Raport final la sfârșit de proiect COBILIRO, proiect component al proiectului complex ReTeRom

Faza de predare: Aprilie 2021
Autori: Anca-Diana Bibiri, Șerban Boghiu, Dan Cristea, Daniela Gîfu, Mihaela Onofrei, Andrei Scutelnicu (UAIC)

I. Rezumatul etapei IV

Proiectul component COBILIRO este coordonat de Universitatea „Alexandru Ioan Cuza" din Iași.

În ultima etapă a primului proiect component COBILIRO, colectivul UAIC implicat în proiect și-a propus următoarele obiective:

- adăugarea a încă două descrieri de proiecte ori aplicații care să exploateze tehnologiile dezvoltate în ReTeRom;
- diseminarea cât mai eficientă a rezultatelor obținute în proiect;
- operarea unor modificări în standardul COBILIRO al resurselor voce-text (standard raportat în livrabilul 1.3 din Faza I);
- îndepărtarea unor bug-uri și îmbunătățirea funcționalității Portalului COBILIRO;
- adăugarea de noi resurse în Platformă.

Toate aceste activități sunt raportate în prezentul livrabil.

II. Rezultatele Etapei IV

1. Noi descrieri de aplicații ale tehnologiilor dezvoltate în proiectul ReTeRom

Aceste descrieri fac obiectul livrabilului A4.1.

2. Diseminarea rezultatelor

Lista completă a acțiunilor de diseminare efectuate în această perioadă este detaliată în livrabilul A4.2.

3. Modificări în standardul COBILIRO

Standardul COBILIRO a fost definit în livrabilul ***Proiectarea funcțională și arhitecturală a infrastructurii care va găzdui resursele și instrumentele de prelucrare și acces ale consorțiului și realizarea unui prototip***, din cadrul Activității A1.3, predat în noiembrie 2018, cu actualizări ulterioare în martie 2020. Standardul COBILIRO descrie structura unui document bimodal incorporabil în Platforma COBILIRO. Orice astfel de document este constituit din două părți: un header și conținutul efectiv.

Prima modificare se referă la modul de adnotare a textului în varianta “file” de memorare a resursei (cea în care fiecare unitate de vorbire este redată într-un fișier separat, textul conținând transcrierea acestor segmente sonore). Până acum standardul prevedea ca segmentele de text să fie reproduse direct între etichete `<text></text>`, ca în Figura 1 de mai jos (care reproduce Figura 4 din livrabilul A1.3):

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

Figura 1: Exemplu de notare a unui semnal de vorbire (formatul aici este WAV) și tipul de segmentare **“file”**: secțiunile `<text>` includ direct secvențele de text

Pe considerentul că anumite resurse sunt constituite din perechi înregistrare vocală - text transcris, în care textele au o întindere semnificativă, cât și pentru a uniformiza notația componentelor vocale și a celor textuale, în noua variantă se permite ca segmentele textuale să fie reproduse și în fișiere distincte, în pereche cu cele de voce. În această variantă, secțiunile `<text>` indică URL-uri în care se află secvențele de text și o uzanță recomandată este ca numele fișierelor de vorbire și de text să fie identice (evident, mai puțin extensiile), ca în Figura 2.

Cea de-a doua modificare a standardului se referă la cazul în care o resursă poate include mai multe variante de segmentare. Acestea pot acum fi descrise ca secțiuni `<units/>` distincte, fiecare fiind caracterizată de un număr de variantă (v. Figura 3). Fișierul `speechFile` este același în toate variantele, pe el fiind “decupate” diferitele transcrieri textuale. Nu este obligatoriu ca segmentarea textuală să acopere complet înregistrarea vocală, ceea ce înseamnă că borna `stop` a unui `<unit>` poate să nu fie identică cu borna `start` a următorului `<unit>` (acestea sunt, de regulă, cazurile în care nu întreaga înregistrare vocală a fost decodificată de modulul S2T). De asemenea, textul unui segment poate fi reprodus direct în `<unit>` (ca în exemplul de mai jos, ori ca în Figura 1), sau indicat ca URL în atributul `textFile` (ca în Figura 2).

