figure 1 shows the XCSL workflow, the process to validate documents' semantics driven by XCSL Constraint Specifications: an XCSL document, describing the constraint to be validated, is given to the XCSL Processor Generator that produces an XSL stylesheet; then, using any standard XSL processor, it is possible to apply that stylesheet to the XML instance we want to check; and as a result we obtain another XML document with the error messages (the doc-status element will be empty if the source document is semantically correct). Figure 1 also describes

[1] http://www.perl.com/
[2] http://www.omnimark.com/

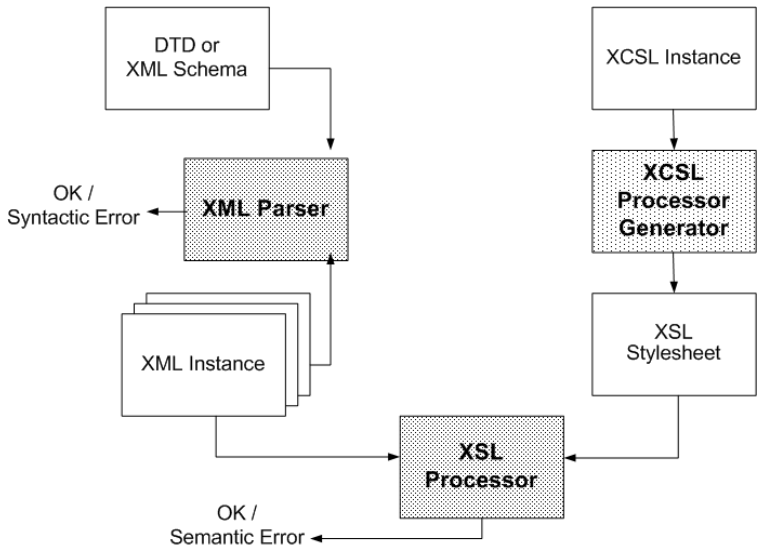

Figure 1: XCSL processing

the traditional structure validation process: the same XML instance, under check, is submitted to an XML parser with its DTD, and the Parser produces the syntactic error messages in a similar way.

The complete details about XCSL will be discussed in the following section.

4 The XCSL Language

Starting from an older idea to add semantics to SGML documents [18], the first version of the Constraint Specification Language is formally defined in [1] as it was already stated. After that very first formalization exercise (that helped us to find the core structure of the language) we refined its definition and improved the processing engine to be able to couple with all practical problems we were faced to.

A specification in XCSL is composed by one or more tuples. Each tuple has three parts [6]:

- **Context Selector**: As the name suggests, this part contains the expression (a tree path) that selects the context where we want to enforce the constraint.
- **Context Condition**: This part corresponds to the condition we want to enforce, within the selected context.
- **Action**: The third part describes the action to trigger every time the condition does not hold.

In a more formal way, we can write the CFG (Context Free Grammar) for that language:

```
ConstraintSpec ::= Header Constraint+
Constraint ::= ContextSelector ContextCondition Action
Header ::= name date version
ContextSelector ::= XPath-Exp Let*
Let ::= Name Value
Value ::= XPath-Exp
ContextCondiction ::= RegExp | XPath-Exp
Action ::= message+
```

The choice of using XSLT to specify the constraints was already justified. In order to be coherent, we needed an XML wrapper for XSLT expressions (like in XSL). So, each XCSL specification is defined as an XML instance and the XCSL language is defined by a DTD and/or an XML-Schema; the present version of the DTD that specifies XCSL (named xcsl.dtd) is shown below.

Nowadays, XML-Schema has overcome the DTD approach to the definition of classes of the XML documents. We also made that upgrade; however, as XML-Schema is much more verbose than the correspondent DTD, we decided to include here the DTD and just a diagrammatic description of the XCSL XML-Schema. This diagram is shown in Figure 2, as obtained with the XML Spy 4.1[3], from Altova.

