Activitatea A1.3

Proiectarea funcțională și arhitecturală a infrastructurii care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului și realizarea unui prototip

Faza de predare: noiembrie 2018

Autori: Dan Cristea, Andrei Scutelnicu

Ultima actualizare: 17 noiembrie 2018

1. Introducere

Prezentăm în această secțiune structura Portalul ReTeRom-COBILIRO[1], care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului. Un prototip al acestei platforme va fi prezentat în workshop-ul ReTeRom din 23 noiembrie 2018, ținut în tandem cu a 13 ediție a conferinței internaționale a Consorțiului de Informatizare pentru Limba Română - ConsILR-2018, *Linguistic Resources and Tools for Processing Romanian Language*, care se va desfășura între 22-23 noiembrie, la Filiala Iași a Academiei Române.

În funcție de deciziile consorțiului, o anumită zonă a acestui Portal va putea fi deschisă și publicului larg, cu respectarea drepturilor de autor și a confidenței datelor. Insistăm ca accesul la Portal să se facă prin site-ul oficial al proiectului ReTeRom/COBILIRO (http://85.122.23.18/COBILIRO/). Considerăm că implicarea unor membri ai echipei în proiectul laboratorului RLP-LeAL@ARFI-IIT[2] este benefică, astfel activitatea consorțiului putând să se sincronizeze cu cea a altor colective care desfășoară cercetări similare în alte centre de cercetare din țară. Din acest punct de vedere considerăm că Portalul pe care îl proiectăm aici ar trebui să permită astfel de colaborări, inclusiv prin partajarea unor secțiuni de date și tehnologii cu parteneri din țară. În aceeași manieră, colaborări viitoare ale colectivului ReTeRom cu alți parteneri va putea fi sincronizată prin intermediul Portalului. De aceea, în viziunea noastră, o trăsătură majoră a funcționalității Portalului va trebui să fie capacitatea de a partaja accesul pe diverse secțiuni ale lui.

2. Structura de pagini a Platformei și interfața utilizator
    a. **Acasă** (în esență o pagină de prezentare a Portalului), cu structura:
        i. **headline**: un titlu care să atragă și care să facă referire la activitatea de bază din proiect;

[1] În proiectarea Platformei am adoptat o viziune apropiată de cea pentru dezvoltarea laboratorului RLP-LeAL@ARFI-IIT (*Romanian Language Processing - Learning Algorithms Laboratory*, din componența Academiei Române Filiala Iași, Institutul de Informatică Teoretică). Acest laborator virtual este conceput ca un spațiu de lucru și interacțiune, în care cercetătorul în LN să găsească metode, algoritmi gata de utilizare, seturi de date de antrenament ori gata formate, lucrări de specialitate, linkuri către tutoriale, lucrări ale membrilor echipei, date rezultate din proiecte mai vechi, call-uri de conferințe etc., respectiv tot ce ține de activitatea cercetătorului în Limbaj Natural și care poate constitui ajutor cercetătorilor în lingvistică computațională și prelucrarea limbii române în activitățile lor.

[2] De exemplu, tema *Resurse electronice și tehnologii ale limbii române: contemporaneitate și diacronie*, din planul de cercetare al Institutului urmărește, printre altele, completarea corpusului CoRoLa cu noi date textuale și înregistrări sonore aliniate cu textul, ce vor putea fi utilizate și în proiectul ReTeRom.

ii. **chemare la acțiune**: o secțiune în care vom descrie pe scurt serviciile pe care le oferă Portalul, împreună cu o chemare (îndemn) de apelare la serviciile Portalului și la contribuirea cu resurse/servicii;

b. **Despre noi**
   i. o scurtă prezentare a instituțiilor partenere în consorțiul ReTeRom;
   ii. membrii consorțiului (scurte biografii, legături către site-urile științifice ale membrilor etc/).

