

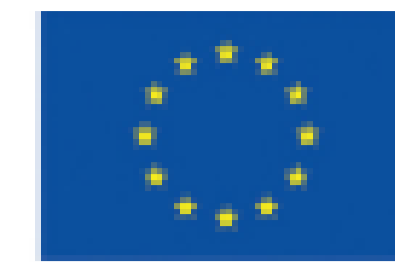
UniDive

2nd General Meeting

University of Naples "L'Orientale"

Naples, Italy, 8-9 February 2024

<https://unidive.lisn.upsaclay.fr/>



Funded by the European Union

REVITALIZING THE HISTORICAL ROMANIAN TEXTS WITH CYRILLIC SCRIPTS

CAFTANATOV OLESEA, MALAHOV LUDMILA AND BUMBU TUDOR

Moldova State University, Vladimir Andrunachievici Institute of Mathematics and Computer Science

The Aim:

The aim of our work is revitalizing the historical Romanian texts with Cyrillic Scripts from the XVII – XX century.

to create linguistic resources of the Historical Romanian

to reissue a folkloric books in latin script that will be used for educational purposes

for philological research, for instance at lexicographical diachronic analysis and others

The Challenges:

We researched various types of historical documents such as: manuscripts, religions books, dialectal text and others. The endeavor to address the linguistic heritage of Romanian history involves tackling several specific challenges, including:

- 1) dealing with a multitude of language evolution periods;
- 2) coping with the scarcity of widely available resources;
- 3) managing the diverse array of alphabets used in historical printings, including mixed Cyrillic-Latin "transition alphabets";
- 4) overcoming the absence of reliable tools for accurately recognizing Cyrillic letters from various historical eras;
- 5) addressing the shortage of lexicons suitable for the time periods of these resources.

А ТРЪЖИЛЕ АДДНЪРЪ ДЕ ЦАРЪ, АТЪТЪ АША НЪМНТЕЛЕ ПО РТЕ ЛЪС ПО РЦИ (ПЕТЕ ТЪТЪ ГЪРЪНА) БОЛНУЪРЕ ЛЪС АПЪРЦИТЪ СПРЕ ФИ-ЕЩЕ КЪЕ КОМНТАТЪ, КЪМШИ АРЪНКАРЪ ШИ ХОТЪРЪРЪ ДЪРН-

XVIII Century

Жълиетта, арътъндвсе iar ла фереастръ.
Треї ворбе лпкъ, избите Ромео, ши апої адио, адио. Дака ведеріле аторълзі тьъ сѣнт вредніче

