

eXtensible Markup Language

# INTRODUCTION À XML

# Introduction à XML

- LE format d'échange sur le web...
- ... de documents et données structurés
- Galaxie XML:
  - Un ensemble de langages complémentaires:  
XML Schema, XLink, XPath, XSL, XQuery...
  - Boîte à outils de l'information
    - Création, mise en forme, utilisation de langages de balisage
- Simplicité & syntaxe stricte

# Balisage XML

- **Balisage**: information ajoutée à un document pour en améliorer la compréhension
- **Langage de balisage**: ensemble de symboles dans le corps du texte qui en délimitent et étiquettent les parties
- Primordial pour le **traitement automatique** de documents  
Le balisage détermine les limites et les fonctions des différentes parties d'un texte
- **Valeur informative** = Contenu + Balisage

# Document XML

- Composé **d'éléments** imbriqués les uns dans les autres pour structurer et étiqueter le contenu
- **Structure arborescente**, l'élément document (racine) contient tous les autres
- **Structure logique** versus physique: un document XML peut être composé de plusieurs fichiers

# Exemple 1

```
<book>
  <head>
    <title>Introduction à XML</title>
    <author>Erik T. Ray</author>
  </head>
  <preface>Depuis son introduction...</preface>
  <chapter>
    <title>Introduction</title>
    <section>XML est une boîte à...</section>
    ...
  </chapter>
  <chapter>...</chapter>
</book>
```

# Example 2

```
<bibliography>
  <book id='x223'>
    <author>
      <firstname>David</firstname>
      <lastname>Lodge</lastname>
    </author>
    <title>Small World</title>
    <publisher>Penguin Books</publisher>
    <year>1995</year>
  </book>
  ...
</bibliography>
```

# Example 3

```
<defclass name="Car">  
  <defattribute name="mark" type="string"/>  
  <defattribute name="age" type="integer"/>  
</defclass>
```

```
<defobject id='123'>  
  <class>Car</class>  
  <attribute name='mark'>Renault</attribute>  
  <attribute name='age'>1992</attribute>  
</defobject>
```

# Exemple 4

```
<rdf:RDF xml:base="http://rdf.insee.fr/geo/2011/"
  xmlns:geo="http://rdf.insee.fr/geo/" xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdf:Description rdf:about="http://rdf.insee.fr/geo/2011/regions-2011.rdf">
  <dc:title xml:lang="fr">Régions et départements de France</dc:title>
  <dc:date>2011-02-21</dc:date>
  <dc:publisher>INSEE</dc:publisher>
</rdf:Description>

<geo:Pays rdf:about="PAYS_FR">
  <geo:code_ISO>FR</geo:code_ISO>
  <geo:nom xml:lang="fr">France</geo:nom>
  <geo:subdivision>
    <geo:Region rdf:about="REG_93">
      <geo:code_region>93</geo:code_region>
      <geo:nom xml:lang="fr">Provence-Alpes-Côte d'Azur</geo:nom>
      <geo:chef-lieu>
        <geo:Commune rdf:about="COM_13055">
          <geo:code_commune>13055</geo:code_commune>
          <geo:nom xml:lang="fr">Marseille</geo:nom>
        </geo:Commune>
      </geo:chef-lieu>
      ...
    </geo:Region>
  </geo:subdivision>
  ...
</geo:Pays>
</rdf:RDF>
```