c. Resurse și servicii de procesare a limbajului natural (existente și care pot fi încărcate de utilizatori autorizați)
   i. Resurse existente:
      1. Un inventar cu toate resursele bimodale, în ordine alfabetică (după <titlu>);
      2. Posibilitatea de căutare după criterii (metadate) sau cuvinte cheie din descrierea resursei;
   ii. Încărcare: poarta de intrare în sistem a unei noi resurse => o schemă a pașilor de urmat pentru *upload*:
      1. Utilizator (U) apasă buton *upload*;
      2. U: completează descrierea resursei;
      3. U (opțional): poate indica referințe bibliografice în care resursa este descrisă și care pot fi citate dacă resursa aparține autorului;
      4. U: completează șablonul de descriere a metadatelor (structură arborescentă);
      5. Platforma (P): verifică câmpul descriere, lista de referințe, corectitudinea și completitudinea metadatelor; dacă e cazul emite mesaje de corectare sau de completare a zonelor imperfecte;
      6. P: solicită U-ului încărcarea resursei;
      7. U: încarcă resursa din discul propriu;
      8. P: verifică consistența datelor care compun resursa cu cele declarate în metadate; emite mesaje de eroare, dacă e cazul;
      9. P: se procedează la conversia resursei la formatul standard agreat în consorțiul ReTeRom; dacă apar probleme la conversie, P poartă un dialog cu U până resursa este complet convertită sau procesul se abandonează;
      10. U: dacă primește mesaje de eroare, corectează resursa și o reîncarcă; se reia apoi dialogul de la 4;
      11. P: aduce la zi statistica generală asupra resurselor.

Observație: acest dialog poate suporta variații, în funcție de tipul de resursă. De exemplu, un serviciu web se "încarcă" în Platformă în mod diferit, pentru că are loc o instalare a ei. De

asemenea, pentru o resursă de utilitate generală (*open source*) se indică la conținut doar adresa web site-ului unde poate fi ea găsită (dar se completează descrierea și metadatele).

iii. Upgradare/updatare a unei resurse: U dorește să încarce o versiune nouă a unei resurse sau să o modifice;
   1. U apasă buton *upgrade*/*update*;
   2. P: afișează fereastră de identificare resursă;
   3. U: completează fereastra pentru identificarea resursei;
   4. P: dacă resursa este identificată în repozitoriu, întreabă U-ul dacă dorește să facă modificări în descriere, în referințele bibliografice, în metadate, în conținut, sau în combinații dintre aceste zone; dacă nu o identifică, afișează avertisment și reia dialogul, de la pct. 1;
   5. P&U: în funcție de opțiunea U-ului, are loc un dialog asemănător celui de la pct. ii prin care U-ul introduce o nouă variantă; P înlocuiește varianta veche cu cea nouă; dacă e cazul, P realizează conversia resursei în formatul standard;
   6. P: aduce la zi statistica generală asupra resurselor.

iv. Exploatare: posibilitate de descărcare/interogare (cu restricțiile IPR).
   1. U apasă buton *exploit*;
   2. (dacă e cazul) P: afișează fereastră de identificare a resursei sau o listă în care resursele pot fi găsite direct;
   3. U completează fereastra cu datele de identificare sau o identifică direct în listă;
   4. P afișează descrierea resursei și referințele bibliografice (care sunt întotdeauna libere la acces);
   5. P verifică compatibilitatea identității U-ului cu restricțiile de acces memorate în metadatele resursei;
   6. dacă e OK, permite accesul în exploatare online (vizualizare sau utilizare serviciu web), sau pentru descărcare executabil (pentru aplicații), sau pentru descărcare a unui segment ori descărcare completă (în caz de resursă de date).

d. Comunicări și publicații

i. Publicații ale consorțiului (comunicări susținute/articole publicate);

ii. Rapoarte de cercetare => accesibil numai persoanelor identificate ca aparținând consorțiului;

iii. Calendar de evenimente (interne dar și externe și de interes pentru consorțiu);

iv. Documentații, liste de lucrări externe colectivului, de interes pentru domeniul prelucrării limbajului natural;

v. Evenimente organizate de institutele consorțiului ReTeRom;

vi. Forum de discuții;
vii. O zonă a presei: o secțiune cu testimoniale / păreri care să credibilizeze serviciile oferite de Portal, precum și ecouri mediatice, linkuri către toate ecourile în media locală, națională, internațională asupra activităților proiectului;
viii. Un newsletter (opțional): o secțiune de unde se poate accesa un newsletter și cu un formular de abonare la el (dacă în consorțiu se va decide publicarea unui newsletter[3]). Ar putea ajuta la dezvoltarea unei baze de date conținând referințe asupra utilizatorilor activi sau potențiali (network) care poate fi o foarte bună resursă de diseminare a rezultatelor echipei, în organizarea de evenimente, în atragerea de studenți către activitatea de cercetare în LN etc.

e. Parteneri externi - clasificați în: educaționali (instituțiile academice), sponsori, parteneri media etc.
i. Parteneri externi (instituții și persoane individuale);
ii. Asociația Română de Lingvistică Computațională (ARLC): o listă a membrilor, cu scurte biografii.

f. Contact - datele de localizare (adresă, telefon, email, site etc.)