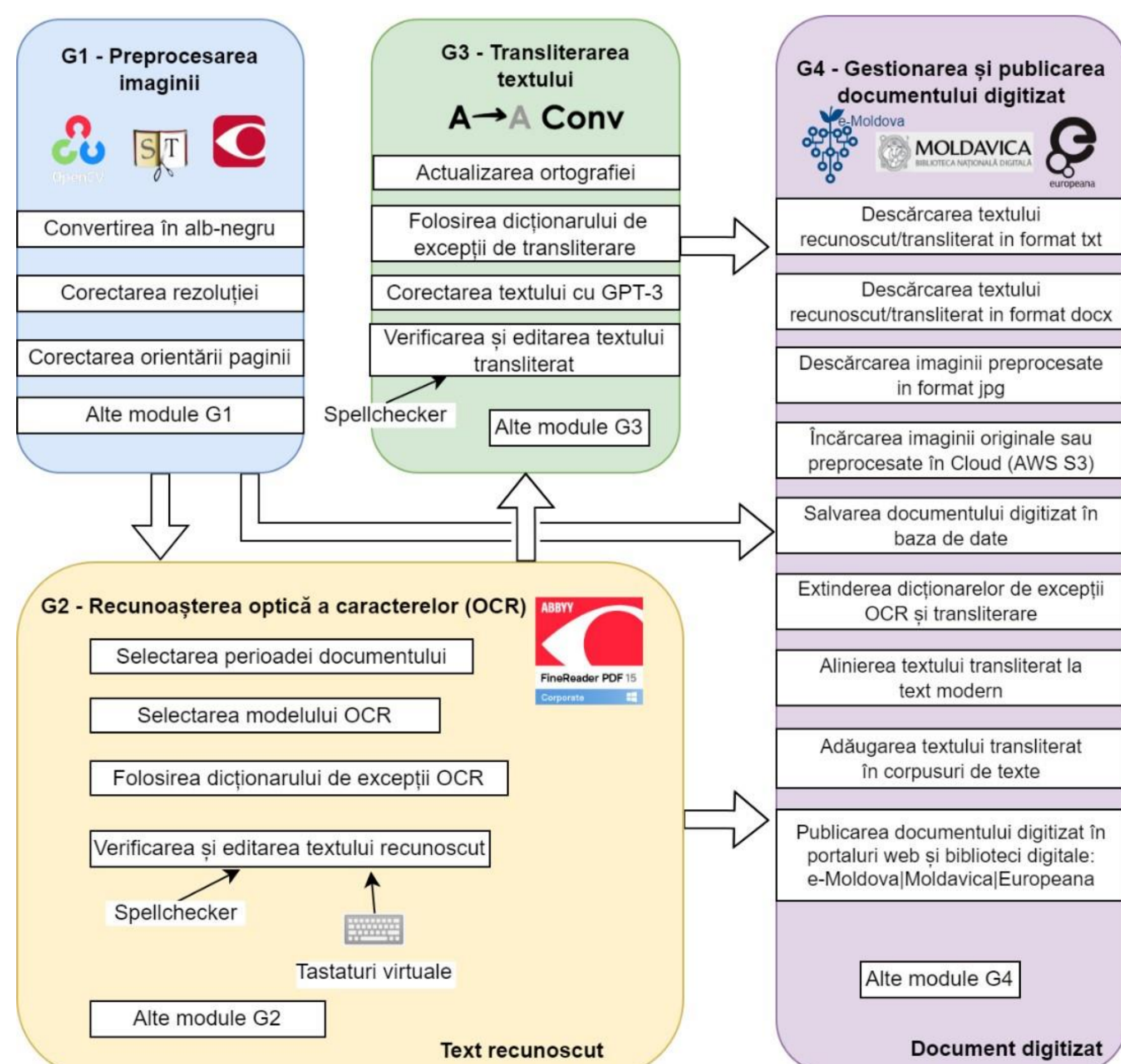
XIX century, mixed alphabet

Уника кестиуне, каре требуе резолватъ, ну аре дежа нич о легътуръ ку кэрэмзиле — кьт де маре поате фи сума нумерелор инверсе челор паре? «Внмулцинд» кестиуня ла дой, обцинем уна

XX Century

The Digitization Platform Architecture:

DP is a web application that features an interactive graphical interface and a set of APIs designed and managed through Django Rest Framework. This application offers seven consecutive steps to enable the recognition of Romanian Cyrillic documents from the XVII-XX centuries, transliteration of the texts into Latin script, editing of recognized/transliterated texts, and downloading or publishing the results. The platform includes: (i) image processing engines, such as ScanTailor, AFR, and OpenCV; (ii) OCR models for Cyrillic characters, trained on datasets gathered from documents printed in the XVII-XX centuries; (iii) a transliteration tool from old Romanian Cyrillic to contemporary Latin script; (iv) virtual keyboards specific to the alphabets used in previously mentioned periods, such as the Romanian Cyrillic alphabet, the transition alphabet, and the Soviet Cyrillic alphabet.