# Modèle de documents

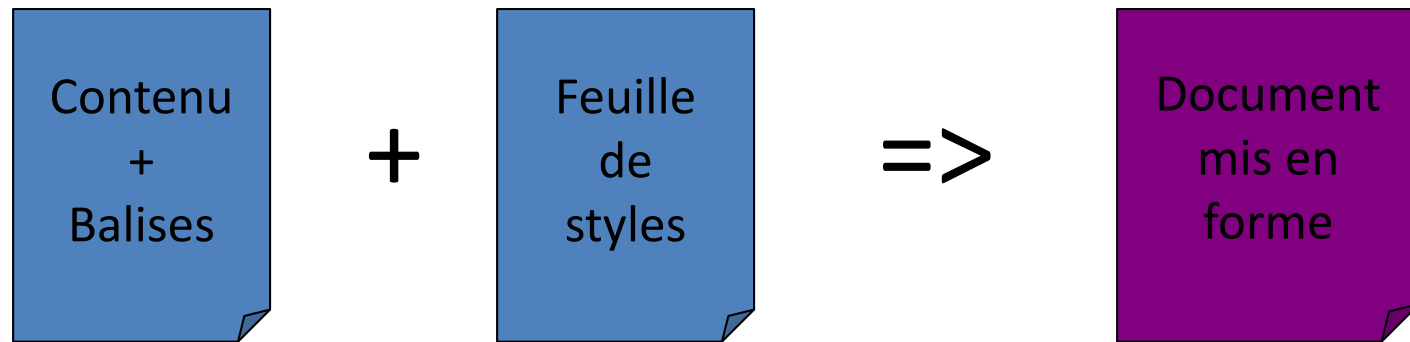
- Document **bien formé**: respecte la syntaxe XML
- Document **valide**: respecte un modèle, les règles d'un langage de balisage (vocabulaire & grammaire)
  - DTD: Document Type Definition
  - XML Schema
- **Application XML** ou Type de document:  
Langage de balisage qui respecte les règles syntaxiques de XML:  
XHTML, MathML, SVG, RDF, RDFS, OWL, etc.
- XML: (**méta**)**langage** de définition de langages

# Modèle de documents

- Feuilles de style CSS pour HTML
- Styles dans Word, galerie de styles
- Styles dans Powerpoint, galeries de styles
- Modèles de documents dans Latex
- ...

# Séparation du fond et de la forme

- Le fond : document XML : contenu + balisage
- La forme de présentation : Feuille de style



# Processeurs XML

Lisent et traitent du XML

- **Parser:** analyser syntaxiquement  
flux de caractères → flux d'atomes → arbre d'objets
- **Valider** (par rapport à un modèle de document)
- Créer, Visualiser
- **Transformer**
- Interroger

# Processeurs XML

- Création: éditeurs de texte, éditeurs dédiés
- Visualisation (avec CSS): IE, Mozilla
- Parsing & validation: JAXP, Xerces, XP
- Transformation: moteurs XT, Xalan, java 1.4
- DOM Document Object Model (W3C)
  - API d'accès aux documents et aux données XML
- SAX Simple API for XML
  - API dirigée par les événements
- Interrogation: XQuery

# CSS et XML

- Pas de display par défaut associés aux éléments d'un document XML
- Déclaration du display des éléments XML *inline* ou *block*
- Association d'une feuille de style CSS à un document XML:  
`<?xml-stylesheet type="text/css" href="mystyle.css"?>`

[http://www.w3schools.com/xml/xml\\_display.asp](http://www.w3schools.com/xml/xml_display.asp)

# **XML IN A NUTSHELL**

# Anatomie d'un document XML

```
<bibliography>
  <book id='x223'>
    <author>
      <firstname>David</firstname>
      <lastname>Lodge</lastname>
    </author>
    <title>Small World</title>
    <publisher>Penguin Book</publisher>
    <year>1995</year>
  </book>
  ...
</bibliography>
```