3. Funcționalitate

Distingem următoarele categorii funcționale ale Portalului:
a. Roluri
i. administrator;
ii. curator resurse - persoană care:
1. dă acceptul de publicare pe site;
2. monitorizează și gestionează resurse noi.
iii. contributor - persoană care donează resurse noi;
iv. utilizator de resurse
1. vizualizare - U pasiv;
2. cercetare teoretică și aplicativă - cu respectarea drepturilor de autor (fiecare resursă este însoțită de o fisă de IPR - copyright, care specifică cine are drept de utilizare și ce fel de drepturi);
3. cercetare cu impact comercial (de analizat fiecare caz în parte și de stabilit condițiile în care resurse produse în proiect pot fi cedate industriei ori serviciilor).

b. Interogarea bazei de date
După ce U formulează o cerere ca o listă de cuvinte cheie sau completează un formular:

[3] Necesită personal dedicat care să-i asigure continuitate.

- pe descriere: P încearcă o potrivire între cuvintele cheie furnizare de U și câmpul *Descriere* al resurselor;
- pe metadate: P încearcă o potrivire între câmpurile de metadate completate de U în cererea sa și cele existente în baza de date;
- pe conținut, prin interfețe specializate web.

c. Lansarea în execuție a unei resurse ca serviciu web;
d. Încărcare a unei resurse;
e. Ștergerea unei resurse;
f. Upgradare/updatare a unei resurse (ștergere + încărcare);
g. Conversia de format a unei resurse;
h. Back-up (salvare periodică a bazei de date a Portalului);
i. Alte funcționalități auxiliare:

v. Panou de administrare securizat;
vi. Design responsiv (adaptat pentru dispozitive mobile);
vii. Newsletter (opțional);
viii. Formular de contact cu o adresă de email unde să poată scrie utilizatorii externi;
ix. Forum de discuții/chat;
x. RSS;
xi. Google Analytics;
xii. Trimitere email la adresa de contact.

Funcțiile principale ale Portalului realizează operații de genul celor descrise mai jos și în Figura 1.

```
*operate_platform:
for each document {
      U: upload document;
      P: verifică compatibilitate de format între resursă și standard;
          if true = insert_into_db;
          else =>*convert_format + *insert_into_db);
                  }
*insert_into_db:
  { separă text de speech;
     trece fiecare componentă prin lanțul de prelucrări specifice =>
      if FAIL: generează listă de erori!
      else: indexează (dacă e cazul) }

*update_global_db_statistics:
// opțiuni: fiecare resursă are datele ei globale și la interogare se adună
SAU există aceste sume deja făcute și la interogare ele doar se afișează
{ Resursa textuală => #elemente lexicale, #verbe etc., #semne de punctuație,
   Resursă vocală => #minute/ore de înregistrări, #persoane vorbitoare,
statistici pe categorie de vârstă, accent, contribuții partener etc.
```

```
}

*convert_format:
// input: un format raportat de partenerul contributor
// output: formatul stardard ReTeRom (v. secțiunea 4)
// prelucrare: convertorul de formate trebuie să apeleze programele de
conversie specifice fiecărei perechi input-output[4]
```