[3] http://www.xmlspy.com/

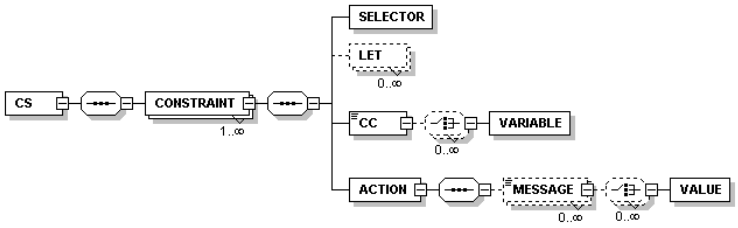

Figure 2: XCSL version 1.0 XML-Schema diagram

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- XCSL: XML Constraint Specification Language -->
    <!ELEMENT cs (constraint)+>
    <!ATTLIST cs
        dtd CDATA #IMPLIED
        date CDATA #IMPLIED
        version CDATA #IMPLIED
    >
    <!ELEMENT constraint (selector, let*, cc, action)>
    <!ELEMENT selector EMPTY>
    <!ATTLIST selector
        selexp CDATA #REQUIRED
    >
    <!ELEMENT let EMPTY>
    <!ATTLIST let
        name CDATA #REQUIRED
        value CDATA #REQUIRED
    >
    <!ELEMENT cc (#PCDATA | variable)*>
    <!ELEMENT variable EMPTY>
    <!ATTLIST variable
        selexp CDATA #REQUIRED
    >
    <!ELEMENT action (message*)>
    <!ELEMENT message (#PCDATA | value)*>
    <!ATTLIST message
        lang CDATA #IMPLIED
    >
    <!ELEMENT value EMPTY>
    <!ATTLIST value
        selexp CDATA #REQUIRED
    >
```

The main elements introduced in the XCSL DTD above are briefly described below.

cs is the root element of an XML Constraint Specification Document. It has three attributes, all of them optional: *dtd* – name of DTD that is to be associated with the constraints being specified; *date* – XCSL document's date; *version* – XCSL document's version.

An XCSL constraint document consists of one or more *constraint* elements. Each constraint element specifies a constraint and the action that will be triggered when this constraint is not respected. The constraint element is composed by a sequence of elements: a *selector* element, zero or more *let* elements, a *cc* element, and an *action* element. These elements are described separately in the following sub-sections.

4.1 Context Selector

The selector element, as the name suggests, selects the context (the element or elements) within which the conditions should be tested. It has a required attribute: *selexp* – this attribute should hold an XPath (XML Path Language[4]) expression that performs the selection.

```
<selector selexp="//iORDERS/itemsREG2"/>
```

let elements have an important role in complex constraints, they allow us to specify multi-step evaluations, enabling the simplification of very complex constraints. With a let element we can save, in a variable (*name*)

[4]http://www.w3.org/TR/xpath

the result of an XPath expression applied to any XPath path, or still the set of values that belong to that context.

let has two required attributes: *name* – it specifies the variable's name; *value* – XPath expression applied to the context or to any XPath path.

```
<let name="keyorderc" value="orderc"/>
```

selector and let are both empty elements, that is, the information is described in their attributes and the elements do not have any content (there is no text between the open and close tags).

4.2 Context Condition

cc is an element that specifies the constraint that has to be verified – regular expression or XPath function. The action will be triggered whenever that expression or function is evaluated to false. It has a mixed content – text in which *variable* elements can occur any number of times.

```
<cc> count(//iORDERS/itemsREG2[orderc=$keyorderc])=1 </cc>
```

The variable element is used when we are enforcing a constraint over an element or attribute and we want to guarantee that this element or attribute is evaluated only if present in the document instance. If the element or attribute in question is absent from the document, the *constraint* will not be evaluated. This element has an attribute; *selexp*, that specifies the XPath path in order to select it.

4.3 Action

action is a required element that specifies the messages to be returned whenever the evaluated expression (that describes the constraint under check) results in a false value. Its content is a sequence of messages.

message has mixed content: text in which the value elements may occur any number of times. It has an optional attribute; *lang*, that identifies the language the *message* is written in.

Finally, the value element is used to include the value of some element, attribute or XPath expression applied to some element/attribute of the XML document, in the message body. It is an empty element. It has a required attribute; *selexp*, that contains the XPath path to the element or attribute which value we want to show or an XPath function applied either to the present context, either to other XPath path.