Case Study:

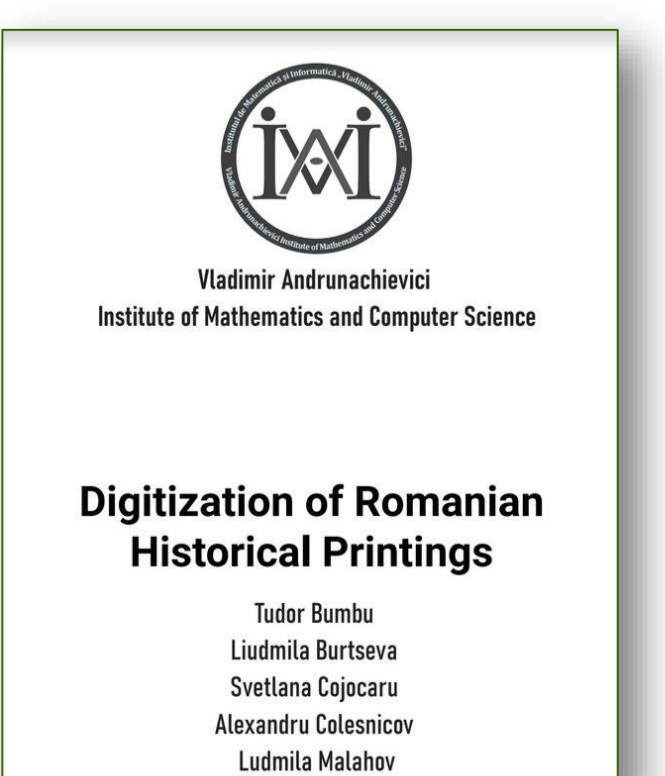
- The main resource is a national folkloric book;
- We used the tool pack developed in our institute to process the book "Folclor din părțile";
- We organized a team of volunteers to illustrate the textual message in order to create and publish the folkloric book
- Recognition accuracy is over 97% in words;
- The transliteration accuracy was about 98%.

Original Text	Recognized Text	Transliterated Text
Кате петричеле'и флнтанъ— Атыя оале ку смънтанъ! Хэй, хэй! Кате лемне суб ватрѣ— Атыя флнтанъ ла фатѣ! Хэй, хэй! Кате пене не куош— Атыя бешь бурдухош! Хэй, хэй! Кате пѣе не касѣ— Атыя галбешь не масѣ! Хэй, хэй! Ешь лелица, ку ковригул, Ко мѣ рупе фригул! Хэй, хэй! Де ла Лухтан Вас. Т., 30 ан, с. Борничен, р-он Кълърань, а Морару С. Г., 223, 12.	Кате петричеле'и флнтанъ— Атыя оале ку смънтанъ! Хэй, хэй! Кате лемне суб ватрѣ— Атыя флнтанъ ла фатѣ! Хэй, хэй! Кате пене не куош— Атыя бешь бурдухош! Хэй, хэй! Кате пѣе не касѣ— Атыя галбешь не масѣ! Хэй, хэй! Ешь лелица, ку ковригул, Ко мѣ рупе фригул! Хэй, хэй! Де ла Лухтан Вас. Т., 30 ан, с. Борничен, р-он Кълърань, Морару С. Г., 223, 12.	Câte petricele'ia flintână— Atâtea sale cu smântână! Hăi, hăi! Câte lemne sub vatră— Atâtea flănci la fată! Hăi, hăi! Câte pene ne cucoș— Atâtea botei burduhoși! Hăi, hăi! Câte pae pe casă— Atâtea galbeni pe masă! Hăi, hăi! Ești leliță, cu covrigul, Că mă rupe frîgul! Hăi, hăi! De la Luchian Vas. T., 30 ani, s. Vornicen, r-mi Călărași, Moraru S. G., 223, 12.



Realization and Future Work:

- A Diachronic Corpus for Romanian (RoDia) - ACL Anthology <https://aclanthology.org/W17-8101/>
- Development of the Digitization Platform HeDy. <https://digitizare.math.md/>
- Manuscript: Digitization of Romanian Historical Printings.



Vladimir Andrunachievici
Institute of Mathematics and Computer Science.
USM
<http://www.math.md>

Caftanator Olesea¹
Malahov Ludmila²
Bumbu Tudor³

1. olesea.caftanator@math.md
2. ludmila.malahov@math.md
3. tudor.bumbu@math.md

5, Academiei street,
MD 2028, Chisinau,
Republic of Moldova