CHALLENGES IN PROOFING THE CYRILLIC MCVRO RESOURCES – EQUABILITY BETWEEN THE TECHNICAL COMPONENT AND THE ROLE OF THE RESEARCHER

DAN BENȚA*, PAULA BUD* ELENA PLATON*, STELIAN PAȘCA-TUȘA*, ELENA ONEŢIU*, ALIN MIHĂILĂ*, FELIX FLOCA*

Abstract The OCR process is a complex one and with interdisciplinary implications. This paper deals with old historical religious manuscripts and old books to offer solutions for the digitalization process. A high rate of success in the OCR process will help researchers from various fields to use relevant content and to value it in their research. When a research in the field of theology or history is conducted, there are many issues, especially when trying to access old manuscripts. There is no digitalized content to access old Romanian manuscripts and researchers need to learn specific old writing. The focus of the present study is on the problematic technical aspects that we encountered in our attempt to establish the equilibrium ratio between the instruments for the capitalization of the information and the contribution of the specialists in such research studies.

Keywords OCR, pattern recognition, historical documents, digital maintenance and control.

Introduction

The results presented in this paper are partial results obtained within MCVRO project – a pioneering one in the Romanian environment, through the characteristics of the research it requires and through the way it values the research. Within this

* *Agora University of Oradea*. benta_dan@yahoo.com.
* *Babes-Bolyai University, Cluj-Napoca*. paulabud_ot@yahoo.fr.
* *Babes-Bolyai University, Cluj-Napoca.* elenaplaton99@yahoo.com.
* *Babes-Bolyai University, Cluj-Napoca*. stelianpascatusa@gmail.com.
* *Babes-Bolyai University, Cluj-Napoca*. elenaonetiu.ot@gmail.com.
* *Babes-Bolyai University, Cluj-Napoca*. alinmihaila@gmail.com.
* *The Lucian Blaga Central University Library of Cluj-Napoca*. fffloca@yahoo.com.

DOI: 10.26424/philobib.2020.25.2.09

project, we plan to exploit the full potential of the Romanian manuscripts and old books from the deposit of the Central University Library from Cluj-Napoca, creating an information instrument that allows both the extraction and the use of the concepts found in the resources that were previously digitized. This desideratum involves a triple purpose: a. creating a virtual library of Romanian manuscripts and old books (MCVRO) at the Central University Library from Cluj-Napoca; b. creating the instruments to extract and adapt the MCVRO content in order to value them within the research of specialists from many fields: theology, economy, philology, history, ethics, philosophy, medicine etc.; c. creating, based on the scanned MCVRO resources, a system to identify the concepts and correlations between them, with an interdisciplinary relevance. These three elements converge in exploiting MCVRO resources as an important element for social activation. To fulfil this desideratum, we considered a complex step of digitizing of the MCVRO resources; the complexity is caused by the heterogeneous nature of the MCVRO corpus which implied multiple and various challenges.

