

Documents structurés multilingues (XML)

Andrei Popescu-Belis
TIM / ETI, Université de Genève

Cours n°1

<http://moodle.unige.ch>

Qu'est-ce que XML ?

- « eXtensible Markup Language »
- Langage de balises permettant de marquer les documents afin de les structurer
- Coder la **structure** ou le **sens** plutôt que la forme d'un document → **DOCUMENTS STRUCTURÉS**
- Langage flexible, expressif, simple, répandu

Premier exemple en XML

- Entrée lexicale d'un dictionnaire bilingue

```
<?xml version="1.0" encoding="UTF-8"?>
<LETTRE VALEUR="x">
  <ENTREE>
    <mot_cle>Xerox</mot_cle>
    <trad pos="nf">photocopieuse</trad>
    <trad pos="nf">photocopie</trad>
  </ENTREE>
</LETTRE>
```

3

Second exemple en XML

- Fiche bibliographique pour un ouvrage

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ouvrage>
  <index valeur="125"/>
  <titre>La légende des siècles</titre>
  <auteur>
    <nom>Hugo</nom>
    <prénom>Victor</prénom>
  </auteur>
</ouvrage>
```

4

Explications (1)

- Première ligne : **déclaration XML**
`<?xml version="1.0" encoding="UTF-8"?>`
 - la version est en général 1.0
 - il existe une version 1.1 (depuis 2004)
 - seulement pour tenir compte de la mise à jour d'Unicode
 - l'encodage des caractères est à préciser si l'on veut

- Puis : une suite **d'éléments** = balises + texte

Voici un élément sans attributs :

```
<trad>photocopieuse</trad>
```

Ou bien, voici un élément avec deux attributs :

