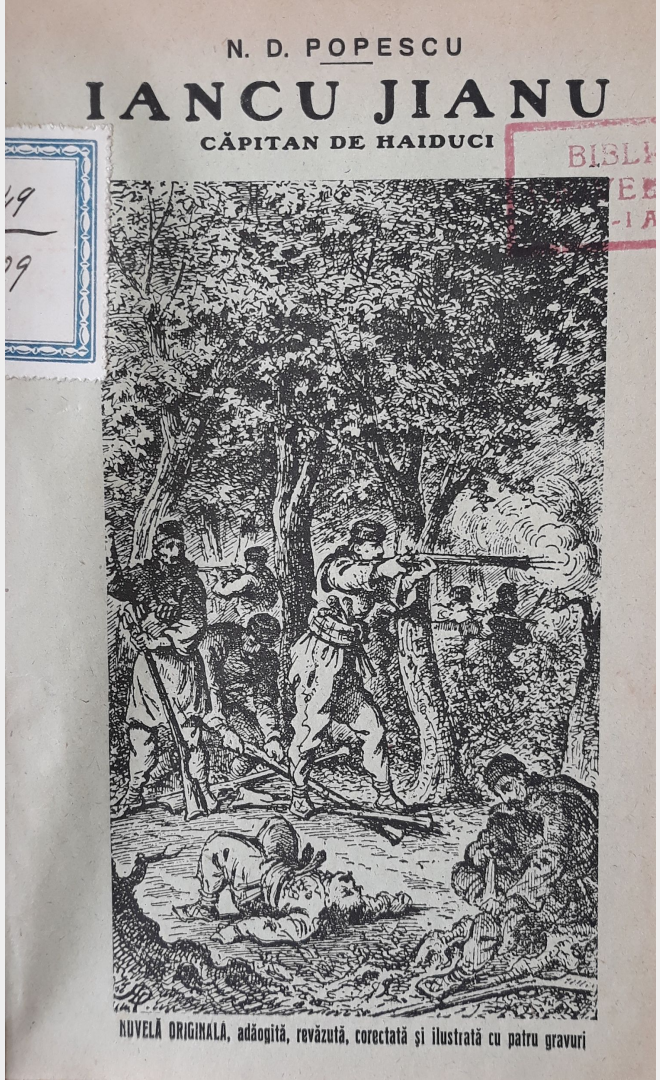
Creating the Romanian ELTeC Collection

-evolution and current status-

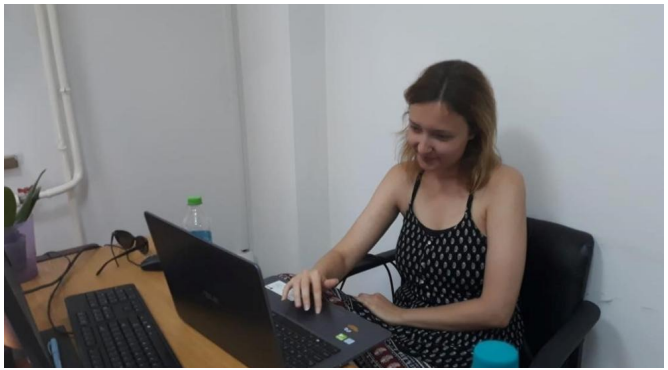
Ioana Lionte and Lucreția Pascariu

- *Organization*
- *Current state*
- *Text selection*
- *Learning TEI*
- *Text processing*
- *Challenges*
- *Future plans*



ORGANIZATION

- Members
- Roxana Patraș (**collection editor and project initiator**)
- Ioana Lionte, Lucreția Pascariu, Alexandra Olteanu (**scanning, OCR, encoding**)
- Alexandra Ruscanu, Laura Ciobanu, Lorena Vlad, Manuela Ursu, Raluca Miron, Emanuela Guțu, Georgiana Apreotesei (**manual cleanup**)



- Step 1: active participation and self-teaching

  ‘See one, do one, teach one’

- Step 2: trainings and short time missions

  Going abroad and learning from professionals

- Step 3: collaborative work

  scanning - OCR - HTML to XML - manual cleanup - encoding

CURRENT STATE

- 80 novels uploaded on GitHub (last release) and Zenodo
- 20 more to go
- E5C of 83,08
- 76 authors (65 male authors, 11 female authors)
- Length (35 short, 29 medium, 16 long)
- Time slot:

  1840-1859: 4 novels

  1860-1879: 14 novels

  1880-1899: 23 novels

  1900-1920: 39 novels

❖ Open issues

title page found in body

lack of author dates

pseudonym mark-up in the TEI header

anonymous writers

```
ERROR: ROM004Dumbravă, Bucura (Fany Seculici/Fanny Szeculicz)(1868-1926)
implausible author dates (Fany Seculici/Fanny Szeculicz)!
WARNING: ROM006 title page found in body : shouldn't it be in front or back?
WARNING: ROM006 title page found in body : shouldn't it be in front or back?
WARNING: ROM006 title page found in body : shouldn't it be in front or back?
WARNING: ROM006 title page found in body : shouldn't it be in front or back?
```