The main goal of the MCVRO project was the digitization of the MVCRO resources, to preserve and as a form to grant access to the content of this type of resources. The digitization process implied the following main stages: scanning, OCR tasks and proofing. We used two main types of resources: documents with Latin writing and documents with Cyrillic writing. Among the documents processed within the project which use Latin characters, we mention as examples: *Schematismus Venerabilis Cleri Dioecesis Magno-Varadinensis Graect-Ritus Catholicorum* (Orade Mare: Typariu lui Joan Tichy, 1822) – BCUCLUJ_FCS_BRV1163; *Schematismus Venerabilis Cleri Dioecesis Magno-Varadinensis Graect-Ritus Catholicorum* (Orade Mare: Typariu lui Joan Tichy, 1824) – BCUCLUJ_FCS_BRV1217. Among the documents processed within the project which use Cyrillic characters, we mention as examples: Ieromonah Constantie, *Întrebărĭ şi răspunsurĭ ṫeologhīċeştĭ ale Sf[â]ntuluĭ Aṫanasīe ċeluĭ Mare*, trans. monk Gherontie from Neamțu Monastery (București, 1821) –BCUCLUJ_FCS_BRV1111 (Cyrillic characters); Episcopul Damaschin, *Învățătura despre șapte Taine* (Râmnic, 1724) – BRV 185 (Cyrillic characters); Gheorghie Râmniceanul, *Rugăċiunĭ cătră Pre Sf[â]nta Născătoarea de D[u]mneẓeu, și pururea Feċioara Marīa* (București: Tipografia Mitropoliei, 1819) – BRV 1048 (Cyrillic characters); Antim al Ierusalimului, *Învățătură părinteasċă*, trans. Meletie of Huși (Iași: Tipografia Mitropoliei, 1822) – BRV 1148; Antonie Monahul, *Mâna lui Damaschin* (Iași, Tipografia Mitropoliei, 1830) – BRV 1487 (Cyrillic characters); *Învățătură pentru ultuirea Vărsatuluĭ* (Chișinău: Tipografia Episcopiei Basarabiei, 1816) –BCUCLUJ_FCS_BRV929C; Eufrosin Poteca, *Cuvinte panigẏrīce și moralnice* (București: Tipografia Mitropoliei, 1826) – BCUCLUJ_FCS_BRV1287; Atanasie al Alexandriei, *Tropariu pentru Melhisedec* (Iași: Tipografia Mitropoliei, 1812) – BCUCLUJ_FCS_BRV814. The stages of scanning and OCR tasks were performed by a library team, and proofing step was assumed mostly by a university team.

1. State of the art

Started in 2012 and with a first deploy of a full-text search engine in 2013, the Austrian National Library conducted a similar OCR process research[1] where the implementation of a full-text-search interface for its historical text collection was defined as a milestone to achieve their strategic goals. This digitalization process began early in other European countries. For example, the National Library of France started mass digitization projects on its collections since 2006. As pointed in A. Ben Salah et al study,[2] their main issue for old documents is that the OCR results are of poor quality and the OCR quality assessment is a real challenge in their field.

There are various studies in the OCR field where different types of documents are used. Authors like B. B. Chaudhuri and C. Adak[3] focus on processing offline handwritten document images where main challenges are that this type of processing produces nonsense character string outputs; their main solutions are: (a) pattern classification and (b) graph-based method for identifying such texts.

The historical text digitalization is considered a real challenge and several issues were identified:

- text is damaged and quality is awfully bad, sometimes with dirt,
- layout is complicated and reading order is different,
- paper is folded or the pages are glued together,
- text is printed on low quality paper,
- text is printed using old fonts and language has different writing variants,
- image quality is not so good.

One of the biggest projects in the OCR field was IMPACT (IMProving ACcess to Text),[4] a project funded by the European Commission under the Seventh Framework Programme (FP7), for 4 years starting from 1st of January 2008 with a 16,5 M Euro budget. It is part of the Cooperation Work Programme for ICT and responds to the fourth challenge in this programme: Digital Libraries and Content. The main objective of the project was to significantly improve the access to historical text and to take away the barriers that stand in the way of the mass digitization of the European cultural heritage. The project started as a collaboration between 26 European Partners to improve OCR software to recognize letters and punctuation

[1] B. K. Und and M. Hintersonnleitner, "Full-text-search in Historical Text Collections. Experiences of the Austrian National Library," *Bibliothek Forschung und Praxis* 39, no 1 (2015): 73-79.

[2] A. Ben Salah et al., "OCR performance prediction using cross-OCR alignment," in *13th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Washington: IEEE Computer Society, 2015), 556-560.

[3] B. B. Chaudhuri and C. Adak, "An approach for detecting and cleaning of struck-out handwritten text," *Pattern recognition* 61, no. 3 (2017): 282-294.