```
<action>
    <message>WARNING: orderc: <value selexp="orderc"/>
             is not unique! </message>
</action>
```

5 The XCSL Processor Generator

The XCSL processor generator is the main piece in our architecture as it can be seen in Figure 1. It takes an XML instance, written according to the XCSL language, and generates an XSL stylesheet that will test the specified constraints when processed by a standard XSL processor like Saxon[5] or Xalan[6].

The first versions of the generator were coded in Perl using an XML down-translation module called XML-DT [8]. The reason was that the task is quite complex and we needed an open tool with strong text processing capabilities. XML-DT is a perl module developed to process transformations over XML documents. It has some specific built-in operators and functions, but we can still use all the functionalities available in Perl.

During the development of this generator we found some problems that had a strong impact in the final algorithm. The most important were:

Optional or non-filled elements – suppose you have specified a constraint for every element named X; suppose that element X is optional and in certain parts of the instance the user did not insert it; the system will trigger the action for every non-existent element X (the absence of an element that is part of a condition will make that condition always false).

[5] http://saxon.sourceforge.net/
[6] http://xml.apache.org/xalan-j/

Ambiguity in context selection - until now, we have just said that an XCSL specification is composed by a set of constraints; we did not say that these constraints are disjoint in terms of context; in some cases there is a certain overlap between the contexts of different conditions; this overlap will cause an error when transposed to XSL; XSL processors can only match one context at a time; there is one solution to this problem that is to run each constraint in a different mode (in XSL each mode corresponds to a different traversal of the document tree).

After some iterations and after solving those problems, the main algorithm is now:

1. Convert each *constraint* element into an *xsl:template*
2. Use attribute *selexp* for the *xsl:template match* attribute
3. Convert each "stamped" path (*variable* element) into a predicate *[...]*, inside *match* attribute
4. Convert each *let* element into an *xsl:variable*
5. Convert *cc* element into an *xsl:if* element (the *test* attribute of the *xsl:if* element is filled with the negation of *cc*'s content).
6. Put *message* contents inside template body converting: - each *value* element into *xsl:variable* and *xsl:value-of*
7. Filter all remaining text nodes

Recently we have been developing the XSL version of the generator. The main function which generates a template for each constraint looks like the following:

```
<xsl:template match="constraint">
  <xsl:variable name="cc">
    <xsl:apply-templates select="cc"/>
  </xsl:variable>
  <xsl:variable name="sel" select="selector/@selexp"/>
  <xsl:variable name="pred">
    <xsl:apply-templates select="cc/variable" mode="pred"/>
  </xsl:variable>
  <xsl:comment>
  ..................NEW CONSTRAINT.................
  </xsl:comment>
  <my:template mode="constraint{count(preceding-sibling::*)+1}"
                   match="{$sel}{$pred}">
    <my:if test="not({$cc})">
      <err-message>
        <xsl:apply-templates select="action"/>
      </err-message>
    </my:if>
  </my:template>
  <my:template match="text()" priority="-1"
              mode="constraint{count(preceding-sibling::*)+1}">
     <!-- strip characters -->
  </my:template>
</xsl:template>
```

Let us now go through a very simple example. Consider the following extract of an XML document that describes a CD in some database:

```
<?xml version="1.0"?>
<cd>
  ...
  <price>32.00</price>
  ...
</cd>
```

We want to ensure that the price of every CD is within a certain range (0 and 100). In order to do that, we specify the following constraint in XCSL:

```
<?xml version="1.0"?>
<cs>
  <constraint>
      <selector selexp="/cd/price"/>
      <cc>(.>0) and (.<100)</cc>
```

```
        <action>
            <message>Price out of range!</message>
        </action>
    </constraint>
  </cs>
```

This semantic specification document will be processed by the XCSL Processor Generator, and the following XSL stylesheet will then be generated:

```
<xsl:stylesheet version="1.0">
  <xsl:template match="/">
    <doc-status>
      <xsl:apply-templates mode="constraint1"/>
    </doc-status>
  </xsl:template>
<!--..................NEW CONSTRAINT.....................-->
  <xsl:template mode="constraint1" match="/cd/price">
    <xsl:if test="not((.&gt;0) and (.&lt;100))">
      <err-message>Price out of range!</err-message>
    </xsl:if>
  </xsl:template>

