UNELE ASPECTE PRIVIND CODIFICAREA ADNOTĂRII

N.Dulgheru
Universitatea Tehnică a Moldovei

1. FORMALISME DE ADNOTARE PENTRU PROCESAREA LIMBAJULUI NATURAL

Adnotarea corpului de date cere alegerea unui format pentru reprezentarea textului şi a adnotărilor sale în formă electronică. Formatul dat trebuie să permită folosirea şi refolosirea corpului de date adnotat de programele soft disponibile în diferite locuri de cercetare. Corpul de date şi documentele de adnotare de asemenea trebuie să fie codate ca să facă accesul datelor cât mai uşor şi flexibil. Limbajul extins de marcări (Extended Markup Language, XML) asigură un cadru de codificări standarde pentru adnotări, care satisface aceste necesităţi. Folosind XML ca bază, Standardul XML de Codificare a Corpului de Date (XML Corpus Encoding Standard, XCES) a fost realizat ca parte a proiectului EAGLES, cu scopul de a oferi un cadru pentru codificarea şi organizarea corpului de date şi a adnotărilor sale într-un format standard, flexibil şi reutilizabil.

Acest raport aduce la cunoştinţa studentului unele standarde de suport şi înrudite cu XML, dezvoltate în cadrul XML, cum sunt XSLT (Limbajul de Transformare XML/ XML Transformation Language) şi RDF (Cadrul de Definiţii Resurse/ Resourse Definition Framework). Aceste implementări oferă împreună atât mijloace de codificare a datelor şi a adnotărilor lor, folosind arhitectura documentului XCES, cât şi posibilitatea de a manipula şi accesa aceste date. În plus, aici, se schiţează problemele şi preocupările pentru reprezentarea datelor adnotate şi se face o privire generală asupra unui model abstract al corpului de date şi asupra adnotărilor lor.

Ce este XML? XML, o abreviere pentru Extended Markup Language (Limbaj Extins de Marcări), este un metalimbaj care permite crearea unui limbaj propriu de marcări. XML prezintă o simplificare a standardului SGML, ultimul fiind destinat reprezentării structurii "logice" a unui document, iar HTML fiind conceput ca o aplicaţie a lui SGML.

Un document XML este un document SGML cu unele diferenţe mici (dar esenţiale). XML este tot atât de expresiv, însă nu atât de complex în comparaţie cu SGML, şi face posibilă transmiterea în reţea a documentelor structurate şi a bazelor de date.

Cu ajutorul XML putem modela date, publica date structurate în reţea, separa structura logică a unui document de prezentarea sa reală, integra date din surse eterogene. Nu putem evita folosirea XML din cauza simplicităţii sale, care face posibilă integrarea acestui limbaj în orice fel de aplicaţii. Plus la aceasta, o varietate largă de aplicaţii este deja implementată în industrie (publicare, baze de date, catalogare, businessul electronic etc.), ştiinţă şi cercetări (astronomie, matematică etc.), iar la dispoziţie avem multe programe soft: editoare, analizatoare sintactice (parsers), punţi de la şi spre mediul de editare existent.

XML are următoarele proprietăţi: accentul trebuie pus pe "semantica" unui document, modelul de bază are structura unui arbore, posibilitatea de a imagina un limbaj-schiţă pentru a accesa orice parte a unui document XML, XML suportă codificările de caractere Unicode.

Un document XML reprezintă o structură ierarhică (fig. 1):

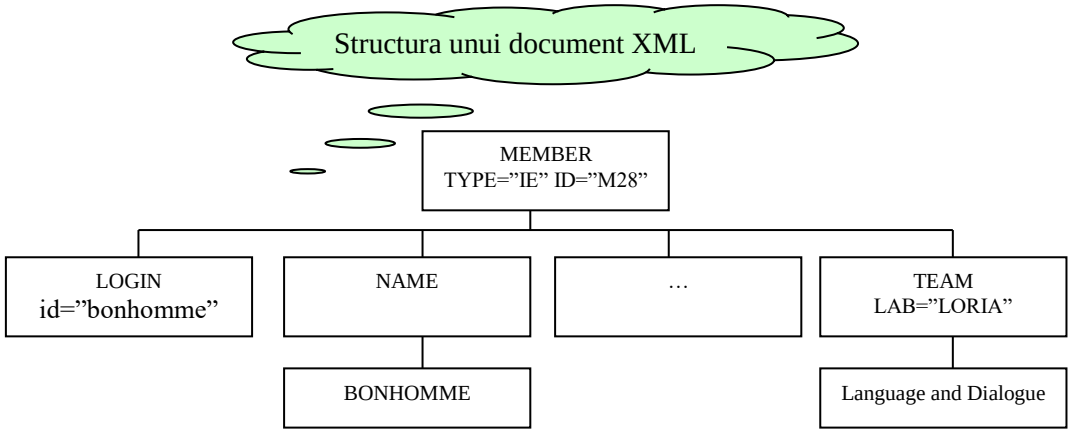


Figura 1

2. CODIFICAREA ADNOTĂRII CU XML

Ce prezintă un corp de date adnotat?