[4] "Impact | Improving access to text," accessed December 12, 2016, http://www.impact-project.eu/home/.

markers to make digital text searchable. Their main guidance was that information is and must remain retrievable. To this respect, a proper way was to transform digital images of the scanned books into electronic text since mass-digitization has become one of the most prominent issues in the library world. IMPACT project identified that there are three main reasons why the digitized material is becoming available too slowly and in too small quantities from too few sources: a) lack of institutional knowledge and expertise, b) the costs for full-featured electronic text of historical documents are much too high and c) automated text recognition, carried out by OCR engines does in many cases not produce satisfying results for historical documents. IMPACT solution was ABBYY. ABBYY was also recognized as the best OCR tool according to the online pool conducted by lifehacker.com.[5] According to their pool with more than 3900 votes, ABBYY received 30.81% of votes while next competitors were Evernote/OneNote with 27.62% votes, Adobe Acrobat with 24.63% of votes, OmniPage with 8.76% of votes, Readiris with 4.14% of votes and Other with 4.04% of votes.

There are situations when alternative solutions are needed for better results. For example, authors like F. Simistira, A. Ul-Hassan, V. Papavassiliou, B. Gatos, V. Katsouros and M. Liwicki[6] report on high-performance OCR experiments using Long Short-Term Memory (LSTM) Networks for Greek poly tonic script and concluded that the character error rate obtained with LSTM varies from 5.51% to 14.68% and it is better than two well-known OCR engines, namely, Tesseract and ABBYY FineReader. LSTM were also preferred by M. R. Yousefi, M. R. Soheili and others[7] to skip the binarization step since in many cases artifacts result in important information loss, by for instance breaking or deforming character shapes. There are papers like those written by C. Clausner and his colleagues[8] where existing OCR systems fit their needs and a baseline for two state-of-the-art OCR systems (ABBYY FineReader Engine 11 and Tesseract 3.03) is given with regard to both text recognition and segmentation/layout analysis performance. Also, additional solutions can be used, such as automatic script identification in archives of documents.[9] So, the approach

[5] "Lifehacker – Tips and downloads for getting things done, Five Best Text Recognition Tools," accessed December 12, 2016, http://lifehacker.com/5624781/five-best-text-recognition-tools.

[6] F. Simistira et al., "Recognition of Historical Greek Polytonic Scripts Using LSTM Networks," in *13th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Washington: IEEE Computer Society, 2015), 766-770.

[7] M. R. Yousefi et al, "Binarization-free OCR for Historical Documents Using LSTM Networks," in *13th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Washington: IEEE Computer Society, 2015), 1121-1125.

[8] C. Clausner et al, "The ENP Image and Ground Truth Dataset of Historical Newspapers," in *13th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Washington: IEEE Computer Society, 2015), 931-935.

[9] S. Kavitha et al., "A Robust Script Identification System for Historical Indian Document Images," *Malaysian Journal of Computer Science* 28, no. 4 (2015): 283-300.

depends on the OCR results obtained since each researcher is looking for results with a lower error rate.

2. Challenges in the proof of MCVRO resources with Cyrillic characters

For the OCR tasks we used ABBYY FineReader (v. 11 and v. 12). The main challenges in this stage concerned unification of the proofreading rules and the technical possibilities of the application regarding the existing fonts. ABBYY FineReader does not have a set of characters to cover the entire range of symbols used within the Romanian Old Books, especially within those with Cyrillic characters. We tested an additional set of fonts, *Cyrillicum Budanum*, created within a project focused on several activities to transpose the Lexicon from Buda into a central database. Furthermore, we contacted the Ukrainian agency of the ABBYY application to initiate a collaboration in order to create a complete set of characters that would include the Cyrillic characters identified in the Romanian old books and which did not match the Cyrillic sets that exist in the sets of this application. After several discussions we had to give up to this collaboration because it would involve additional costs that are not supported by the project. Given the situation it was necessary to accept other solutions, such as:

- Creation of a special font based on the characteristics of the MCVRO resources processed within the project, a font that would include all or the majority of the necessary characters;
- Combination of sets that belong to a font from Symbols;
- Insertion of new fonts created in Abby FineReader;
- Definition of the settings for OCR step to include several languages, which, through their combination, would lead to a higher percent in the recognition of the text.