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

Figura 2: Variantă de notare a unui semnal de vorbire (formatul aici este WAV) și tipul de segmentare **“file”**: secțiunile `<text>` indică URL-uri în care se află secvențele de text

```
<tei>
<teiHeader>
     <speechSection speechSegmentation=”start-stop”
speechFileType=”wav” speechFile=”my-file.wav”/>
</teiHeader>
<units variant-ID=”...”>
  <unit>
    <speech start=”0.0000” stop = ”8.5673”/>
    <text>...</text>
  </unit>
  <unit>
    <speech start=”10.2233” stop=”15.6652”/>
    <text>...</text>
  </unit>
</units>
<units variant-ID=”...”>
  <unit>
    <speech start=”0.0000” stop = ”7.2345”/>
    <text>...</text>
  </unit>
  <unit>
    <speech start=”7.2345” stop=”15.6652”/>
    <text>...</text>
  </unit>
</units>
</tei>
```

Figura 3: Exemplu de notare a unui semnal de vorbire (formatul aici este WAV) și tipul de segmentare **“start-stop”**, cu două variante de segmentare

În Figura 4 este redat un exemplu de marcaje XML care însoțesc o resursă (resursa exemplificată aici are numele “SoRoEs Timișoara”).

```
<tei>
  <teiHeader Description="SoRoEs Timișoara" Collection="SoRoEs"
Keywords="Timișoara" Language="ro"
Contribuitor="anca.bibiri@gmail.com" Distribution="intern">
    <speechSection SpeechCreator="UAIC_ICI_DISU" AcousticMedia="în
spațiu liniștit" Duration="5 minute" SamplingFrequency="-"
Resolution="-" RecordDate="2016-07-14" RecordTime="11:00"
Equipment="Professional Marantz recorder" Broadcast="Înregistrare pe
teren_Timișoara" SpeechFileType="wav" SpeechSegmentation="File">
      <speaker SpeakerName="9T3a" SpeakerAccent="bănățean"
Gender="Female" Age="Between40And50" />
    </speechSection>
    <textSection TextCreator="UAIC_ICI_DISU"
TextFormat="UTF8"></textSection>
    <dataSection MetadataCreator="UAIC_ICI_DISU"
AnnotationCreator="UAIC_ICI_DISU" AnnotationLevel="SentenceAlign" />
    <title Ident="S5_SoRoEs-Timișoara" LongTitle="SoRoEs-Timișoara:
9T3a" />
  </teiHeader>
  <units>
    <unit>
      <speech SpeechFile="9T3a_55.wav" />
      <text TextFile="9T3a_55.txt">
    </unit>
    <unit>
      <speech SpeechFile="9T3a_56.wav" />
      <text TextFile="9T3a_56.txt">
    </unit>
    ...
    <unit>
      <speech SpeechFile="9T3a_62c.wav" />
      <text TextFile="9T3a_62c.txt">
    </unit>
  </units>
</tei>
```

Figura 4: Marcaje <teiheader> și <units> ale unei resurse stocate pe Portalul COBILIRO

4. Îmbunătățiri aduse Portalului COBILIRO

Pentru îndepărtarea unor bug-uri, pentru îmbunătățirea funcționalității Platformei, cât și pentru ergonomizarea accesului utilizatorilor la serviciile Platformei, în această perioadă au fost operate peste 30 de intervenții în codul interfeței și în structura bazei de date, cele mai importante fiind descrise mai jos:

4.1 Îndepărtarea unor erori și îmbunătățiri în funcționarea Platformei și a Interfeței COBILIRO

a. O salvare comandată după o editare provoca un cod de eroare

Eroarea a fost îndepărtată.

b. Interfața afișa în duplicat câmpul "Reprezentarea semnalului"

Eroarea a fost îndepărtată.

c. Interfața indica "access blocat" la tentativa de editare a resurselor

Eroarea a fost îndepărtată.

d. Vizualizare eronată a unui câmp

Indicatorul "Nivel de adnotare" a fost redenumit "Nivel de segmentare".

e. Validări de tip la încărcarea resurselor

La încărcarea resurselor se filtrează extensiile fișierelor text/sunet, astfel încât doar fișierele cu extensiile corecte sunt acum afișate pentru a fi încărcate.