# Anatomie d'un document XML

## Imbrication d'éléments

<bibliography>

<book id='x223'>

<author>

<firstname>David</firstname>

<lastname>Lodge</lastname>

</author>

<title>Small World</title>

<publisher>Penguin Book</publisher>

<year>1995</year>

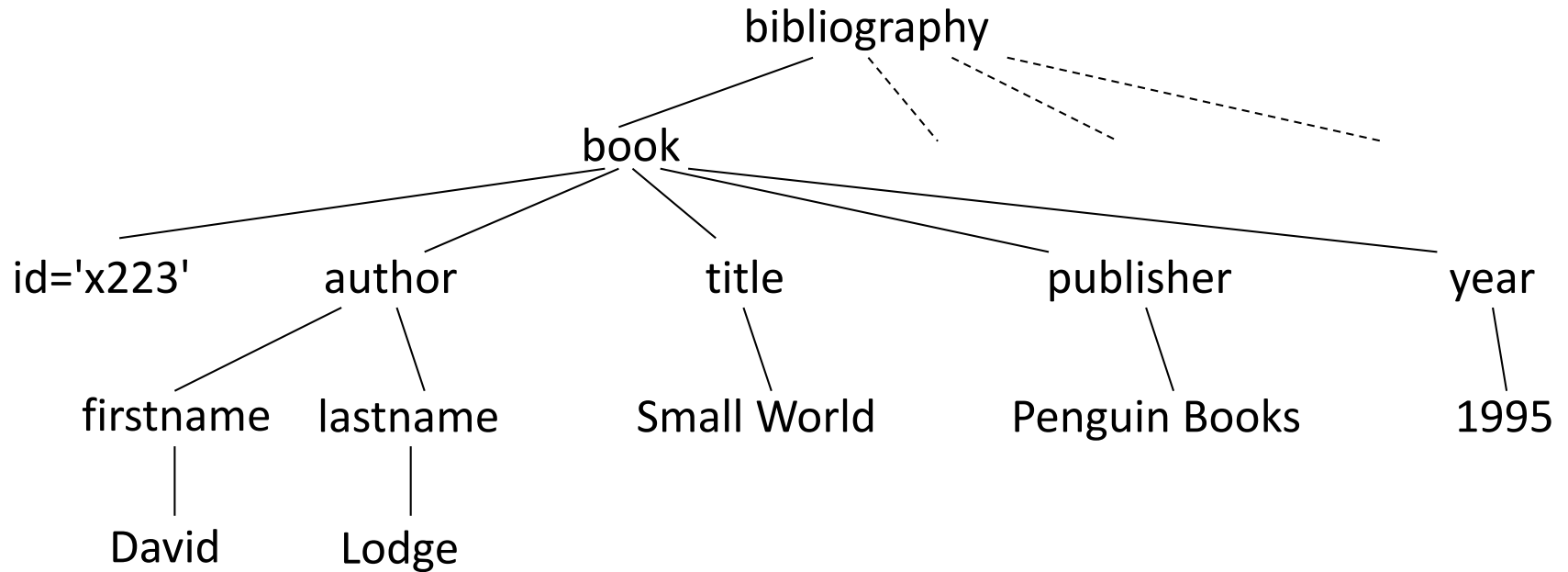
</book>

...