Unfortunately, all these technical efforts had unsatisfying results after the OCR step, the text of the resources was recognized in such a small percent that it would not justify the subsequent intervention for proofreading without requiring an enormous and unjustified period of time.

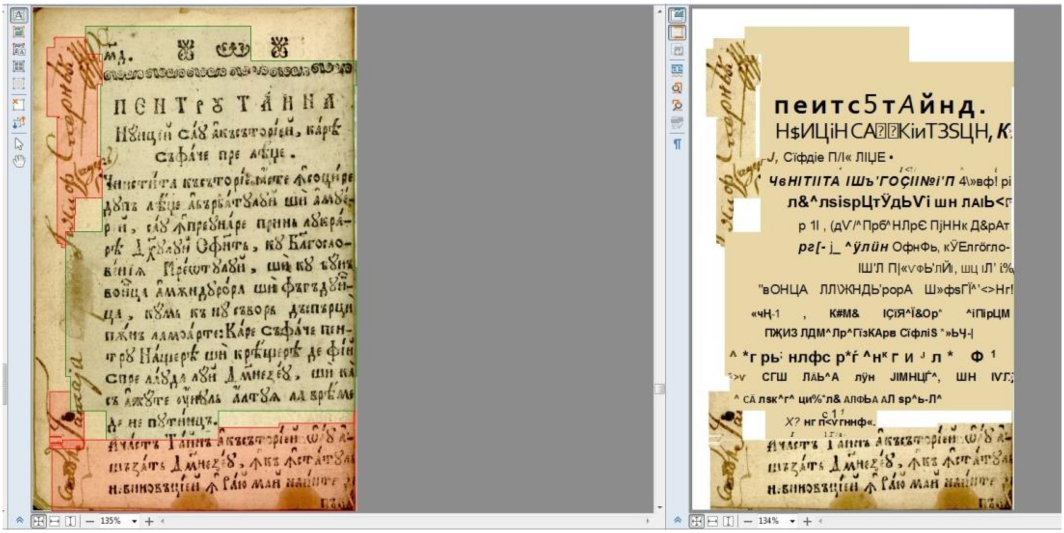
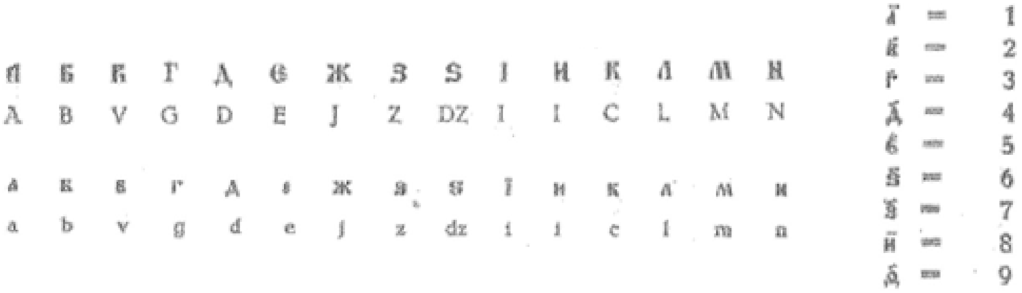


Figure 1. Text with many mistakes combined with image fragments (BRV 185)[10]

In this situation of poor recognition of MCVRO resources with Cyrillic characters, we decided to transliterate the text and attach it as meta-text to scanned image, which became a searchable pdf document.