  <xsl:template match="text()" priority="-1" mode="constraint1"/>
  <xsl:template match="text()" priority="-1"/>
</xsl:stylesheet>
```

This stylesheet, when applied to an XML source document by an XSL processor, will emit error messages whenever the constraint does not evaluate to true.

6 A Case Study

In this section we show a real life case study taken from a set of examples of official documents produced in the context of Portuguese laws. That set of legal documents was chosen to make a comparative study between XCSL and alternative approaches, as it will be discussed in the next section.

6.1 Second Conference for a divorce

In Portugal, to get divorced, a couple has to deliver two documents in court. The first one is called the *first request for divorce*, and is written when they decide to state their will to get divorced, assuring that in the presence of the judge. The second one, which we are focusing on in this example, is called the *second conference requirement* and the earlier it can be submitted is 90 days after the first one. This last one is to ask for a new meeting where the couple will state they still wish to get divorced, after which the couple will be officially divorced. At the moment of designing the DTD for this family of documents, we have two options: the number of days passed since the first conference is directly stated; or just the date of the first conference is written. It makes much more sense adopting the second one. Therefore, the DTD will be:

```
<!ELEMENT div_2c (header, body, ending)>
<!ELEMENT header (sender, addressee)>
<!ELEMENT sender (#PCDATA | cdepart)*>
<!ELEMENT cdepart (#PCDATA)>
<!ELEMENT addressee (#PCDATA | court)*>
<!ELEMENT court (#PCDATA)>
<!ELEMENT body (requesters, request)>
<!ELEMENT requesters (#PCDATA | name)*>
<!ELEMENT name (#PCDATA)>
<!ELEMENT request (#PCDATA | date | article)*>
<!ELEMENT date (#PCDATA)>
<!ATTLIST date
    value CDATA "19000101"
>
<!ELEMENT article (#PCDATA)>
<!ELEMENT ending (text, place, date, signature, signature)>
<!ELEMENT place (#PCDATA)>
<!ELEMENT signature (#PCDATA)>
<!ELEMENT text (#PCDATA)>
```

The following XML instance is valid and illustrates a marked up *second conference requirement* according to Portuguese laws:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE div_2c SYSTEM "div_2c02.dtd">
<div_2c>
  <header>
    <sender>
      Action of Divorce, Proc. Ner 2001/4/99
      <cdepart>2nd Department</cdepart>
    </sender>
    <addressee>
      Most worthy doctor of Laws, Judge of the
      <court> Judicial Court of the District of
               Caldas da Rainha  </court>
    </addressee>
  </header>
  <body>
    <requesters>
      <name>MARIETA MENDES </name>
      and husband
      <name>RICARDINO MENDES</name>
    </requesters>
    <request>
      identified in the referred Action of Divorce
      official papers, having accomplished the first
      conference in the
      <date value="20010406">6th of April of 2001</date>
      and both maintaining their will to divorce, come,
      by this means to require to be convoked for the
      second conference, according to the
      <article>1423th article of the Code of Civil Law
      </article>
      in order to the definite divorce be decreed.
    </request>
  </body>
  <ending>
    <text> Ask that their request be granted </text>
    <place>Caldas da Rainha</place>
    <date value="20010506">6th of May of 2001</date>
    <signature>The first requester's Lawyer</signature>
    <signature>The second requester</signature>
  </ending>
</div_2c>
```

Structurally, the document will be validated by any standard XML parser, but that check should not be successful unless the time between the present date and the date of the first conference is greater or equal to 90 days. It is essential that the requirement for the second conference occurs at least 90 days after first conference. For that, we need to compare both the dates; out of this comparison, we shall get the number of days in which they differ. To achieve this, we may use a function [11] that receives a Gregorian date and returns the Julian day for each date. Afterwards, it is enough to subtract them to get the number of days. Finally, comparing that result with the number 90, we know exactly whether the document is semantically correct or not.

