Normes basées sur XML pour la gestion de ressources terminologiques et lexicales

Andrei Popescu-Belis
TIM / ETI, Université de Genève

Cours n°13

Plan du cours

1. Normes pour la terminologie
2. Normes pour les lexiques électroniques

Outils terminologiques

- Exemples d'applications
  - extraction de termes
  - création / maintenance de glossaires
  - recherche terminologique
  - validation de la terminologie après traduction
- Les **standards**
  - importants pour l'échange des banques de données terminologiques
  - utilisés surtout par des outils qui permettent de définir, maintenir, vérifier et traduire la terminologie d'un domaine

Standards XML en terminologie (1/2)

- ISO 1087
  - concepts utilisés en terminologie
- ISO 12618
  - principes de création des banques de données terminologiques
- ISO 12620
  - catégories de données en terminologie
- ISO 12200 ou MARTIF
  - *Machine-Readable Terminology Interchange Format*
  - principes de conversion et encodage XML de banques de données terminologiques
- XLT ou DXLT
  - développement basé sur MARTIF
  - projet SALT

Standards XML en terminologie (2/2)

- ISO 16642 ou TMF
  - *Terminological Markup Framework*
  - méta-format d'annotation XML
  - deux applications: MARTIF et GENETER
  - projet européen
- TBX
  - *Term Base eXchange*
  - groupe OSCAR de LISA
- OLIF2
  - *Open Lexicon Interchange Format*
  - consortium OLIF
  - structure et encodage XML de lexiques multilingues

XLT : XML representation of Lexicons and Terminology

- Principe
  - fournir un format d'échange universel pour les données lexicales et terminologiques
  - utiliser la norme MARTIF: standard ISO 12200:1999 (*MAchine-Readable Terminology Interchange Format*)
  - … elle-même basée sur ISO 12620 (*Computer applications in terminology - Data categories*)
  - se concerter avec le format OLIF pour les lexiques
    - XLT reste focalisé sur l'aspect terminologique
    - OLIF se focalise sur l'échange de données lexicales, p.ex. utilisées par des systèmes de TAO

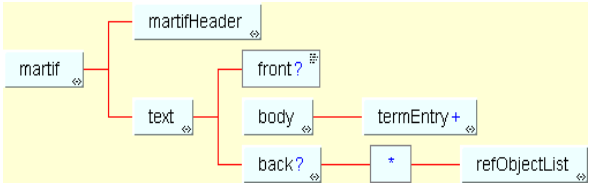
Description du projet SALT

- Organisation
  - projet transatlantique (2000-2002)
    - http://www.loria.fr/projets/SALT Europe
    - http://www.ttt.org/salt/ Etats-Unis
  - Standards-based Access to Lexicographical and Terminological multilingual resources
  - consortium d'universités, associations, sociétés, institutions publiques aux E.-U. et en Europe
- Objectifs
  - tester et développer XLT (d'abord DXLT, Default XLT)
  - développer un site Internet avec des utilitaires
  - développer une boîte à outils XLT
    - conversion OLIF / XLT, Geneter / XLT, etc.
    - manuel de définition d'outils de conversion

7

Description de DXLT (1)

- But : encoder en XML un ensemble de termes avec leurs définitions, etc.
- Représentation des données à encoder :



8

Description de DXLT (2)

```
<?xml version='1.0'?>
<!DOCTYPE martif PUBLIC "ISO 12200:1999A//DTD MARTIF core (MSCcdV04)//EN">
<martif type='DXLT' lang='en' >
<martifHeader>
        <fileDesc><sourceDesc><p>from an Oracle termBase</p></sourceDesc></fileDesc>
        <encodingDesc><p type='DCSName'>MSCdmV04</p></encodingDesc>
</martifHeader>
<text> <body>
        <termEntry id='ID67'>
                <descrip type='subjectField'>manufacturing</descrip>
                <descrip type='definition'>A value between 0 and 1 used in …</descrip>
                <langSet lang='en'>
                <tig>       <term>alpha smoothing factor</term>
                            <termNote type='termType' >fullForm</termNote> </tig>
                </langSet>
                <langSet lang='hu'>
                            <tig><term>Alfa simítási tényezõ </term></tig>
                </langSet>
        </termEntry>
</body> </text>
</martif>
```

9

Description de DXLT (3)

