jInfer BasicXSDExporter Module Description

Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, Matej Vitásek
Advisors: RNDr. Irena Mlýnková, Ph.D., Martin Nečaský, Ph.D.

Praha, 2011

Target audience: developers willing to extend jInfer, specifically hack the XSD export.

Responsible developer:	Mário Mikula
Required tokens:	none
Provided tokens:	cz.cuni.mff.ksi.jinfer.base.interfaces.inference.SchemaGenerator
Module dependencies:	Base
Public packages:	none

1 Introduction

This is an implementation of a *`SchemaGenerator`* exporting the inferred schema to XSD, supporting basic features of the language.

2 Structure

The main class implementing `SchemaGenerator` inference interface and simultaneously registered as its service provider is `SchemaGeneratorImpl` in package `cz.cuni.mff.ksi.jinfer.basicxsd`. Process of export consists of two phases described in detail in later sections.

1. Preprocessing.
2. The export to a string representation itself.

Method `start()` first creates an instance of `Preprocessor` class supplied by rules (elements) it got in the simplified grammar on input. Phase of preprocessing is done by creating that instance (calling its constructor) and its purpose is to discover information such as which elements should be globally defined and which element is the root element.

Afterwards, `start()` method uses instances of classes derived from `AbstractElementsProcessor` class to export elements of input grammar.

2.1 Preprocessing

Code to handle preprocessing resides in package `cz.cuni.mff.ksi.jinfer.basicxsd.preprocessing`.

2.1.1 Purpose

As mentioned above, preprocessing is implemented in `Preprocessor` class and its functions are following.

- Decide which elements should be defined globally.
- Remove unused elements.
- Find the top level element.
- Find an instance of element by its name.

2.1.2 TODO spravny vyraz sem

Constructor of `Preprocessor` class gets a list of elements and a number, defining minimal number of occurrences of an element to be defined globally. It first topologically sorts input elements to decide which of the elements is the root element. Afterwards, it counts occurrences of the elements and removes unused ones (those which did not occur). Finally, for each element it decides whether to mark it as a global one or not. An element is considered global if its occurrence count is greater than or equal to the number of occurrences provided on input.

2.1.3 Running preprocessor and obtaining its result

As described above, preprocessing is performed by creating an instance of `Preprocessor` class.

Information discovered by the preprocessing can be obtained by calling `Preprocessor`'s `getResult()` method. This method returns an instance of `PreprocessingResult` class. Purpose of this class is to provide of what the preprocessor has discovered and to provide an easy way to search the input grammar for an element by its name. For details see JavaDoc of `PreprocessingResult`'s public methods.

2.2 Export

XSD export itself is performed using classes derived from `AbstractElementsProcessor` class and a helper class named `Indentator`.

2.2.1 Indentation

To generate a human readable XSD output, it is necessary to apply correct indentation of XSD elements and their content. This is handled by `Indentator` class. This classes also serves as a buffer for string representation of XSD that the exporter is creating.

Instance of this class holds text appended to it and keeps indentation level state. Text can be appended without indentation (method `append()`) or indented (method `indent()`). Level of indentation can be incremented or decremented by methods `increaseIndentation()` and `decreaseIndentation()`. At the end of export, when textual representation of each element has been appended to the `Indentator`, `Indentator`'s method `toString()` will return string representation of the resulting XSD.

2.2.2 Definition of elements

Before we describe the export of elements, let's take a look on how we define elements and their attributes using XSD language.

Element is defined by XSD element `element`, specifying its name and type. Let `xs` be XMLSchema namespace.

```
<xs:element name="Person" type="...
```

Type of an element is one of following.

- XSD *built-in type*. One of types like `xs:string`, `xs:integer`, `xs:positiveInteger`, etc.

  ```
  <xs:element name="Person" type="xs:string"/>
  ```

- `simpleType`. Actually, exporter does not support any XSD features which are defined in `simpleType`. This means, that these types will not occur in result XSDs.

  ```
  <xs:element name="Person">
    <xs:simpleType>
      ...
    </xs:simpleType>
  </xs:element>
  ```

- `complexType`. This type can contain XSD element `xs:sequence` or `xs:choice`. Each of these elements can contain definitions of elements, `xs:sequences` and `xs:choices` again.

```
<xs:element name="Person">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Name" type="xs:string"/>
      <xs:element name="Surname" type="xs:string"/>
      <xs:choice>
      ...
      </xs:choice>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

An *empty element* is defined as empty `complexType`.

```
<xs:element name="EmptyElement">
  <xs:complexType>
  </xs:complexType>
</xs:element>
```

Named type of an element is defined by XSD element `simpleType` (lack of features mentioned above) or `complexType` with specified attribute `name`. Its content is exactly the same as described above. There is of course no need to define *built-in types*.

```
<xs:complexType name="PersonType">
...
</xs:complexType>
```

Element of this type can be then defined by specifying name of the type.

```
<xs:element name="Person" type="PersonType"/>
```

XSD elements `xs:element`, `xs:sequence` and `xs:choice` can have attributes `minOccurs` and `maxOccurs`. These attributes defines interval of number of instances of a particular element. Legal values of these attributes are non-negative integers.

```
<xs:element name="Person" type="PersonType" minOccurs="1" maxOccurs="3"/>
```

Default values for `minOccurs` and `maxOccurs` attributes are "1", if they are not specified. So the example above has the same meaning as the following one.

```
<xs:element name="Person" type="PersonType" maxOccurs="3"/>
```

Exporter supports types of *mixed elements*. *Mixed element* is an element that contains other elements as well as some text.

```
<mixedElement>
  some text
  <anotherElement/>
  another text
</mixedElement>
```

Mixed element type is defined as `complexType` with attribute `mixed="true"`. Definition of the element from the last example may be as following.