We specify this semantic constraint in XCSL as follows (we only show the *constraint* element, as the rest is trivial):

```
<constraint>
  <selector selexp="//div_2c"/>
  <let name="a" value="(floor((14-substring(ending/date/
                      @value,5,2)) div 12))"/>
  <let name="y" value="(substring(ending/date/@value,1,4)
                     + 4800 - $a)"/>
  <let name="m" value="(substring(ending/date/@value,5,2)
                     + 12 * $a - 3)"/>
  <let name="t" value="(substring(ending/date/@value,7,2)
```

```
                    + floor((153 * $m + 2) div 5) +
                    (365 * $y) + floor($y div 4) -
                    floor($y div 100) +
                    floor($y div 400) - 32045)"/>
  <let name="a2" value="(floor((14-substring(body/request/
                     date/@value,5,2)) div 12))"/>
  <let name="y2" value="(substring(body/request/date/
                     @value,1,4) + 4800 - $a2)"/>
  <let name="m2" value="(substring(body/request/date/
                     @value,5,2) + 12 * $a2 - 3)"/>
  <let name="t2" value="(substring(body/request/date/
                     @value,7,2)+floor((153*$m2+2)
                     div 5) + (365 * $y2) +
                     floor($y2 div 4) -
                     floor($y2 div 100) +
                     floor($y2 div 400) - 32045)"/>
  <cc>
    ($t - $t2) >= 90
  </cc>
  <action>
    <message lang="en">
       Only <value selexp="($t - $t2)"/> days undergone
       since the first conference...
       You will have to wait a little longer!!
    </message>
    <message lang="pt">
       Só passaram <value selexp="($t - $t2)"/> dias desde
       a primeira conferência...
       Têm que esperar mais algum tempo!!
    </message>
  </action>
</constraint>
```

We use 8 elements *let* as they provide the modularity needed to specify this constraint in less lines. The context is the root element (*div_2c*). The first four *let* elements are applied to the *ending* branch of the document's tree (where we can find the date of the requirement itself) — we use the names a, y and m for the intermediary calculations and, finally, t to keep the Julian day of that date. The second set of *let* elements is applied to the *body* branch of the document's tree (where we can find the date of the first petition) — we now use the names *a2*, *y2*, *m2* and *t2*, with the same meaning.

After defining all this variables, the *cc* element itself is as simple as subtracting *t2* out of t and comparing the result with 90.

To provide a personalized result (error message) in case of an invalid petition, we can simply use a *value* element inside *message* element. We use the variables t and *t2* defined previously, avoiding the repetition of all the code.

The XML instance we show is structurally correct but semantically not correct once the first date is the 6th of April and the second one the 6th of May, both of the year 2001. Therefore, while validating this document against the constraint document specified above, we would get the error message shown below (the default language is English and that is the language of the message shown if no other information is provided):

```
<?xml version="1.0" encoding="iso-8859-1"?>
<doc-status>
   <err-message>
       Only 30 days undergone
       since the first conference...
       You will have to wait a little longer!!
   </err-message>
</doc-status>
```

where 30 is the number of days between the two dates, generated automatically according to the *value* element specified in the constraint.

The *action* element has two *message* sub-elements, meaning that we could ask the output to be given in both (*lang="all"*) the languages or in a particular one. For the message in Portuguese, it would be enough

to specify *lang="pt"* while invoking the tool, receiving the output:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<doc-status>
   <err-message>
       Só passaram 30 dias desde
       a primeira conferência...
       Têm que esperar mais algum tempo!!
   </err-message>
</doc-status>
```

7 Related Approaches

In this section, we compare XCSL with Schematron [9] and XML-Schema [10], two other semantic specification approaches developed meanwhile with similar aims.

7.1 Constraining with Schematron

Schematron is an XML schema language that combines powerful validation capabilities with a simple syntax and implementation framework. At Schematron's design and specification time, there were several aims, from which we highlight: to promote natural language descriptions of validation failures; to allow a more human-readable answer as the validation result; aim for a short learning curve by layering on existing tools (XPath and XSLT); support workflow by providing a system which understands the phases through which a document passes in its lifecycle.

By time being, the only possibility of specifying in Schematron the constraint we specified in subsection 6.1 is writing the whole equation each time we need to use it. In Schematron we do not have the *let* statement, so we can not use intermediate variables; this is due to the fact that, in Schematron language, only XPath functions are available and the one used in XCSL, that allows variables to be instantiated, is an XSL function. The equivalent constraint in Schematron would, therefore, be,