- Explications sur le tutoriel
- http://www.ttt.org/oscar/xlt/webtutorial/

- Eléments
  - structure d'ensemble de la base de termes
  - structure d'un terme
    - définition du concept
    - termes dans plusieurs langues

10

Prolongements

- Le projet SALT s'est terminé en 2002
  - résultats : gratuits, transparents (open-source)
  - XLT a été publié, converti en (pré-)normes
- Suite du travail
  - repris par LISA : Localisation Industry Standards Association, groupe OSCAR
  - DXLT a été repris dans le TBX, Terminology Base eXchange
  - travail encore en cours
- Intérêt d'une norme pour l'échange terminologique
  - insertion rapide de nouveaux termes dans une base
  - uniformité accrue entre les documents
  - amélioration de la traduction automatique

11

TBX

- Term Base eXchange
  - reprend le XLT par défaut, DXLT, de SALT
- Organisation
  - LISA
  - OSCAR : sous groupe de LISA
    - Open Standards for Container/Content Allowing Re-use
- Version 1.0 publiée en mai 2002
  - http://www.lisa.org/standards/tbx

12

Encore plus de généralité : TMF

- Terminological Markup Framework
  - standard de l'ISO, 16642 (*presque*)
  - objectif : décrire des méta-contraintes applicables à toute application d'XML pour le marquage terminologique
  - deux instanciations décrites : MCS (MARTIF with Specified Constraints) et Geneter
- Utilise surtout ISO 12620 sur les catégories de données
- Assez abstrait (méta-langage). Voir : http://www.loria.fr/projets/TMF/

Plan du cours

1. Normes pour la terminologie
2. Normes pour les lexiques électroniques

Standard pour l'échange de lexiques : OLIF

- Open Lexicon Interchange Format
  - d'abord dans le cadre d'un projet européen OTELO
  - actuellement : Consortium OLIF dirigé par SAP
    - industriels : TAO, ingénierie linguistique
    - instituts de recherche
    - utilisateurs des technologies

  → http://www.olif.net
- Actuellement
  - OLIF 2 et 2.1 sont disponibles (DTD et schéma XML)
    - OLIF 2.1.1_10, Novembre 2006
  - à venir : des mécanismes de conversion OLIF/XLT

Principes d'OLIF (1)

- Objectifs
  - OLIF : versant lexicographique de XLT
  - format d'échange de lexiques
  - destiné aux outils de TAL
  - utilisation paradigmatique : outils de TAO
- Equilibre lexique / termes
  - ni trop « lexical » (orienté vers les lexèmes)
  - ni trop terminologique (orienté vers les concepts)
    - pour cela : cf. DXLT / TBX
  - format structuré selon les divers sens de chaque mot

Principes d'OLIF (2)

- Structure d'une banque lexicale (XML, compatible TMF)
  - en-tête, *puis* :
  - liste d'entrées monolingues (sens par sens)
  - liens entre les entrées :
    - intra-langue (synonymes, etc.)
    - inter-langues (possibilités de traduction)
- DTD modulaire + Schema XML
  - mécanisme de définition assez complexe

→ http://www.olif.net

Exemple d'entrée OLIF

```
<entry>
 <mono>
  <keyDC>
   <canForm>table</canForm>
   <language>en</language>
   <ptOfSpeech>noun</ptOfSpeech>
   <subjField>general</subjField>
   <semReading>86</semReading>
  </keyDC>
  <monoDC>
   <monoAdmin>
    <originator>Weber</originator>
    <adminStatus>ver</adminStatus>
   </monoAdmin>
   <monoMorph>
   <inflection>like: book,books
    </inflection>
   </monoMorph>
   <monoSem>
   <definition>An arrangement of words,
    numbers, or signs or combinations of
    them, as in parallel columns, …
   </definition>
   <semType>inform</semType>
   </monoSem>
  </monoDC>
 </mono>

 <crossRefer>
  <keyDC>
   <canForm>row</canForm>
   <language>en</language>
   <ptOfSpeech>noun</ptOfSpeech>
   <subjField>general</subjField>
   <semReading>69</semReading>
  </keyDC>
  <crLinkType>has-meronym</crLinkType>
 </crossRefer>

 <transfer>
  <keyDC>
   <canForm>Tabelle</canForm>
   <language>de</language>
   <ptOfSpeech>noun</ptOfSpeech>
   <subjField>general</subjField>
   <semReading>86</semReading>
  </keyDC>
 </transfer>

 </entry>
```