```
<xs:element name="mixedElement">
  <xs:complexType mixed="true">
    <xs:sequence>
      <xs:element name="anotherElement" type="..."/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

2.2.3 Definition of attributes

Attributes are defined by XSD element `xs:attribute` with attributes `name`, `type` and optional `use`. Elements `xs:attributes` have to be placed at the end of a `complexType` definition.

```
<xs:element name="Person">
  <xs:complexType>
  ...
  </xs:complexType>
  <xs:attribute name="age" type="xs:positiveInteger"/>
  <xs:attribute name="id" type="xs:string" use="required"/>
</xs:element>
```

Attribute `type` is one of a built-in types. If an attribute is obligatory, this is defined by specifying `use="required"`.

2.2.4 Export of elements

Classes to handle export of elements are in package `cz.cuni.mff.ksi.jinfer.basicxsd.elementsexporters`. Basic common logic is implemented in `AbstractElementsExporter` class. This class is supposed to be extended by classes with a particular purpose. Its constructor signature is defined as following.

```
public AbstractElementsExporter(final PreprocessingResult preprocessingResult,
                                final Indentator indentator)
```

Names of parameters are self explanatory. Parameter `preprocessingResult` is a result of preprocessing. Parameter `indentator` is an instance of `Indentator` class to be used to buffer and indent output of exporter. This instance has not to be empty. Output of exporter is appended at the end of text held by the `indentator`. This behaviour is convenient when chaining output of several elements exporters.

There are two classes extending `AbstractElementsExporter`. `GlobalElementsExporter` and `RootElementExporter`.

First, global elements are exported. For each element, its type is defined as a global type. TODO rio example. This is done by passing global elements to `processGlobalElement` method.

After global elements, others are exported by calling of method `processElement`, supplied by the top level element as its argument. Reference to a global element type is done simply by declaring element's name and type.

```
<xs:element name="Person" type="PersonType"/>
```

To sa deje rekurzivnym priechodom so startom v korenovom elemente. Elementy, ktorych typy su definovane globalne a elementy so vstavanymi typmi, su exportovane jednoducho, uvedenim ich mena a typu. Ostatne elementy su definovane na mieste. TODO rio priklady. Na toto je potrebne zavolat metodu `processElement` s korenovym elementov, ako argumentom.

Code exporting attributes is in `attributeToString()`. First thing this method does is to assess the domain of a particular atribute: this is a map indexed by attribute values containing number of occurences for each such attribute. Type definition of an attribute is generated in the `DomainUtils.getAttributeType()` method. Based on a user setting, this might decide to enumerate all possible values of this attribute using the `(a|b|c)` notation, otherwise it just returns `#CDATA`.

Attribute requiredness is assessed based on `required` metadata presence. If an attribute is not deemed required, it might have a default value: if a certain value is prominent in the attribute domain (based on user setting again), it is declared default.

2.3 Preferences

All settings provided by *`BasicDTDExporter`* are project-wide, the preferences panel is in `cz.cuni.mff.ksi.jinfer.basicdtd.properties` package. As mentioned before, it is possible to set the following.

- Maximum attribute domain size which is exported as a list of all values (`(a|b|c)` notation).
- Minimal ratio an attribute value in the domain needs to have in order to be declared default.

3 Data flow

Flow of data in this module is following.

1. `SchemaGeneratorImpl` topologically sorts elements (rules) it got on input.
2. For each element, relevant portion of DTD schema is generated.
3. String representation of the schema is returned along with the information that file extension should be "dtd".

References

[Aho96] H. Ahonen. *Generating grammars for structured documents using grammatical inference methods*. PhD thesis, Department of Computer Science, University of Helsinki, Series of Publications A, Report A-1996-4, 1996.

[Bou] Ronald Bourret. Dtd parser, version 2.0. `http://www.rpbourret.com/dtdparser/index.htm`.

[HMU01] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation (2nd Edition)*. Addison-Wesley, 2001.

[HW07] Yo-Sub Han and Derick Wood. Obtaining shorter regular expressions from finite-state automata. *Theor. Comput. Sci.*, 370(1-3):110–120, 2007.

[jun] Java universal network/graph framework. `http://jung.sourceforge.net/`.

[KMS+a] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer Architecture*.

[KMS+b] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer AutoEditor automaton visualization and editor module*.

[KMS+c] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer Base Module Description*.

[KMS+d] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer BasicDTDExporter Module Description*.

[KMS+e] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer BasicIGG Module Description*.

[KMS+f] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer BasicRuleDisplayer Module Description*.

[KMS+g] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. jinfer javadoc. `http://jinfer.sourceforge.net/javadoc`.

[KMS+h] Michal Klempa, Mário Mikula, Robert Smetana, Michal Švirec, and Matej Vitásek. *jInfer TwoStep simplifier design and implementation*.

[log] Apache log4j™. `http://logging.apache.org/log4j/`.

[loo] org.openide.util.class lookup. `http://bits.netbeans.org/dev/javadoc/org-openide-modules/org/openide/modules/doc-files/api.html`.

[mod] Module system api. `http://bits.netbeans.org/dev/javadoc/org-openide-modules/org/openide/modules/doc-files/api.html`.

[Nor] Theodore Norvell. A short introduction to regular expressions and context free grammars. `http://www.engr.mun.ca/~theo/Courses/fm/pub/context-free.pdf`.

[VMP08] Ondřej Vošta, Irena Mlýnková, and Jaroslav Pokorný. Even an ant can create an xsd. In *DASFAA'08: Proceedings of the 13th international conference on Database systems for advanced applications*, pages 35–50, Berlin, Heidelberg, 2008. Springer-Verlag.

[wik] Regular expression. `http://en.wikipedia.org/wiki/Regular_expression`.