```
<title>Request for the 2nd conference of divorce</title>
<diagnostics>
 <diagnostic id="01">
   Less than 90 days undergone since the first
   conference...
   You will have to wait a little longer!!
 </diagnostic>
</diagnostics> <pattern name="Days since the First Conference">
 <rule context="//div_2c">
  <assert test="
   (((substring(ending/date/@value,7,2)+
   floor((153*(substring(ending/date/@value,5,2)+12*
     (floor((14-(substring(ending/date/@value,5,2)))
     div 12))-3)+2) div 5)+
   (365 * (substring(ending/date/@value,1,4)+4800-
     (floor((14-(substring(ending/date/@value,5,2)))
     div 12)))))+
   floor((substring(ending/date/@value,1,4)+4800-
     (floor((14-(substring(ending/date/@value,5,2)))
     div 12))) div 4)-
   floor((substring(ending/date/@value,1,4)+4800-
     (floor((14-(substring(ending/date/@value,5,2)))
     div 12))) div 100)+
   floor((substring(ending/date/@value,1,4)+4800-
     (floor((14-(substring(ending/date/@value,5,2)))
     div 12))) div 400)-32045)
   - ...)
   &gt;= 90" diagnostics="01">
  </assert>
 </rule>
</pattern>
```

Where we put "...", we mean that all the code for the evaluation of the Julian day is repeated, now for the branch *body*. We don't repeat it to keep the example as light as possible. Notice that in this approach

we are not providing the personalized output that we did with XCSL. To do this, it would be necessary to repeat all the code listed above for the evaluation of both the dates. This is clearly a disadvantage once the number of lines we need to specify the Schematron's constraint is huge when compared with XCSL's one.

This does not mean that we can not use a *key* element in Schematron, but that it can not be used for variables. The key element and can only be used to keep a set of values in a list against which we will be able to compare other values — as we did to solve the *unicity problem* in databases.

With Schematron we can not have several output languages.

Our experiments allowed us to say that Schematron is clearly more complex than XCSL, and even if it is true that the first one has some possibilities inexistent in the last one (like documentation or the ability of entitling the whole restrictions' document), it is also true that the over all effort to learn those facilities overweighs the advantage of using them.

7.2 Constraining with XML-Schema

XML-Schema was created once the syntax of DTDs fell short of the requirements of the XML users. The aims of the *W3C XML Schema Working Group* were to create a language that would be more expressive than DTDs and written in XML Syntax. In addition, it would also allow authors to place restrictions on the elements' content and attributes' value in terms of primitive datatypes found in most languages. While using DTDs, to specify constraints, even if very simple, we need to use a constraining language (such as XCSL or Schematron) and, consequently, need two documents to completely validate an XML instance. Constraining with XML-Schema, on the other hand, means, for a particular set of constraints, using only one document to validate XML instances instead of using both a DTD and a constraint document. Unfortunately, the range of constraints we can validate with an XML-Schema is far from the set we specified above.

In the case-study of subsection 6.1, we had to decide whether to write the number of days passed since the first petition or the second conference request's date. Back then we used the date (*date*), which would not be the best choice if we were to use the XML-Schema validator; choosing to use the number of days instead, we would be able to specify the constraint by using just an XML-Schema. For that, rather than the *date* element, we have a *howmany* element, which is of type *tahowmany*:

```
<xs:simpleType name="tahowmany">
    <xs:restriction base="xs:integer">
        <xs:minInclusive value="90"/>
    </xs:restriction>
</xs:simpleType>
```

This subtype is a subset of *integer* defined, by restriction, allowing just values greater than 90 or equal. As simple as this, we can specify all Domain Range checking constraints with XML-Schemas.

With the XML-Schema we may specify three kinds of constraints: Domain Range checking, Pattern Matching against a Regular Expression and a particular case of Complex Constraints — mixed content. Nome of these could be specified with DTDs.

However, other constraints that we can specify, and check, in XCSL and Schematron, are impossible to deal with in XML-Schema — Dependencies between two elements/attributes, and every other complex constraint (for instance unicity constraints).

8 Conclusion

We have been benchmarking XCSL, Schematron and XML-Schema; its discussion is out of the scope of this paper. The results attained so far are presented and compared in another paper [7]. However, the referred experiments permit us to say that XCSL was approved: we succeeded in applying this approach to all the case studies, virtually representative of all possible cases.

It means that: on one hand, we were able to describe the constraints required by each problem in a direct, clear and simple way; on the other hand, the semantic validator could process every document successfully, that is: keeping silent when the constraints are satisfied; and detecting errors, reporting them properly, whenever the contextual conditions are broken.

XCSL can either be used as a standalone validating application or together with DTDs or XML-Schema (the most frequent option). For instance when we just want to validate a particular invariant or context condition, we do not need an XML-Schema at all and it is advisable to use XCSL as a standalone application.

Some reasons to use XCSL are:

- XCSL is XML. So we can use all the available tools to manipulate and process XCSL specifications.

- XML-Schema or DTDs are not always necessary, a simple XCSL specification can solve the problem.

With XCSL, some problems that couldn't be solved are now solvable in a simple way, nevertheless using existing technology.

9 Future work

In the future, our work will pursue in two directions:

1. The generation of Perl (XML::DT) instead of XSL. This will allow to use Perl operators in Contextual Conditions (patterns and actions will be written in Perl) making the system more expressive, powerful.

2. The reverse engineering of the whole system: trying to find a suitable abstract representation for these constraints and the whole model (probably High Order Attribute Grammars).

References

[1] Ramalho, J.C.: Anotação Estrutural de Documentos e sua Semântica. Universidade do Minho - Portugal. **1** (2000)

[2] Knuth, D.: Semantics of Context Free Languages. In Mathematical Systems Theory Journal (1968)

[3] Henriques, P.R.: Atributos e Modularidade na Especificação de Linguagens Formais. Universidade do Minho - Portugal, **1** (1992)

[4] Robie, J., Lapp, J., Sach, D.: XSL Transformations (XSLT) - version 1.0. In http://www.w3.org/TR/1998/WD-xsl-19980818 (2000)

[5] Robie, J., Lapp, J., Sach, D.: XML Query Language (XQL). In QL'98 - The Query Language Workshop (1998)

[6] Ramalho, J.C., Henriques, P.R.: Constraining Content: Specification and Processing. In XML Europe'2001, Internationales Congress Centrum (ICC), Berlin, Germany (2001)

[7] Jacinto, M.H., Librelotto, G.R., Ramalho, J.C., Henriques, P.R.: Constraint Specification Languages: comparing XCSL, Schematron and XML-Schemas. In XML Europe'2002, Barcelona, Spain (2002)

[8] Ramalho, J.C., Almeida, J.J.: XML::DT - a Perl Down Translation Module. In XML Europe'99, Granada, Espanha (1999)

[9] Dodds, L.: Schematron: Validating XML Using XSLT. In XSLT UK Conference, Keble College, Oxford, England (2001)

[10] Duckett J., Griffin, O., Mohr, S., Norton, F., Stokes-Rees, I., Willians, K., Cagle, K., Ozu, N., Tennison, J.: Professional XML Schemas. Wrox Press (2001)

[11] Tondering, C.: Frequently Asked Questions about Calendars - Version 2.3. In http://www.tondering.dk/claus/calendar.html (2000)

[12] World Wide Web Consortium: Extensible Stylesheet Language (XSL) - Version 1.0. In http://www.w3.org/TR/xsl/ (2001)

[13] World Wide Web Consortium: Element Sets: A Minimal Basis for an XML Query Engine. In http://www.w3.org/TandS/QL/QL98/pp/sets.html (1998)

[14] World Wide Web Consortium: Extensible Markup Language (XML). In http://www.w3.org/XML (2001)

[15] Abiteboul, S., Quass, D., McHugh, J., Widom, J., Wiener, J.: The Lorel Query Language for Semistructured Data. In International Journal on Digital Libraries, 1(1):68-88 (1997)

[16] Clark, J.: Document Style Semantics and Specification Language (DSSSL). In http://www.jclark.com/dsssl/ (2001)

[17] Harold, E.R., Means, W.S.: XML in a Nutshell. O'Reilly & Associates (2001)

[18] Herwijnen, E.: Practical SGML. Kluwer Academic Publishers (1994)

[19] Ceri, S., Comai, S., Damiani, E., Fraternali, P., Paraboschi, S., Tanca, L.: XML-GL: A Graphical Language for Querying and Reshaping XML Documents. In http://www.w3.org/TandS/QL/QL98/pp/xml-gl.html (1998)