4.2 Modificări pentru îmbunătățirea ergonomiei utilizării Platformei

a. Încărcare resursă prin parcurgerea unui director de perechi de fișiere voce-text

La urcarea unei resurse în interfață (*upload*), utilizatorul selectează o secvență de fișiere text și sunet din computerul propriu (minimum o pereche). Aceste fișiere trebuie să aibă nume identice pentru componentele vocale și textuale, diferențiate prin extensiile WAV/MP3, respectiv TXT. După urmarea în Platformă resursei i se va crea un unic `<teiheader>`, fiecare pereche fiind marcată între etichete `<unit></unit>`, ca în Figura 2.

b. Completarea automată la upload a unor câmpuri din `<teiHeader>`

Figura 5 indică câmpurile secțiunii `<teiHeader>` => `<speachSection>` care sunt completate automat la upload-ul resursei.

```
<teiHeader
  <speechSection Duration SamplingFrequency Resolution SpeechFileType/>
</teiHeader>
```

Figura 5: Câmpuri calculate și completate automat în operațiunea de upload a unei resurse

c. Indicare operațiune în derulare

Pentru a nu induce impresia de blocare a Platformei după lansarea unor anumite operațiuni din interfață, utilizatorul este acum atenționat (printr-un *spinner*) că Platforma lucrează; util cu precădere în operațiuni de încărcare ori descărcare a unor fișiere de dimensiuni mari.

d. Facilitare descărcări masive

În loc de a selecta fiecare fișier în parte, utilizatorul poate acum să bifeze o casetă pentru a marca toate fișierele în vederea descărcării lor în bloc.

e. Micșorarea duratei de reacție a Interfeței (*caching la deschiderea unei resurse*)

Anumite date afișate de Interfață în secțiunea Detalii (durata unei înregistrări vocale, frecvența de eșantionare a semnalului, rezoluția) erau calculate dinamic, ceea ce provoca o reacție întârziată a Interfeței.

Deschiderea secțiunii Detalii a unei resurse se face acum quasi-instantaneu, pentru că operațiunea nu mai presupune calcularea instantanee a parametrilor numerici ai componentelor resursei, afișarea lor făcându-se din valorile memorate în metadate. Înregistrarea ori actualizarea acestor mărimi în metadate se face la încărcarea/editarea resursei.

f. Câmpurile interfeței sunt afișate acum cu diacritice

g. Secvența de câmpuri care apar acum în interfață a fost actualizată

Ordinea în care sunt acum afișate câmpurile în Interfață este cea din Figura 6, urmărind îndeaproape structura specificată în standardul CoBiLiRo.