- Nişte date primare, iniţiale…
  - un text, un semnal de vorbire etc.
- …asociate cu o însemnare (notă explicativă) despre proprietăţile sale lingvistice;
  - categoria morfo-sintactică pentru fiecare cuvânt, categoriile sintactice şi structura, structura discursului, coreferinţe etc.

Unde sunt adnotările?

- Deseori, în acelaşi document /fişier cu date primare

Penn Treebank

```
((S  ((NP-SBJ-1 Jones)
     (VP followed)
        (NP him)
        (PP-DIR into)
                (NP the front room))
                ,
        (S-ADV     (NP-SBJ* -1)
                    VP closing
                    NP the door)
                   (PP behind
                       (NP him))))).))
```

```
<TEXT>
<p>
  <s>
    <lex pos=DT>The</lex>
    <lex pos=NNP>Federal</lex>
    <lex pos=NNP>Aviation</lex>
    <lex pos=NNP>Administration</lex>
    <lex pos=VBD>underestimated</lex>
    <lex pos=DT>the</lex>
    <lex pos=NN>number</lex>
    <lex pos=IN>of</lex>…
```

De ce aceasta nu este o idee prea bună?

- Este greu de avut câteva tipuri diferite de adnotări în acelaşi corp de date;
- Este greu de avut adnotări alternative de acelaşi tip;
- Este greu de menţinut;
- …şi multe alte cauze.

XCES rezolvă problema!

- XCES = Standardul XML de codare a corpului de date;
- "Însemnări aparte"(stand-off markup);
  - Adnotări în documente XML separate, legate de original (fig. 2).

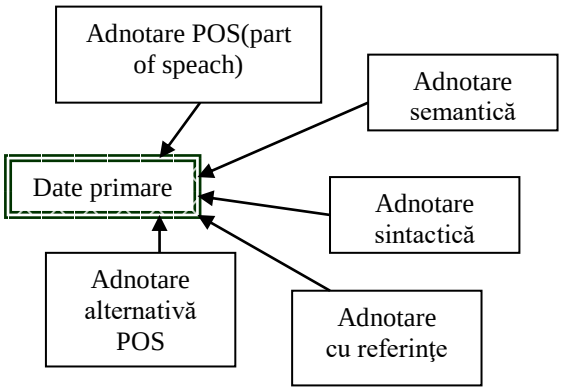


Figura 2

Pointerii XML uşurează lucrul

- Se poate de indicat un element, caracter, şir de caractere etc. în acelaşi document sau în altul (fig. 3).

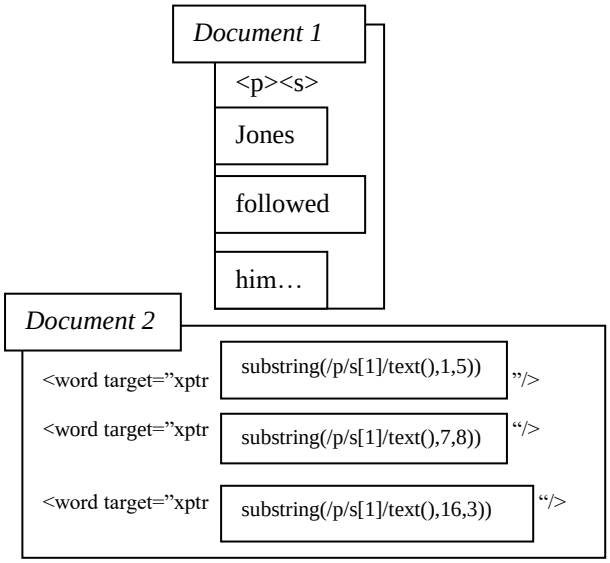


Figura 3

Ce fel de notări se fac asupra datelor primare?