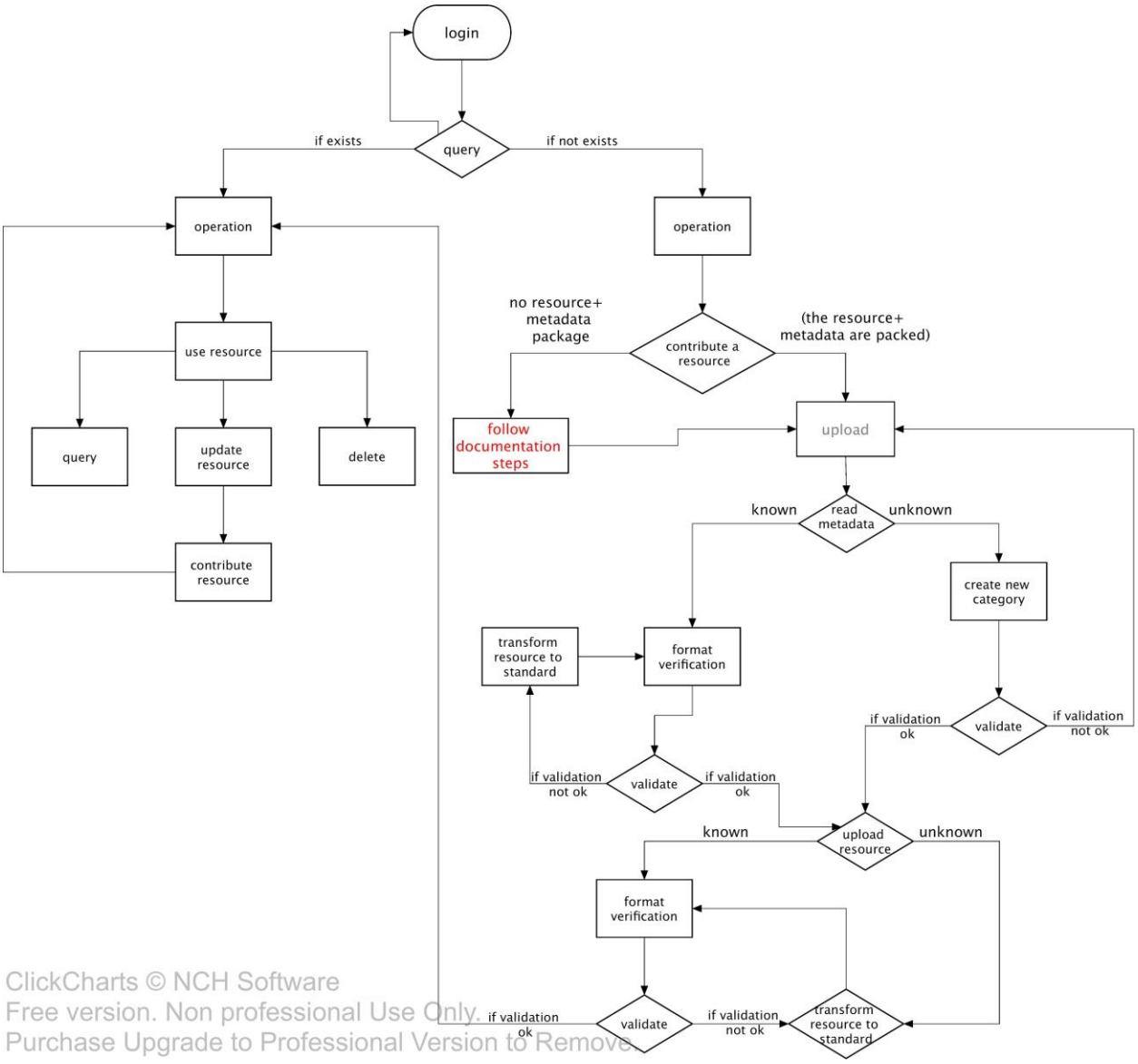

Figura 2: Funcționalitatea Platformei - variantă (preliminară)

4. Propunere de standard COBILIRO

[4] Despre convertoare de formate: http://oxgarage.tei-c.org/#

Pentru compatibilitate cu resurse internaționale, dorim ca formatul de reprezentare al datelor să fie cât mai apropiat de standardul TEI-P5[5] (Sperberg-McQueen and Burnard, 2018) sau de baseline-ul definit în proiectul Parthenos (Romary *et al*., 2017).

Ce urmează reprezintă o sugestie de standard pentru înregistrările corpusului care vor popula platforma ReTeRom/COBILIRO. Acest standard este necesar pentru uniformizarea contribuțiilor membrilor consorțiului ReTeRom și în vederea proiectării tehnologiei de conversie de metadate/înregistrări sonore/înregistrări textuale.

