Documents structurés

Cours 1

Nassim ZELLAL

2020/2021

Documents structurés, documents semi-structurés et documents non structurés

- Un document non structuré est aussi appelé « plat ». Il n'intègre aucune marque explicite d'élément de structure. C'est une suite de caractères (plein texte/texte brut).
- Un document semi-structuré contient des éléments de structure partiels, e.g., un document au format Lexico3 ne contient que des balises ouvrantes (<xxx>….).
- Un document structuré contient, entre autres, des éléments de structure complets, e.g., un document XML (Extensible Markup Language) <xxx>donnée</xxx> = élément.
- Certains considèrent qu'un document structuré peut également contenir d'autres types d'éléments de structure, comme des virgules ou des tabulations, e.g., un fichier CSV (Comma-separated values) ou un fichier TSV (Tab-separated values).

Le format XML (Extensible Markup Language)

- XML (Extensible Markup Language) permet de produire des documents structurés.
- XML est un métalangage qui permet de structurer de l'information textuelle.
- XML est un langage à balises ("markup language") utilisé pour transférer des données sur le web.
- En XML, un article aura un titre, un auteur, une date, des chapitres, des section à l'intérieur des chapitres, des paragraphes à l'intérieur des sections.
- Les balises permettent de structurer et d'organiser les données.
- Les balises sont appelées « métadonnées-objets ». Elles donnent des informations sur ce qui constitue le document, sur les « objets » qui le composent.
- Dans un document XML, la structure logique composée de balises et les données qu'elles encapsulent est appelée « élément » (cf. infra, Voiture.xml).

Versions du langage XML

- **XML 1.0** : c'est la version publiée par le W3C (World Wide Web Consortium) en 1998, c'est la version la plus rependue du langage XML.
- **XML 1.1** : c'est la version publiée par le W3C en 2004. Elle apporte, entre autres, des améliorations dans la gestion et le support de différentes versions d'UNICODE.

Où trouve-t-on du XML ?

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<w:document xmlns:ve="http://schemas.openxmlformats.org/markup-compatibility/2006" xmlns:o="urn:schemas-microsoft-com:office:office" xmlns:r="
http://schemas.openxmlformats.org/officeDocument/2006/relationships" xmlns:m="http://schemas.openxmlformats.org/officeDocument/2006/math" xmlns:v="urn:schemas-microsoft-com:vml" xmlns:wp=
"http://schemas.openxmlformats.org/drawingml/2006/wordprocessingDrawing" xmlns:w10="urn:schemas-microsoft-com:office:word" xmlns:w="
http://schemas.openxmlformats.org/wordprocessingml/2006/main" xmlns:wne="http://schemas.microsoft.com/office/word/2006/wordml">
    <w:body>
        <w:p w:rsidR="00612783" w:rsidRDefault="005D6C91">
            <w:r>
                <w:t>Ceci est un document structuré.</w:t>
            </w:r>
        </w:p>
        <w:sectPr w:rsidR="00612783" w:rsidSect="00612783">
            <w:pgSz w:w="11906" w:h="16838"/>
            <w:pgMar w:top="1417" w:right="1417" w:bottom="1417" w:left="1417" w:header="708" w:footer="708" w:gutter="0"/>
            <w:cols w:space="708"/>
            <w:docGrid w:linePitch="360"/>
        </w:sectPr>
    </w:body>
</w:document>
```

Document WORD
.docx

Où trouve-t-on du XML ?

boitedialog.fxml

```
<?xml version="1.0" encoding="UTF-8"?>
<?import java.lang.*?>
<?import java.util.*?>
<?import javafx.geometry.*?>
<?import javafx.scene.control.*?>
<?import javafx.scene.image.*?>
<?import javafx.scene.layout.*?>
<?import javafx.scene.paint.*?>
<?import javafx.scene.text.*?>
<GridPane hgap="14.0" maxHeight="+Infinity" maxWidth="+Infinity" minHeight="-Infinity" minWidth="-Infinity" vgap="20.0" xmlns="
http://javafx.com/javafx/8.0.40" xmlns:fx="http://javafx.com/fxml/1" fx:controller="test.TestController">
  <children>
    <ImageView fitHeight="60.0" fitWidth="60.0" pickOnBounds="true" preserveRatio="true" GridPane.columnIndex="0" GridPane.halignment=
    "CENTER" GridPane.rowIndex="0" GridPane.valignment="TOP">
      <image>
        <!-- place holder -->
      </image>
    </ImageView>
    <VBox maxHeight="+Infinity" maxWidth="+Infinity" minHeight="-Infinity" prefWidth="400.0" spacing="7.0" GridPane.columnIndex="1"
    GridPane.rowIndex="0">
      <children>
        <Label fx:id="messageLabel" text="message" textAlignment="LEFT" wrapText="true">
```

JAVAFX

Où trouve-t-on du XML ?

activity_main.xml