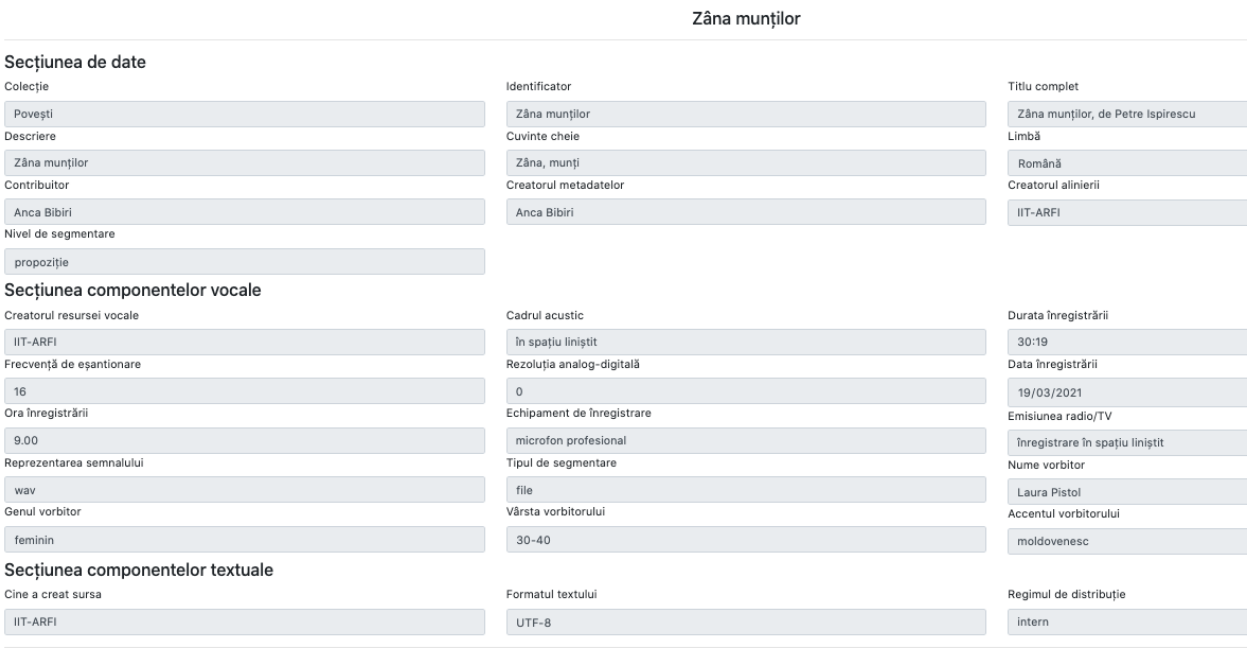

Figura 6: Captură de ecran: Detalii ale unei resurse

h. Statistici

Un număr de valori statistice (raportate în livrabilul A3.3) sunt acum afișabile de interfață. Staticile sunt grupate în valori calculate pe întregul corpus Cobiliro (globale) și pe colecțiile componente (locale).

5. Noi resurse încărcate în Portalul COBILIRO

Următoarele resurse voce-text au fost adăugate în această perioadă în Portalul COBILIRO:

Tabelul 1: Colecțiile recent incluse în Portalul CoBiLiRo-ReTeRom

Nr.	Numele colecției	Numele componentei	Durata (mm:ss)
1		Marin Burlea	56:38
2		Alexandru Călinescu	51:00
3		Theodor Codreanu	55:19
4		Traian Cohal	55:31
5		Dan Cristea	55:21

6	*Ghici cine (re)vine la cină* Total: **16:52:08**	Norina Forna	53:54
7		Ioan Holban	55:34
8		Mircea Radu Iacoban	54:42
9		Aurica Ichim	58:58
10		Alexandru Jula	56:11
11		Sabin Păutza	53:20
12		Aurel Placinschi	54:50
13		Dumitru Popescu	54:00
14		Liviu Suhar	55:53
15		Grigore Tinică	56:14
16		Constantin Tofan	58:07
17		Ella Urmă	56:14
18		Lucian Vasiliu	54:52
19		Sofia Vicoveanca	53:03
20	*Scaune de pluș* Total: **01:37:04**	Ion	42:03
21		Bătrâna	24:25
22		Discuția	11:37
23		Uranus	18:59
24	*Povești* Total: **10:20:50**	Amintiri din copilărie	17:50
25		Cele douăsprezece fete de împărat și palatul cel fermecat	31:03
26		Ciobănașul cel isteț sau Țurloaiele Blendei	23:05
27		Fata babei și fata moșneagului	15:49
28		Fata cu pieze rele	18:37
29		Fata de împărat și fiul văduvei	11:24
30		Fata săracului cea isteață	17:37

