UPMC/Licence/Info/2I013
Flowdroid – XML

Janvier 2015

1 Structure XML

Extensible Markup Language héritier de SGML (*Standard Generalize Markup Language*) est un language formel de description de *documents* normé par le W3C (*World Wide Web Consortium* – `http://www.w3.org/TR/REC-xml/`).

Les *balises* (*Markup*) du langage permettent d'annoter le contenu d'un document distinguant ainsi ses divers éléments, ce qui lui confére une *structure* (par exemple, on distingue dans un texte les titres et les paragraphes). Ces indications de structures sont destinées au traitement informatique du document. C'est donc une *donnée* pour un traitement par programme informatique.

Un usage du format XML peut être plus simplement de décrire une *structure de données*. On retrouve d'ailleurs en XML les structures de données classiques en programmation : les listes et les arbres.

XML et les bons parenthésages Le principe générique de structuration d'un document XML est le *bon parenthèsage*.

Un bon parenthèsage est une suite judicieusement ordonnée de parenthèses ouvrantes et de parenthèses fermantes. On peut facilement en donner une définition *récursive* :

– la suite vide est un bon parenthèsage ;
– si p est un bon parenthèsage, alors `(`p`)` est un bon parenthèsage ;
– si p_1 et p_2 sont des bons parenthèsages, alors la suite p_1p_2 est un bon parenthèsage.

On peut utiliser n'importe quelle paire de symboles pour fabriquer des bons parenthèsages, par exemple `[` et `]` ou `begin` et `end`. Ce qui importe, c'est de respecter l'imbrication d'ouvrantes et de fermantes induit par la définition récursive ci-dessus.

XML utilise des *balises* qui sont des suites de caractères délimitées par des *chevrons* : les symboles `<` et `>`. Pour distinguer une balise fermante d'une balise ouvrante, on utilise le symbole `/` que l'on place directement après le `<`. Si n est un *nom* (suite de caractères répondant aux critères du *Name* du W3C) alors les suites `<`n`>` et `</`n`>` forment une paire de balises, respectivement ouvrante et fermante.

Balises XML Une basile ouvrante peut être plus complexe qu'un simple nom placé entre chevrons. Il est possible d'enrichir une balise ouvrante d'*attributs*. On les donne sous forme d'une liste de noms (d'attributs) auxquels on peut associer une valeur. Le symbole `=` est placé entre le nom d'un attribut et la valeur qu'on lui donne. La valeur est souvent mises entre guillements (symbole `"`). Une paire de balises ouvrante et fermante prend alors la forme `<`n n_1`="`v_1`"` $\ldots$ n_k`="`v_k`">` et `</`n`>`, Où n est le nom de la balise, $n_1 \ldots n_k$ sont les noms des attributs et $v_1 \ldots v_k$ dénotent leur valeur.

Il est possible qu'une balise soit à la fois ouvrante et fermante. Dans ce cas, le chevron `>` est précédé de `/`. On aura, par exemple, `<`n n_1`="`v_1`"` $\ldots$ n_k`="`v_k`"/>`.

XML et arbres La structure que donne à un document un balisage XML est une structure d'*arbre général*. L'imbrication des balises donne les niveaux des éléments correspondant dans l'arborescence.

2 Une API JAVA pour XML

Il y a deux étapes dans le traitement par programme d'une donnée XML

1. analyser le fichier texte XML pour construire sa représentation arborescente en mémoire ;
2. exploiter la structure arborescente pour en extraire les informations voulues.

Il existe en JAVA des bibliothèques permettant de traiter les documents textes XML et d'explorer leurs structures. Elles fournissent des fonctionnalités d'analyse du texte XML pour construire une représentation équivalente en mémoire. Cette représentation répond aux éléments de spécification du DOM (*Document Object Model*). Le DOM est également un standard du W3C (`http://www.w3.org/DOM/`). Ce standard définit une API qui offre des primitives de navigation et d'extarction de données pour la structutre arborescente du DOM.

Analyse du fichier XML On utilise les fonctionnalités des bibliothèques `javax.xml.parsers.DocumentBuilderFactory` et `javax.xml.parsers.DocumentBuilder` pour créer un analyseur qui construira une instance de `Document` (interface fournie par `org.w3c.dom.Document`).

Exemple de méthode de lecture d'un document XLM

```
public Document readXMLFile(String fname) {
  try {
    DocumentBuilderFactory factory = DocumentBuilderFactory.newInstance();
    DocumentBuilder builder = factory.newDocumentBuilder();
    return builder.parse(fname);
  } catch (Exception ex) {
    return null;
  }
}
```

La méthode `parse` de `DocumentBuilder` peut également prendre en argument un `InputStream`.

Le DOM et son exploitation Le DOM est une structure arborescente de *nœuds* qui peuvent être des *éléments*, un simple bloc de *texte* ou d'autres choses encore. Les *éléments* peuvent à leur tour contenir d'autres éléments, blocs de texte ou autre. Les blocs de texte sont considérés comme des *feuilles* de la structure arborescente. Un *document* DOM contient un *élément racine* unique.

Lorsqu'un DOM provient de l'analyse d'un flux XML, les *éléments* correspondent aux balises et les blocs de textes au texte simple que l'on trouve entre les balises.

Les fonctionnalités d'exploitation de la structure DOM sont fournies par les interfaces données par `org.w3c.dom`. On pourra insérer

```
import org.w3c.dom.*;
```

dans le code chargé de l'exploitation du DOM : on se servira à peu près de tout. On y trouve, en particulier, les interfaces `Document` et `Element` qui implémentent toute deux l'interface `Node`.

Si `theDoc` est un `Document`, alors `theDoc.getDocumentElement()` permet de récupérer l'élément racine. Le résulat est une instance de `Element`.

Si `node` est un `Node` alors `node.getChildNodes()` donne la liste des instances de `Node` qu'il contient. Cette liste est une instance de `NodeList`.

Ceci vaut également si `node` est une instance de `Element`. Dans ce cas, on a également que `node.getNodeName()` donne le nom de l'élément (i.e. du point de vue XML : le nom de la balise) et que `node.getAttributes()` donne la liste des couples d'attributs et valeurs associés à cet élément. Le résultat obtenu est une instance de `NamedNodeMap`.

Si `nodes` est un `NodeList`, alors `nodes.getLength()` donne le nombre d'objets que contient `nodes` ; si `i` est un entier (compris entre 0 et `nodes.getLength()`, strictement) alors `nodes.item(i)` donne le `i`ème élément de la liste – instance de `Node`.

Si `attrs` est un `NamedNodeMap`, alors `attrs.getLength()` donne le nombre d'objets contenus dans `attrs` ; si `i` est un entier (compris entre 0 et `attrs.getLength()`, strictement) et si `name` est une chaîne de caractères, alors `attrs.getNamedItem(name)` donne l'association nom-valeur de l'attribut `name`. C'est une instance de `Node`. Pour obtenir la valeur de l'attribut, on peut forcer le type du résultat vers `Attr`. Si `attr` est un `Attr` alors `attr.getValue()` donne la valeur de l'attribut comme une chaîne de caractères (instance de `String`).

ATTENTION : `getChildNodes` donne TOUS les nœuds contenus dans un élément. En particulier, les tabulations, les espaces ou les passages à la ligne peuvent constituer de tels nœuds. Il faut avoir cela en tête lorsque nous explorerons les DOM construits à partir de nos fichiers XML.