La nivelul metadatelor, această propunere combină informații preluate din (Li and Yin, 2007), cu rezultatul analizei făcute în secțiunea 3 a livrabilului A1.1 și în livrabilul A1.2, pe un schelet TEI P5 (Sperberg-McQueen and Burnard, 2018). La nivelul adnotării înregistrărilor, convențiile noastre urmează îndeaproape indicațiile TEI, cu câteva adaosuri necesare descrierii explicite a corpusului bimodal, tip de document care nu are o secțiune dedicată în ghidul TEI. Limbajul de notare este XML. Atributele unui element XML, pot fi scrise și ca subelemente, caz în care valorile lor apar drept conținut. În general însă vom adopta o astfel de scriere numai în cazurile în care e nevoie ca un atribut să aibă marcate alte atribute. Pentru a distinge între elemente XML, atribute și valorile lor, atunci când apare referit în textul acestui referat, vom nota un nume de element între paranteze unghiulare, un nume de atribut - prefixat cu @ și o valoare - între ghilimele. Pentru a simplifica notațiile, nu vom adopta o ierarhie de clase, precum cea recomandată de TEI P5.

În interesul proiectului ReTeRom este ca înregistrările vocale să fie prezentate în corpus în pereche cu transcrierile lor textuale. Unitatea de bază în alinierea vorbire-text este fraza (ori întinderea ei minimală - propoziția). Inferior limitei de frază/propoziție, vorbirea poate fi segmentată în unități morfologice (cuvânt), fonologice (fonem), prozodice (*pitch*, creștere ori descreștere a frecvenței fundamentale) ori sintactice (grup nominal, clauză etc.), deși ghidul TEI nu precizează nume pentru astfel de segmente.

Un document (element <TEI>) trebuie să conțină o secțiune <teiHeader>, care furnizează informații contextuale detaliate, cum ar fi: sursa obiectului, identitatea participanților (în condițiile respectării constrângerilor de confidențialitate), tipul de vorbire (spontană sau voce-în-citire), aspecte privind condițiile tehnice ale înregistrării etc. (în general, ceea ce în secțiunile 2 și 3 am numit metadate). Fiecare obiect are în constituția sa o secvență de perechi <speech> <text>, unde <speech> reprezintă înregistrarea sonoră, iar <text> - transcrierea ei textuală. Înregistrările sonore pot fi: conversații între un număr mic de persoane, prelegeri, piese de teatru, emisiuni Radio sau TV difuzate, interviuri efectuate pe stradă, în casă, în mijloace de transport etc, înregistrări speciale realizate în condiții de laborator (camere anecoide), audiobook-uri etc.

TEI Guidelines recomandă ca un document:

- să conțină un <text> coeziv;
- să conțină o înregistrare care să acopere o întindere de timp contiguă, fără întreruperi semnificative (microfon deschis o singură dată);
- să poată fi descrisă printr-un singur <teiHeader>.

Se întâmplă însă adesea ca aceste cerințe să nu poată fi îndeplinite simultan. De exemplu, audio-book-urile, înregistrările din piese de teatru și chiar unele emisiuni radio/TV pot fi realizate în sesiuni de lucru succesive, deci încalcă cerința întinderii de timp contigue. Alte documente pot include culegeri de fraze disparate, fără legătură între ele, prea multe pentru a li

[5] http://www.tei-c.org/

se atribui fiecăruia un header propriu, deci încalcă cerința textului coeziv. Ca urmare, cerințele de mai sus trebuie considerate doar orientative și nu vor fi impuse în corpusul ReTeRom.

Corpusul ReTeRom este format din documente sau obiecte (<tei>), fiecare în parte fiind, la rândul lor, segmentat în unități. Exemplificăm în Figura 3 structura recomandată a unui document/obiect al corpusului bimodal.

```
<tei>
<teiHeader/>
<unit>
   <speech>...</speech>
   <text>...</text>
</unit>
<unit>
   <speech>...</speech>
   <text>...</text>
</unit>
…
</tei>
```

Figura 2: Structura unui document bimodal

Uneori una sau alta dintre modalități poate fi segmentată suplimentar față de nivelul alinierilor <speech><text>. Vom utiliza marcajele **<s></s>** (fiecare identificate printr-un @**id** unic) pentru a delimita astfel de subunități. Spre exemplu, să considerăm cazul unui document în care alinierile sunt la nivel de paragraf, dar componenta <text> conține o segmentare la nivelul de frază. Reprezentarea unui document de acest tip document poate fi realizată astfel:

```
<tei>
<teiHeader/>
<unit>
   <speech>...</speech>
   <text>
     <s id=”id1”>prima frază</s>
     <s id=”id2”>a doua frază</s>
     <s id=”id3”>a treia frază</s>
   </text>
</unit>
<unit>
   <speech>...</speech>
   <text>...</text>
</unit>
…
</tei>
```