```xml
<?xml version="1.0" encoding="utf-8"?>
<RelativeLayout xmlns:android="http://schemas.android.com/apk/res/android"
    android:layout_width="match_parent"
    android:layout_height="match_parent"
    android:background="#434343">
    <EditText

        android:layout_width="wrap_content"

        android:layout_height="wrap_content"

        android:id="@+id/editText"

        android:layout_marginTop="44dp"

        android:layout_alignParentTop="true"

        android:layout_alignParentLeft="true"

        android:layout_alignRight="@+id/button"

        android:inputType="text"

        android:textColor="#ffffff" />
```

ANDROID

Où trouve-t-on du XML ?

pizza.owl

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://www.co-ode.org/ontologies/pizza/pizza.owl#"
     xml:base="http://www.co-ode.org/ontologies/pizza/pizza.owl"
     xmlns:pizza="http://www.co-ode.org/ontologies/pizza/pizza.owl#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:terms="http://purl.org/dc/terms/"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:skos="http://www.w3.org/2004/02/skos/core#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:dc="http://purl.org/dc/elements/1.1/">
    <owl:Ontology rdf:about="http://www.co-ode.org/ontologies/pizza">
        <owl:versionIRI rdf:resource="http://www.co-ode.org/ontologies/pizza/2.0.0"/>
        <dc:title xml:lang="en">pizza</dc:title>
        <terms:contributor>Nick Drummond</terms:contributor>
        <terms:license rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Creative Commons Attribution 3.0 (CC BY 3.0)</terms:license>
        <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">pizza</rdfs:label>
```

ONTOLOGIE (OWL)

Où trouve-t-on du XML ?

calculatrice.cbp

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<CodeBlocks_project_file>
    <FileVersion major="1" minor="6" />
    <Project>
        <Option title="calculatrice" />
        <Option pch_mode="2" />
        <Option compiler="gcc" />
        <Build>
            <Target title="Debug">
                <Option output="bin/Debug/calculatrice" prefix_auto="1" extension_auto="1" />
                <Option object_output="obj/Debug/" />
                <Option type="1" />
                <Option compiler="gcc" />
                <Compiler>
                    <Add option="-g" />
                </Compiler>
            </Target>
            <Target title="Release">
                <Option output="bin/Release/calculatrice" prefix_auto="1" extension_auto="1" />
                <Option object_output="obj/Release/" />
                <Option type="1" />
                <Option compiler="gcc" />
                <Compiler>
                    <Add option="-O2" />
                </Compiler>
                <Linker>
```

CODE::BLOCKS

La syntaxe d’un document XML

- XML est un langage strict. Un document XML doit impérativement respecter la syntaxe du XML. On dira alors que le document est "bien formé" (Well-formed). Seuls les documents "bien formés" seront affichés correctement.

Écrire les balises et les attributs en minuscules

- <Document> …</Document> **!=** <document> …</document>
- Écrire : <Document> …</Document> au lieu de <Document> …</document>
- <adresse pays="Portugal"/> **!=** <adresse PAYS="Portugal"/>

- Le XML est sensible à la casse (case sensitive).
- Pour éviter les erreurs, on a tendance à écrire les balises et les attributs en minuscules.

Toute balise ouverte doit impérativement être fermée

-
  ...
  ...

Les balises doivent être correctement imbriquées

- Écrire :
- <e>…</e>

- au lieu de

- <e>… </e>

Tout document XML doit comporter une racine

-
  <e>
  <p_e> ... </p_e>
  </e>

Tous les attributs doivent avoir une valeur d'attribut

- <date anniversaire="071185">
- Au lieu de
- <date anniversaire>

Les valeurs des attributs doivent toujours être mises entre des guillemets

- <date anniversaire="071185">
- Au lieu de
- <date anniversaire=071185>

Les balises uniques doivent également comporter un slash / de fin

- **<document />**
- **<adresse />**
- **<personne/>**

Les commentaires en XML

- <!-- ……. -->

Voiture.xml

```
<?xml version="1.0"?>
<Voiture marque="Renault" modèle="Safrane">
<Carosserie couleur="rouge">
<Capot>Un peu cabossé</Capot>
</Carosserie>
<Moteur>
<!-- Ceci est un document XML -->
<Cylindres />
<Allumage>Défectueux</Allumage>
</Moteur>
<Transmission type="automatique" nb_vitesses="5">
<Boîte />
<TrainAV />
<TrainAR />
</Transmission>
</Voiture>
```

Éditeurs XML

- XMLCooktop (Cooktop)
- **XML Copy Editor**
- **Exchanger XML Editor**
- Notepad++ (XML Tools)

Mon courriel

zellal.nassim@gmail.com