- XCES: "Trăsături sugerate de tipografie"
  - Structura logică generală (capitol, paragraf, titlu, referinţă etc.)
  - Propoziţii?
  - Cuvinte?

Unde se începe "adnotarea?"

Deci cum să alcătuiesc o schemă de codificare pentru adnotări?

- Răspuns simplu:
  - *Oricum vrei, doar să urmezi un model coerent de date.*

Ce este de dorit să conţină o schemă de adnotări?

- Cuprinde caracteristicile de care ai nevoie;

- Uşor de procesat atât pentru tine, cât şi pentru alţii;
- Compatibilă cu alte scheme, ca să permită fuzionarea, compararea etc.;
- Permite ca adnotările să poată fi *utilizate din nou* şi să fie *extensibile.*

Pasul întâi

- Găseşte *categoriile lingvistice şi trăsăturile* pe care doreşti să le reprezinţi, fără să iei în considerare problema *formatului fizic*
  - E.g.
    - ❖ Parte de vorbire: cât de detaliată este informaţia despre categorii? Care sunt lemele specificate?
    - ❖ Sintaxa: care sunt categoriile (NP,VP)? Este funcţia sintactică (subiect, complement) sau informaţia tematică (agent, pacient) inclusă?

Pasul doi

- Dezvoltă o modalitate de a reprezenta informaţia;
  - Preferabil în XML, dar nu numaidecât;
  - Care poate să fie procesată de software-ul tău;
  - Nici o informaţie pierdută;
  - Atât timp cât informaţia este reprezentată, nu este nevoie sa te îngrijorezi de formatul specific XML.

Exemplu (Parte de vorbire)

```
<w><orth>dogs</orth>
    <cat>NNP</cat>
    <lemma>dog</lemma>
</w>
```

```
<seg target="xptr(substring(/p/s[1]/text(),1,5)"/>
    lex="MyCats#DX51"/>
```

```
<w cat="NNP" lemma="dog">dogs</w>
```

Figura 4

De ce nu este important formatul exact (modul de aranjare şi prezentare)?

- Pentru că există XSLT!
  - Este uşor de a trece de la un format la altul (…Atât timp cât informaţia există);
  - Adnotatorii pot să continue să lucreze după formatele lor interne;
  - Trebuie să fie construit după un model de date comun;

De ce avem nevoie?

- De un model comun, abstract care poate exprima (reprezenta) informaţia adnotării, ne luând în considerare codarea fizică;
- De o exemplificare concretă generică XML a modelului, spre care şi din care pot fi traduse formate concrete folosind XSLT;
- De un set comun de categorii de date, la care adnotatorii să facă referinţă şi de care să se folosească.

Deci cum lucrează aceasta? (fig. 5)

Registrul Categoriilor de Date (RCD)

- Inventarul categoriilor de date;
  - Specificate de colectivul de cercetători;
  - Poate fi în continuu extinsă;
- Menţinut pentru referire şi folosire universală;
  - Nu este un standard, ci un punct de referinţă;
  - Arată cum pot fi găsite subcategoriile specifice în categoriile din RCD:

Instanţierea Registrului Categoriilor de Date

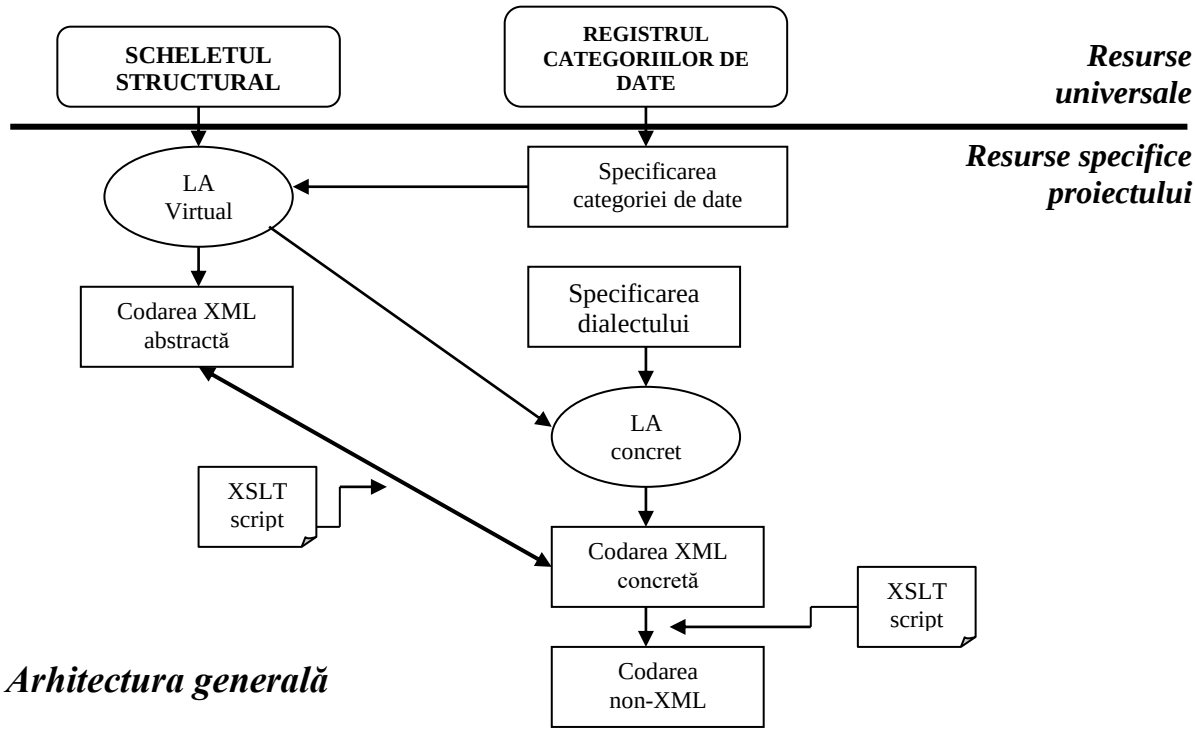
- Este definit folosind Cadrul de Definiţii Resurse (Resource Definition Framework, RDF);
  - Realizat de comunitatea XML (Consorţiul W3);
  - Schemele RDF specifică semantica datelor bazându-se pe XML (obiecte şi relaţiile lor, ierarhii ale tipurilor derivate;
  - Descrierile RDF dau exemple de obiecte.

Scheletul structural

- Un model general, fundamental care informează practica curentă;
- Un Instrument de trasare generică (Generic Mapping Tool, GMT) furnizează realizarea XML a scheletului structural;
  - *Lingua franca* folosită să compare şi să unească adnotările, să facă posibil designul instrumentelor generice pentru vizualizare, editare, extragere etc.

Specificarea categoriei de date

- Set de categorii de date într-o schemă de adnotare dată;
  - Subclasă de categorii din RCD + (opţional) categorii specifice în aplicaţie definite suplimentar;
- Descrie constrângerile asupra categoriilor de date;
  - Restricţii asupra valorilor;



Arhitectura generală

Figura 5

- Restricţii asupra locului unde pot apărea categoriile de date (de ex., nivelul într-o ierarhie structurală).

Specificarea dialectului

- Defineşte formatul XML specific proiectului;
- Foloseşte scheme XML, XSLT scripts, foi de stil XSL;
- Include;
- *Stiluri de instanţiere a categoriilor de date*

```
<NounPhrase>
<cat type="NounPhrase">
```

- *Stiluri ale vocabularului de categorii de date*

```
<cat type="NounPhrase">
<cat type="NP">
<cat type="SN">
```

- *Structuri de lărgire (expansiune)*
  - ❖ Modifică structura exprimată folosind scheletul structural;
  - ❖ Ex. Creează două subnoduri sub un nod dat pentru a grupa diferite tipuri de informaţie.

Limbajul de Adnotare (LA) / (Annotation Markup Language, AML)

- Scheletul structural + SCD(Specificarea Categoriilor de Date)=**LA Virtual;**
- LA Virtual + Specificarea Dialectului folosită pentru a genera în mod automat **LA concret**
  - Se conformează formatului specific proiectului în Specificarea Dialectului;
  - Filtrele XSLT traduc
    - ❖ Reprezentările adnotării în LA concret şi în Instrumentul de Trasare Generică (ITG) / (GMT, Generic Mapping Tool);
    - ❖ Din LA concret în formate non-XML (ex. Penn Treebank).

Bibliografie

1. A.Mengel, W.Lezius. *An XML –based representation format for syntactically annotated corpora/ Proceeding of the 2nd Conference an Language Resources and Evaluation LREC, Athens, Greece, 2000.*

2. *XCES.2000. Corpus Encoding Standard for XML.2000. Vassar College. LORIA/CNRS (http://www.cs.vassar.edu/XCES/).*

3. *W.3C.2000.XSL Transformation (XLST). Version 1.0 (http://www.w3c.worg/TR/xlst.*