```
<trad pos="nf" dom="tech">photocopieuse</trad>
```

5

Explications (2)

- Un fichier XML est donc:
 - un fichier texte = suite de caractères lisibles sur tout ordinateur
 - *pas* un fichier Word, PDF, ou autre format propriétaire
- Les noms des éléments (balises) sont à définir selon les besoins de chaque application

6

Intérêt d'XML

- Structurer l'information
- Rendre explicite le contenu des documents
- Echanger de l'information entre humains et/ou entre programmes informatiques
- Traiter l'information structurée en XML grâce aux nombreux outils existants
- Applications (exemples):
<http://xml.coverpages.org/xmlApplications.html>

7

La « galaxie XML » (1)

- **XML** : la syntaxe de base des balises
- **Espaces de noms** : éviter les conflits entre noms de balises pour différentes applications
- **XSL** (eXtensible **S**tylesheet **L**anguage)
 - **XSLT** : transformations d'un document XML
 - **XSL-FO** : formatage en vue de l'affichage
- **XPath** : mécanisme pour se référer à une partie d'un document XML

8

La « galaxie XML » (2)

- **XLink, XPointer** : connexions entre documents
- **XML Query** : langage de requêtes
- **XForms** : langage de définition des données pour des formulaires
- **WSDL** : langage de définition pour les services web

- Les **DTD** et **XML Schema**
 - définir la syntaxe des balises pour une application
 - DTD : **D**ocument **T**ype **D**efinition (ex-SGML)
 - XML Schema: plus riche, plus précise, écrite en XML

9

Bref historique (1)

- **Ancêtre : SGML** (Standard Generalized...)
 - début des années 1980
 - 1986 : standard ISO 8879
 - mêmes principes (structuration sémantique des fichiers), mais plus complexe que XML
- **Application de SGML: HTML ***
 - le langage des pages web
 - seulement intéressé par l'aspect graphique des documents hypertexte

* Hyper-text markup language

10

Bref historique (2)

- 1996: initiative visant à alléger SGML → XML
- 1998: spécification **XML 1.0**, très appréciée
- 2004: **XML 1.1** pour une mise à jour sur les caractères spéciaux en lien avec Unicode

- Organisme fédérateur: le W3C (www.w3.org)

- Spécifications ouvertes, format transparent
- Développement d'outils et de fonctionnalités complémentaires
- Très nombreuses **applications**, i.e. particularisations des balises à un problème précis

11

XML vs. HTML

- Quelques points communs mais des différences importantes
- Noms de balises
 - XML : à définir selon les besoins
 - HTML : fixés en vue de l'affichage
- Syntaxe des balises
 - XML : plutôt stricte
 - HTML : assez tolérante, dépend du navigateur
- Nombreux outils associés à XML, applications XML possibles, etc.

12

XML et le traitement multilingue

- Structuration de documents : XML
- Support des caractères spéciaux : XML
- Interfaçage avec Unicode : XML
- Format de localisation : XLIFF < XML
- Echange de mém. de trad. : TMX < XML
- Applications en terminologie : TMF < XML
- Gestion de sites web multilingues
- Systèmes d'enseignement à distance
 - <http://tecfu.unige.ch/guides/xml/applications.html#section5>

13

La documentation XML

- Recherche des principaux sites définissant les normes XML et fournissant de l'aide:
 - www.w3.org www.w3schools.com
 - ibiblio.org/xml xml.coverpages.org
- Plus technique:
 - xml.com, xml.org, xml.apache.org
- Regardez ces sites et notez leurs URLs

14

Intérêt du cours d'XML (1)

- Introduction aux outils XML
 - la syntaxe XML
 - définir des balises XML personnalisées : DTD
 - gérer/transformer des documents XML : XSLT
- Introduction aux applications d'XML
 - annotation de textes: XCES
 - format d'échange pour
 - la localisation : XLIFF
 - les ressources terminologiques : MARTIF, XLT, TMF, TBX
 - les mémoires de traduction : TMX
 - méta-données: DublinCore, RDF, RSS

15

Intérêt du cours d'XML (2)

- Pratique des langages de balises
 - liens avec HTML, XHTML
- Avenir : XML est la base du « **Semantic Web** »
 - informellement appelé « **Web 2.0** »
- Pratique de l'informatique et introduction à la programmation
 - nombreux exercices
 - utilisation d'un outil de référence
 - version professionnelle (NB: la version de base est gratuite)
 - écriture d'éléments de programmes, voire plus!

16

Organisation du cours

- Une heure de théorie
- Une heure de pratique
 - exercices sur ordinateur (salle 6289)
- Les questions et les discussions sont fortement encouragées

17

Utilisation de Moodle comme support

<http://moodle.unige.ch>

- Inscription individuelle
- Consultation des transparents
 - en général disponibles quelques jours avant le cours
- Téléchargement des exercices
- Messages, questions, sondages, etc.
- **Mais n'oubliez pas de venir aux cours 😊**

18

Evaluation : projet personnel

- Contenu du projet
 - approfondir une fonctionnalité, une commande, un outil, etc. (vus ou non en cours)
 - analyse théorique et illustration pratique
 - écrire/utiliser quelques courts fichiers XML, une DTD/XSD, une feuille de style XSLT/CSS (selon les besoins du projet)
- Présentation
 - courte description écrite, en 1-2 pages
 - exposé + démo + discussion, en 20-30 minutes

19

Premiers exercices

Application choisie: stocker des fiches bibliographiques (livres, articles, etc.) sous la forme de documents XML

Comment procéder ?

Outils d'édition XML

- Editeur XMLSpy : le plus répandu
 - gratuit dans sa version « Home Edition »
- http://www.altova.com/download_spy_home.html
- version installée en salle 6289:
Enterprise XML Suite 2006
 - nombreuses possibilités d'affichage et de test
 - seules les fonctions de base seront utilisées
 - ... heureusement ☺

21

XMLSpy : Fonctionnalités de base (1/2)

- Préambule
 - nombreuses commandes → viser les plus simples
 - commandes accessibles par les menus ou par les boutons en dessous (pour les plus importantes)
 - semblable à Word / personnalisation possible
- Fenêtres et cadres
 - garder seulement la fenêtre avec les documents
 - noter (à gauche) les numéros de ligne, les signes +/-, les alignements
- Menu 'File'
 - ouvrir / enregistrer fichiers XML
 - un onglet (*tab*) par fichier
 - plusieurs modes de visualisation → Text / Grid / Browser

22

XMLSpy : Fonctionnalités de base (2/2)

- Menu 'Edit'
 - commandes d'édition habituelles
 - noter : 'Pretty-print XML text'
 - formatage graphique
 - plusieurs modes de visualisation → Text / Grid / Browser
 - G/B possibles seulement si le fichier est bien formé
- Menu 'XML'
 - commandes pour la création de fichiers XML
 - surtout : 'Check well-formedness' (ou F7)
 - ou bouton marqué avec ✓
- Menu 'View'
 - options d'affichage diverses; noter : 'Expand fully +'

23

XMLSpy : utiliser l'aide → menu 'Help'

- Table of contents
 - User Manual
 - Introduction
 - Tutorials
 - XMLSpy Tutorial > Creating an XML document
 - Text View
 - Enhanced Grid View
 - User Reference

24

Tâches de la séance (1/2)

- Connectez-vous à Moodle, inscrivez-vous au cours
- Créez un dossier local **H:\prive\xml**
- Enregistrez dans ce dossier (*sans l'ouvrir*) le fichier **livre1.xml** disponible sur Moodle (Cours 1)
 - utilisez la commande ‘Enregistrer la cible du lien sous...’ du menu contextuel du navigateur

25

Tâches de la séance (2/2)

- Lancez XMLSpy (raccourci sur le bureau)
- Ouvrez le fichier **livre1.xml** dans XMLSpy
- Testez les différentes options d’affichage
 - en mode ‘Text’ ou ‘Grid’ ou ‘Browser’
- Modifiez l’auteur et le titre, supprimez les informations sur l’édition et les informations administratives
 - enregistrez le fichier obtenu sous un autre nom : **livre2.xml**
 - s’il y a des erreurs XML, on doit forcer l’enregistrement
 - (la syntaxe XML sera le sujet des prochains cours)

→ **Soumettre livre2.xml sur Moodle avant mardi prochain 12h00**

26