31		Prâslea cel voinic și merele de aur	29:00
32		Fata moșului cea cuminte	09:00
33		Făt-Frumos cel rătăcit	30:20
34		Făt-Frumos cu carâta de sticlă	21:52
35		Făt-Frumos cu părul de aur	30:56
36		Făt-Frumos din lacrimă	51:06
37		Găinăreasa	15:37
38		George cel viteaz	28:47
39		Greuceanu	24:00
40		Hoțu Împărat	29:39
41		Lupul cel năzdrăvan și Făt-Frumos	21:00
42		Prâslea cel voinic și merele de aur	29:00
43		Sarea în bucate	16:28
44		Soacra cu trei nurori	16:00
45		Tinerețe fără bătrânețe și viață fără de moarte	22:14
46		Voinicul cel cu cartea în mână născut	39:00
47		Voinicul cel fără de tată	12:47
48		Zâna munților	30:19
49	George Orwell	*1984*	**01:18:00**
50	ROBIN-RACAI	Vox Populi I Vox Populi II	**85:17:17**
	TOTAL ORE		**115:25:19**

În total, în ultima etapă (septembrie 2020 - aprilie 2021) au fost incluse în Platforma CoBiLiRo un număr de aproximativ 120 de ore de înregistrări vocale dublate de texte transcrise. În anumite cazuri, cum sunt interviurile din colecția *Ghici cine (re)vine la cină*, transcrierile sunt post-editate după banda sonoră (publicate în cartea cu același nume, autor Vasile Arhire, apărută la Editura. Junimea din Iași, în luna aprilie 2021) și reprezintă variante aproximative ale

originalului sonor. Chiar și în aceste condiții, considerăm că ele sunt utile pentru îmbunătățirea tehnologiilor S2T dezvoltate în ReTeRom, pentru că aplicațiile de aliniere dezvoltate de colectivul TADARAV sunt acum capabile să identifice secvențele vocale care corespund îndeaproape textului transcris, să le marcheze și să le folosească ulterior pentru reantrenarea tehnologiei de recunoaștere vocală. Operațiunea, aplicată în mod repetat, poate duce inclusiv la identificarea unor segmente din ce în ce mai lungi de suprapuneri voce-text.

6. Concluzii

Acum, la sfârșitul proiectului complex ReTeRom, platforma construită în cadrul proiectului component CoBiLiRo se constituie într-un mediu online având o funcționalitate și un conținut deosebit de util dezvoltării tehnologiilor de interpretare și generare a vocii umane exprimate în limba română. Utilizatorul cercetător are acum posibilitatea să procedeze la încărcări de resurse bimodale, editări de metadate, descărcări și ștergeri a colecțiilor de resurse. Formatul de stocare a datelor bimodale este unul făcut în standardul intern proiectului, numit standard Cobiliro, el putând fi diferit de cel al datelor de intrare, pentru că interfețe de conversie efectuează automat operațiile de transformare. Contribuitorul de resurse poate încărca printr-o singură operație colecții mari de fișiere constituite din zeci, sute, mii de fișiere perechi voce-text, ele fiind împerecheate corespunzător prin metadate. Portalul oferă priviri globale asupra colecțiilor memorate, sintetizând totaluri și calculând statistici de natură lexicală, iar aplicarea unor filtre în metadatele care însoțesc colecțiile ajută efectuarea de selecții după variate criterii. Pentru realizarea statisticilor de natură lexicală, Platforma apelează lanțuri de prelucrări textuale, corespunzător a două tehnologii diferite, una externă proiectului și cealaltă internă, lanțul creat în proiectul component TEPROLIN.

Cantitativ, colecțiile existente acum în Portal sunt cele promise la începutul proiectului, adică în jur de 500 de ore de înregistrări vocale românești, dublate de texte transcrise. Portalul este operațional (la adresa http://platforma.cobiliro.info.uaic.ro/), utilizarea lui rămânând deschisă celor interesați.

Toate obiectivele incluse în plan la această activitate au fost realizate.