Figura 3: Exemplu de notație în care componenta <text> este segmentată suplimentar față de componenta <speech>

La nivelul fiecărui document, înregistrările sonore pot fi depozitate în corpus ca fișiere WAV sau MP3. Tipul acestora este marcat în atributul @**speechFileType** incorporat structurii

elementului <teiHeader>. Segmentarea semnalului vocal poate fi realizată în mai multe moduri: prin separarea în fișiere WAV/MP3 a fiecărei componente <speech> (să numit acest mod **“file”**), sau prin marcarea bornelor temporale de început-sfârșit ale componentelor <speech> într-un fișier unic asociat unui obiect <tei> (modul **“start-stop”**). Elementele <speech> trebuie să includă: în primul caz - atributul @**speechFile**, care identifică numele fișierului conținând înregistrarea vocală, iar în cel de al doilea caz - atributele @**start** și @**stop**, cu valori reale, reprezentând momentele de început și de sfârșit ale unității vocale. Dacă modul de segmentare este **start-stop**, numele fișierului în care referă bornele @**start** și @**stop** ale elementelor <speech> sunt date într-un atribut @**speechFile** incorporat elementului <teiHeader>. Pentru a identifica în corpusul ReTeRom prima ori a doua opțiune de marcare a segmentării, recomandăm ca elementele <teiHeader> să includă un atribut @**speechSegmentation**, cu una din valorile: **“file”**, respectiv **“start-stop”**. Evident, tipul de segmentare trebuie să fie unitar în lungimea unui obiect.

```
<tei>
<teiHeader>
     <speechSection speechSegmentation=”file” speechFileType=”wav”/>
</teiHeader>
<unit>
   <speech speechFile=”file1.wav”>...</speech>
   <text>...</text>
</unit>
<unit>
   <speech speechFile=”file2.wav”>...</speech>
   <text>...</text>
</unit>
…
</tei>
```

Figura 4: Exemplu de notare a unui semnal de vorbire în format WAV și tipul de segmentare **“file”**

```
<tei>
<teiHeader>
     <speechSection speechSegmentation=”start-stop”
speechFileType=”wav” speechFile=”my-file.wav”/>
</teiHeader>
<unit>
   <speech start=”0.0000” stop = ”8.5673”>...</speech>
   <text>...</text>
</unit>
<unit>
   <speech start=”8.5673” stop=”15.6652”>...</speech>
   <text>...</text>
</unit>
…
</tei>
```

Figura 5: Exemplu de notare a unui semnal de vorbire în format WAV și tipul de segmentare **"start-stop"**

Dacă un obiect ReTeRom include și alinieri la nivel subfrastic (propoziție, grup, cuvânt, fonem) acestea trebuie detaliate ca atare, modelul de aliniere fiind unul derivat din cel start-stop, sugerat mai sus.

Mai jos dăm o propunere de detaliere a metadatelor din <teiHeader>. Următoarele subelemente sau atribute pot fi incorporate elementului XML <teiHeader> (cele care nu necesită niveluri incluse pot fi reprezentate ca atribute):

@**colection**: (opțional) dacă obiectul face parte dintr-o colecție mai mare de obiecte similare, numele acestei colecții;

<**title**>: (obligatoriu) titlul resursei, cu atribute:
@**ident** (obligatoriu) - identificator sau acronim unic,
@**longTitle** (opțional) - un nume lung;

@**description**: (opțional) descriere generală a resursei, a aplicabilității, a modului recomandat de folosire;

@**language**: (obligatoriu) limba în care a fost creată resursa, valoare în ReTeRom: **"ro"**;

@**contributor**: (obligatoriu) partenerul care a urcat resursa în platforma ReTeRom;

<**addDataSection**>: (opțional) secțiunea în care se descriu componentele metadata și adnotări ale obiectului, cu atributele:
@**metadataCreator**: (opțional) entitatea (persoană, organizație) care a creat metadatele;
@**annotationCreator**: (opțional) entitatea (persoană, organizație) care a creat adnotările, inclusiv alinierea vorbire-text;
<**annotationLevel**>: (opțional) în lipsă, valoarea implicită e: **"sentenceAlign"**; alte valori: **"wordAlign"**, **"phonemAlign"**, **"prosody"** etc.