</bibliography>

# Arbre XML

- Feuilles: contenu ou attributs
- Nœuds internes: balises



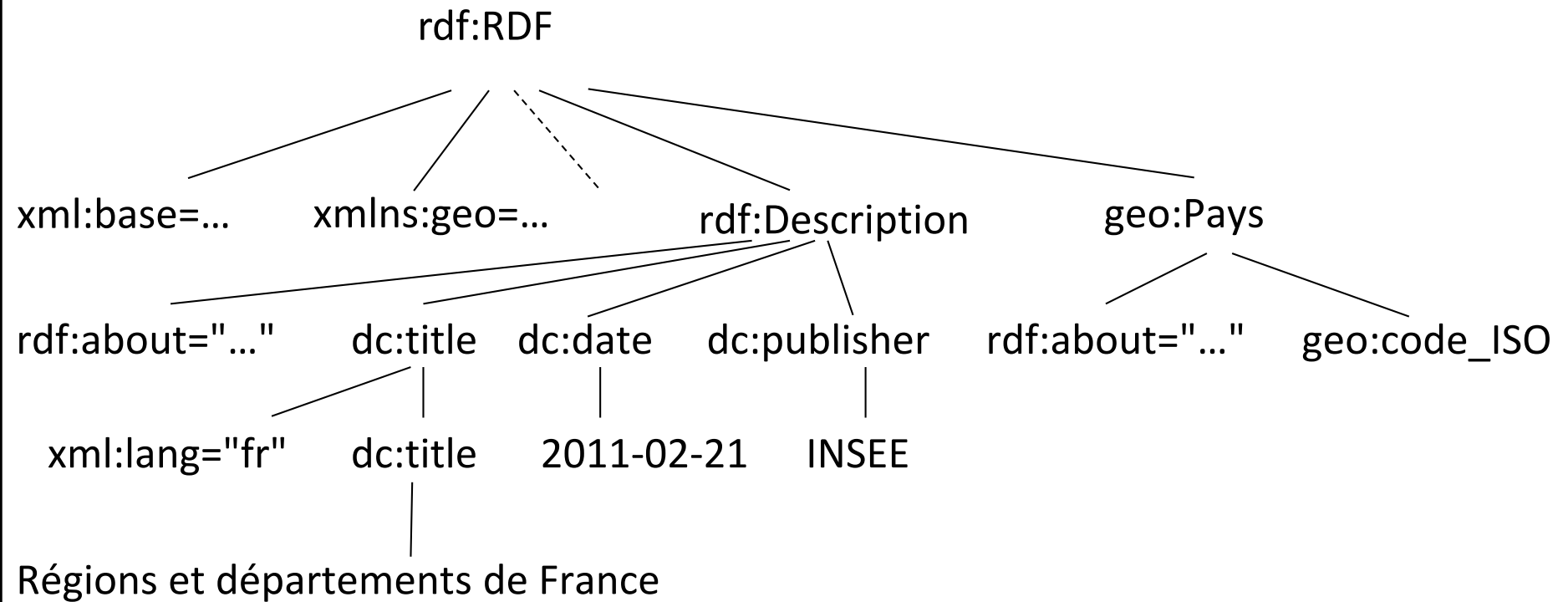
# Anatomie d'un document RDF/XML

```
<rdf:RDF xml:base="http://rdf.insee.fr/geo/2011/"
  xmlns:geo="http://rdf.insee.fr/geo/" xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

  <rdf:Description rdf:about="http://rdf.insee.fr/geo/2011/regions-2011.rdf">
    <dc:title xml:lang="fr">Régions et départements de France</dc:title>
    <dc:date>2011-02-21</dc:date>
    <dc:publisher>INSEE</dc:publisher>
  </rdf:Description>

  <geo:Pays rdf:about="PAYS_FR">
    <geo:code_ISO>FR</geo:code_ISO>
    <geo:nom xml:lang="fr">France</geo:nom>
    <geo:subdivision>
      <geo:Region rdf:about="REG_93">
        <geo:code_region>93</geo:code_region>
        <geo:nom xml:lang="fr">Provence-Alpes-Côte d'Azur</geo:nom>
        <geo:chef-lieu>
          <geo:Commune rdf:about="COM_13055">
            <geo:code_commune>13055</geo:code_commune>
            <geo:nom xml:lang="fr">Marseille</geo:nom>
          </geo:Commune>
        </geo:chef-lieu>
        ...
      </geo:Region>
    </geo:subdivision>
    ...
  </geo:Pays>
</rdf:RDF>
```

# Arbre RDF/XML



# Prologue

- Déclaration XML

<?xml version="1.0"?>

<?xml version='1.0' encoding="US-ASCII" standalone='yes'?>

<?xml version="1.0" encoding='iso-8859-1' standalone="no"?>

- Déclaration de type de document

<!DOCTYPE book SYSTEM "Usmlstuff/dtds/barebonesdb.dtd"

[

<!ENTITY companyname "Cybertronic">

<!ENTITY productname "Tournevis Sonic 2000">

} Déclaration  
d'entités

# Entités

- Réserves de contenu
  - Lisibilité, maintenance
- Appel d'une entité: **&nom;**
- Déclaration d'une entité: **<!ENTITY nom "valeur" >**