❖ Encoding decisions

Letters: <quote type=”letter”></quote>

```
<quote type="letter">
<p><hi>„Iubite Herdelea, doresc să am răspunsul d-tale precis, în trei zile. Țin mult să te numesc pe d-ta la Pripas. Salutări cordiale. — Cernatony.”</hi></p>
</quote>
```

Diary: <quote type=”diary”></quote>

Quoted poems: <quote type=”verse”></quote>

Volumes:

Chapters:

Subchapters: <milestone unit=”section” n=”x”/>

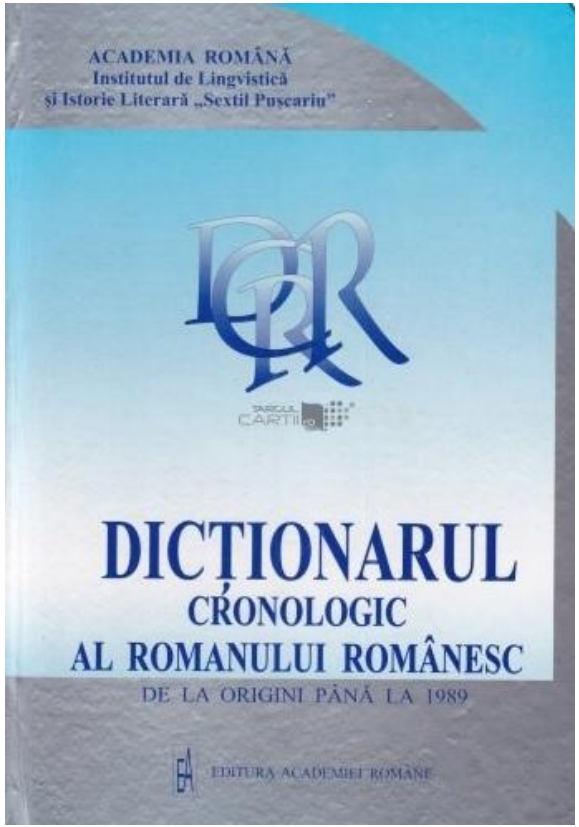
- Embedded tags (volume – chapters – subchapters)

```
<body>
      <div type="group" n="1">
            <div type="titlepage">
                  <p>GLASUL PĂMÎNTULUI</p>
            </div>
            <pb/>
            <div type="chapter" n="1">
    <head>Capitolul I</head>
    <head>ÎNCEPUTUL</head>
            <milestone unit="section" n="1"/>
    <p>1</p>
    <p>Din șoseaua ce vine de la Cîrlibaba, întovără-șind Someșul cînd în dreapta, cînd în stînga, pînă la Cluj și chiar mai departe, se desprinde un drum alb mai sus de Armadia, trece rîul peste podul bătrîn de lemn, acoperit cu șindrilă mucegăită, spintecă satul Jidovița și aleargă spre Bistrița, unde se pierde în cealaltă șosea națională care coboară din Bucovina prin trecătoarea Bîrgăului.</p>
```

TEXT SELECTION

HOW?

- According to the time-slots
- Using DCRR (*The Chronological Dictionary of the Romanian Novel from Origins to 1989*)
- Adapting to ELTeC criteria
- Copyright issues



SOURCES

The Metropolitan Library of Bucharest

"Mihai Eminescu" University Library of Iasi

The County Library of Botosani

Personal micro-collections uploaded on Zenodo under the following labels: "Hajduks Library", "RomanianNovel Library"; "CityMysteries Library"; "BibliotecaDHL_Iasi"

LEARNING TEI

- Self-study
- Training (Würzburg, Galway, Paris)

TEXT PROCESSING

from print to digital

- Started with a small number of already scanned novels
- CZUR scanner (purchased by the university)
- Abbyy FineReader 15
- Oxygen XML Author 21.1
- Notepad++
- jEdit

CHALLENGES

1. infrastructure-bound (underdeveloped library services and librarians' training)

2. digitization policies-bound (scarce open access resources and available formats)

3. text processing-bound (OCR and POS tagging sub-optimal performance on diachronic varieties of Romanian);

4. editing-bound (data on booklength available in page numbers but not in word count);

5. culturally and literary tradition-bound (unbalance of T1, T2, T3, T4; the percentage of female-authored novels)

6. the Romanian team's low training level and lack of experience with xml, epub, hml formats, with collaborative work on platforms such as github, with TEI markup and annotation in general, and with editing software

FUTURE PLANS

- Projects stemming from the ELTeC collaboration

  HAI-RO, PopLite

- The creation of a DHL (Digital Humanities Laboratory) hub at the “Alexandru Ioan Cuza” University of Iasi

Facebook Page:
https://www.facebook.com/DigitalHumanitiesLaboratoryUAIC