<**speechSection**>: (oligatoriu) secțiunea în care se descrie componenta vorbire a obiectului, cu detalierea:
@**speechCreator**: (obligatoriu) entitatea (persoană, organizație) care a creat resursa vocală;
@**acusticMedia**: (opțional) **"inStudio"**, **"inTheOpen"**, **"inCar"** etc.
@**duration**: (obligatoriu) durata în ore/minute/secunde a înregistrării vocale;
@**samplingFrequency**: (obligatoriu) frecvența de eșantionare a semnalului vocal (în KHz);
@**rezolution**: (obligatoriu) rezoluția analog digitală a convertorului, numărul de biți întrebuințat pentru reprezentarea semnalului analogic (în bps);
@**recordDate**: (opțional) o dată a înregistrării, în orice format;
@**recordTime**: (opțional) momentul ori intervalul de timp al înregistrării, în orice format;

@**equipment**: (opțional) detalii tehnice despre echipamentul folosit în înregistrarea audio;

@**broadcast**: (opțional) detaliază emisiunea radio/TV în care înregistrarea a fost difuzată;

@**speechFileType**: (obligatoriu) modul de reprezentare a semnalului vocal, cu valorile **“wav”**, **“mp3”**;

@**speechSegmentation**: (obligatoriu) cu una din valorile: **“file”**, respectiv **“start-stop”**;

@**speechFile**: (opțional) numai pentru cazul @speechSegmentation = “start-stop”, conține numele fișierului în care pointează bornele @start și @stop ale elementelor <speech> ale unităților;

<**speaker**>: (opțional) conține informații despre vorbitori. Fiecare vorbitor care apare într-o înregistrare are dedicat un element cu următorul conținut:

@**speakerName**: (opțional) un nume utilizator (nu neapărat cel real), pentru identificarea vorbitorului; un ID de acest tip trebuie referit în secțiunile <speech> ale fiecărui <unit>;

@**speakerGender**: (opțional) **“male”**, **“female”**;

@**speakerAge**: (opțional) **“<10”**, **“10-20”**, **“20-30”**, **“30-40”**, **“40-50”**, **“50-60”**, **“60-70”**, **“70-80”**, **”80-90”**, **“>90”**;

@**speakerAccent**: (opțional) **“Moldavian-Romania”**, **“Moldavian-RepMoldova”**, **“Transylvanian”**, **“Oltenian”** etc. Accentul poate fi identificat și printr-o localitate;

<**textSection**>: (opțional) secțiunea în care se descrie componenta textuală a obiectului, cu detalierea:

@**textCreator**: (opțional) entitatea (persoană, organizație) care a creat resursa textuală;

@**respStmt**: (opțional) o declarație de responsabilitate privind conținutul intelectual al textului, în cazul tipurilor de înregistrări voce-în-citire; în caz de nevoie acest atribut poate fi transformat într-un element <respStmt> care să detalieze informații despre un grup de persoane, editură ori organizații care dețin drepturi de autor privind producerea ori utilizarea materialului;

@**textFormat**: (opțional) formatul (obligatoriu TXT) al reproducerii textuale, cu valorile: “UTF8”, “UTF16”; dacă acest atribut lipsește, valoarea implicită e “UTF8”;

@**distribution**: (obligatoriu) regimul de distribuție, valori: tipuri de contracte (ex. https://creativecommons.org/licenses/by-nc/3.0/);

5. Concluzii

Prezentul raport descrie o propunere de structură și funcționare a Portalului proiectului ReTeRom/COBILIRO, care să facă față cerințelor de stocare și acces a resurselor bimodale deținute de partenerii proiectului, precum și în perspectiva utilizării lui deschise pentru activități de cercetare dedicate prelucrării limbajului natural. În partea a 3-a a raportului propunem un standard de reprezentare a datelor corpusului bimodal, inspirat din standarde internaționale și care să ofere un echilibru între completitudinea informațiilor și simplitate.

Referințe

Li, Ai-jun and Zhi-gang, Yin (2007). Standardization Of Speech Corpus. In: Data Science Journal, Volume 6, Supplement, 18 November.

Sperberg-McQueen, C.M. and Burnard, L. (2018). Original editors, revised and expanded under the supervision of the Technical Council of the TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 3.3.0, last update: 31st January 2018, revision: f4d8439.