**<!ENTITY cpr "Copyright UNS 20017">**

**<text>du texte libre ... &cpr; </text>**

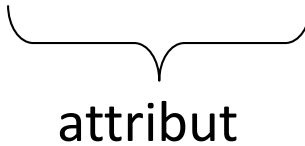
équivalent à

**<text>du texte libre ... Copyright UNS 2017 </text>**

# Éléments et attributs

- Élément conteneur

`<nom att1 = "val1" att2 = 'val2'>`  
contenu  
`</nom>`



attribut

- Élément vide

`<nom att1 = 'val' att2 = "val2"/>`

- Règles syntaxiques:

- La balise de début d'un élément précède celle de fin
- Balises de début et fin sont dans le même élément parent

# Balisages particuliers

- Commentaires

`<!-- ceci est un commentaire -->`

- Sections CDATA

`<exampleOfACDATA>`

`<![CDATA[`

Comme c'est une section CDATA on peut utiliser toutes sortes de caractères réservés comme `>`, `<`, `"` et `&`, ou écrire des choses comme `<foo></bar>` et le document XML reste bien formé!

`]]>`

`</exampleOfACDATA>`

- Instructions de traitement pour le processeur

`<?xml-stylesheet type='text/xsl' href='talk.xsl'?>`



# Arbre XML (revisit )

- N uds internes:
  - Noms de balises
- Feuilles:
  - Attributs (dont d clarations de namespaces)
  - Texte et/ou entit s
  - Commentaires
  - Instructions de processing

# Document XML bien formé

- Tout élément contenant a une balise de début et une balise de fin
- Tout élément vide a une barre oblique en fin de sa balise
- Toutes les valeurs d'attribut sont entre guillemets
- Les éléments ne se chevauchent pas
- Les caractères de balisage <, > et & n'apparaissent pas dans le contenu textuel d'un élément
- Les noms d'éléments commencent par une lettre ou un caractère souligné et comportent des lettres, des chiffres, des tirets, des points, des caractères soulignés. Les deux-points sont réservés aux espaces de nom.

# XML Namespaces

<http://www.w3.org/TR/REC-xml-names/>

- Eviter les conflits de noms entre DTD:
  - outil:fraise ≠ fruit:fraise
- Traiter différemment les objets selon leurs namespaces
- Noms qualifiés: **ns-prefix:nom-local**
- Déclaration d'un espace de noms: **xmlns:nom = "url"**
  - Portée: les éléments fils de l'élément contenant l'attribut xmlns

# XML Namespaces

Namespaces només

```
<RDF xmlns:rdfs="&rdfs;"  
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
  xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#">  
  <rdfs:Class rdf:ID='Object' />  
</RDF>
```

Namespace par défaut

Par exemple,

**rdfs:Class** désigne la ressource

"http://www.w3.org/2000/01/rdf-schema#Class"

**RDF** désigne la ressource

"http://www.w3.org/1999/02/22-rdf-syntax-ns#RDF"

# Attribut xml:base

<http://www.w3.org/TR/xmlbase/>

```
<rdf:RDF xmlns:rdfs="&rdfs;"  
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
  xml:base="http://www.polytech.unice.fr/si/xml">  
<rdfs:Class rdf:ID='Object'/>  
</rdf:RDF>
```

**Object** désigne ensuite la ressource

"http://www.polytech.unice.fr/si/xml#Object"

Document Type Definition

# **DTD IN A NUTSHELL**

# Modèle de document

- La grammaire et le vocabulaire d'une **application XML**
- Permet de déterminer si un document XML est **valide**, i.e. conforme à une application XML
- Dans le prologue, **standalone=yes** indique aux processeurs XML que le document peut être traité sans document extérieur, et en particulier sans modèle de document :
  - le vocabulaire est illimité,
  - il n'y a pas de règle de grammaire,
  - Il n'y a pas de restrictions sur les attributs
- **Un document qui déclare un modèle doit s'y conformer**