3. Transliteration of the MCVRO resources – approach and challenges

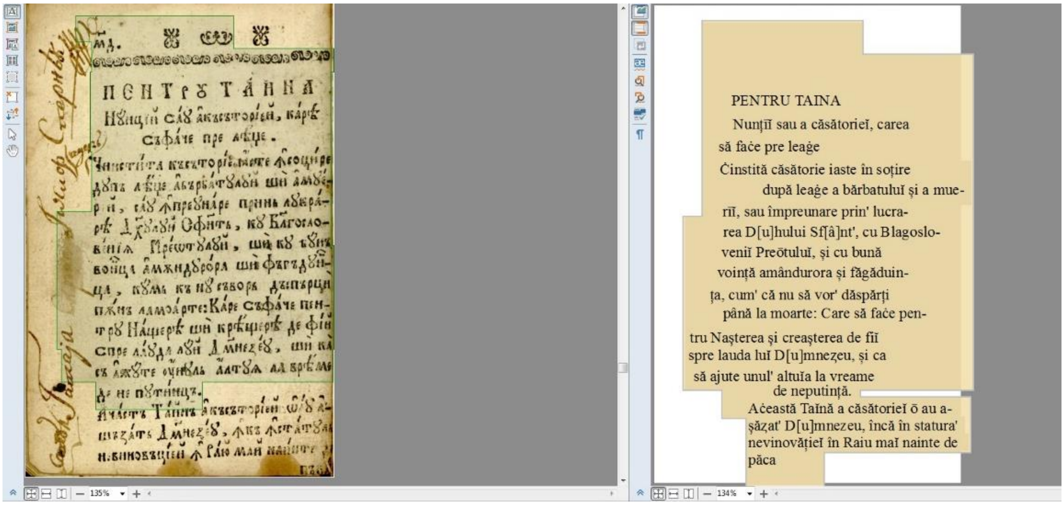
The transliteration was uniform according to the following correspondences **for capital letters, for small letters and for numbers:**



А	Б	В	Г	Д	Є	Ж	З	Ѕ	І	И	К	Л	М	Н
A	B	V	G	D	E	J	Z	DZ	I	I	C	L	M	N
а	б	в	г	д	є	ж	з	ѕ	ї	и	к	л	м	н
a	b	v	g	d	e	j	z	dz	i	i	c	l	m	n

а̃	=	1
в̃	=	2
г̃	=	3
д̃	=	4
є̃	=	5
ѕ̃	=	6
з̃	=	7
и̃	=	8
ѳ̃	=	9

[10] Episcopul Damaschin, *Învățătura despre șapte Taine* (Râmnic, 1724), 49.



PENTRU TAINA
Nunțiĭ sau a căsătorieĭ, carea
să faċe pre leaġe
Ċinstită căsătorie iaste în soțire
după leaġe a bărbatuluĭ și a mue-
riĭ, sau împreunare prin' lucra-
rea D[u]hului Sf[â]nt', cu Blagoslo-
veniĭ Preōtuluĭ, și cu bună
voință amândurora și făgăduin-
ța, cum' că nu să vor' dăspărți
până la moarte: Care să faċe pen-
tru Naşterea și creașterea de fiĭ
spre lauda luĭ D[u]mnezeu, și ca
să ajute unul' altuĭa la vreame
de neputință.
Aċeastă Taĭnă a căsătorieĭ ō au a-
şăzat' D[u]mnezeu, încă în statura'
nevinovățieĭ în Raiu maĭ nainte de
păca

Figure 2. Example of transliteration, from BRV 185, previously presented.[11]

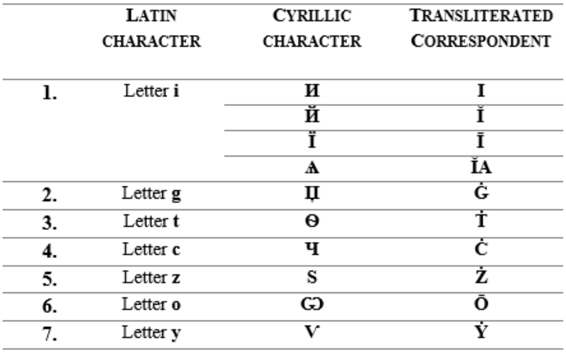
Although it seems a desperate solution in the given situation, the transliteration offers, at a close analysis, more advantages to the researcher than the rewriting of the original text:

- from the perspective of the researchers that may use these digitized resources, it is more likely to initiate a search based on a word written with Latin characters than one written with Cyrillic characters, especially taking into account the variations of words; all these words may be identified easier in the text in a search using Latin characters;
- based on the rules of transliteration available to the researchers together with the access to the digitized resources, the original writing can be recomposed in the case of extremely important words for certain researches;
- the relationship between *transliterated text* and *image* also offers easy access to the original text, at the same time giving the possibility to identify interesting thematic aspects with a simple search in the original text;
- supposing that researchers who access the resources are not all familiar with the Cyrillic characters, the transliteration opens easy access to the content of MCVRO resources, thus offering new possibilities of exploiting the content in a research; otherwise, the activation of the cultural production would be limited to the knowledge of Cyrillic of each user.

Adding character by character from *Symbols* required a huge effort from the team and an extended period. The necessity of this action was imposed by the fact that the default Cyrillic fonts that can be added from the keyboard were incomplete

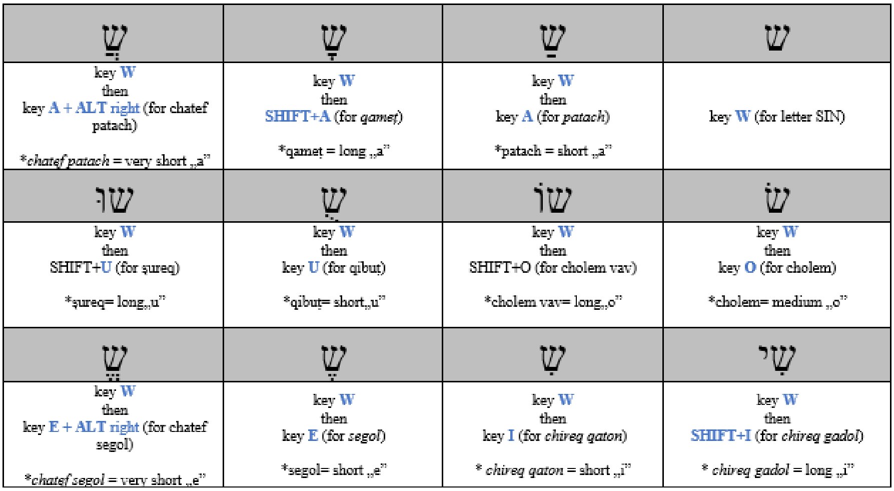
[11] Damaschin, *Învățătura despre șapte Taine*, 49.

and do not cover all the symbols identified in the MCVRO resources. *Symbols* was used for the following characters:



	LATIN CHARACTER	CYRILLIC CHARACTER	TRANSLITERATED CORRESPONDENT
1.	Letter **i**	**И**	**I**
		Й	**Ĭ**
		Ї	**Ī**
		Ѧ	**ĬA**
2.	Letter **g**	**Џ**	**Ġ**
3.	Letter **t**	**Ѳ**	**Ṫ**
4.	Letter **c**	**Ч**	**Ċ**
5.	Letter **z**	**Ѕ**	**Ż**
6.	Letter **o**	**Ѡ**	**Ō**
7.	Letter **y**	**Ѵ**	**Ẏ**

The experience acquired within this approach determined us to develop, as a future solution for this type of process of transliteration, the creation of a special font where additional characters, besides the main corpus of the letter, may be added separately from the keyboard, as is the case of the fonts used for Semitic languages. We offer here an example from Hebrew, where the original text uses consonants, and the vocalic system is superposed additionally over them, as follows:



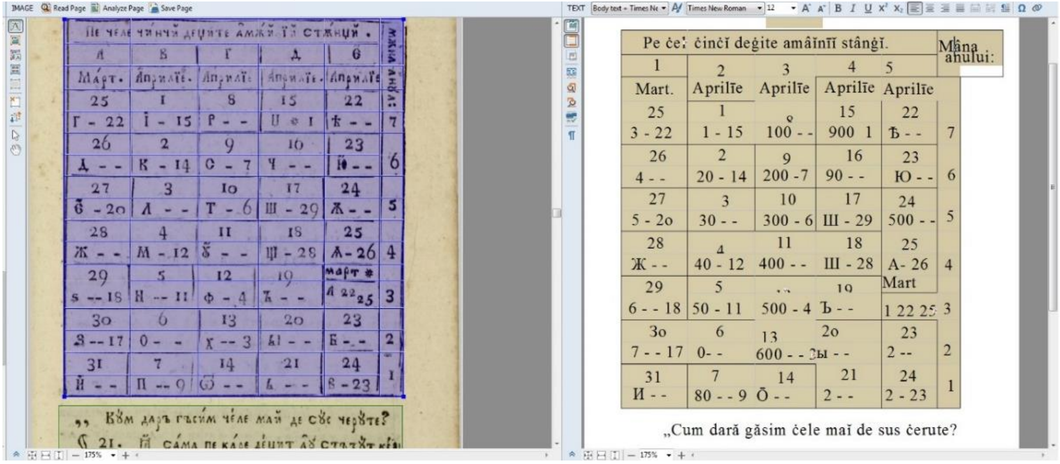
שֲ	שָ	שַ	ש
key W then key A + ALT right (for chatef patach) *chatef patach* = very short „a"	key W then SHIFT+A (for *qameṭ*) *qameṭ = long „a"	key W then key A (for *patach*) *patach = short „a"	key W (for letter SIN)
שוּ	**שֻ**	**שוֹ**	**שֹ**
key W then SHIFT+U (for ṣureq) *ṣureq= long„u"	key W then key U (for qibuṭ) *qibuṭ= short„u"	key W then SHIFT+O (for cholem vav) *cholem vav= long„o"	key W then key O (for cholem) *cholem= medium „o"
שֱ	**שֶ**	**שִ**	**שִי**
key W then key E + ALT right (for chatef segol) *chatef segol* = very short „e"	key W then key E (for *segol*) *segol= short „e"	key W then key I (for *chireq qaton*) * *chireq qaton* = short „i"	key W then SHIFT+I (for *chireq gadol*) * *chireq gadol* = long „i"

The table presents the same letter with different possibilities of vocalization, where all the vocals were introduced from the keyboard using simple or combined keys. We observed that the vocals are positioned next to the body of the letter, similarly with the Cyrillic letters presented in the previous table. We consider that such a font would be extremely useful for the Cyrillic alphabet as well, resulting in a considerable shortening of the time needed for transliteration.

4. Transliteration of the Cyrillic numbers

Another challenge in the stage of transliteration was caused by numbers. In Cyrillic, numbers are represented with letters. The main difficulties that occurred in such cases are:

a. In the MCVRO resources we found cases where we were not able to identify the standard correspondences presented in the table:



Пе челе чинчи дєџитє амѫнїи ѧ стѫнџи.					Мѫна анѫлѫи:
А	В	Г	Д	Є	
Мáрт.	Априлїє.	Априлїє.	Априлїє.	Априлїє	
25 Г - 22	1 Ї - 15	8 Р - -	15 Ц ✱ 1	22 Ѣ - -	7
26 Д - -	2 К - 14	9 С - 7	16 Ч - -	23 Ю - -	6
27 Є - 20	3 Л - -	10 Т - 6	17 Ш - 29	24 Ѫ - -	5
28 Ж - -	4 М - 12	11 Ѹ - -	18 Щ - 28	25 Ѧ - 26	4
29 Ѕ -- 18	5 Н -- 11	12 Ф - 4	19 Ъ - -	марТ ✱ А 22 25	3
30 З -- 17	6 О - -	13 Х -- 3	20 Ы - -	23 Б - -	2
31 Й - -	7 П -- 9	14 Ѡ - -	21 Ь - -	24 8 - 23	1

„ Кѫм дарѫ гѫсим челе маи де сѫс черѫте?

§ 21. Ꙟ сáма пе кáре деџит аѫ стѫтѫт кѣ

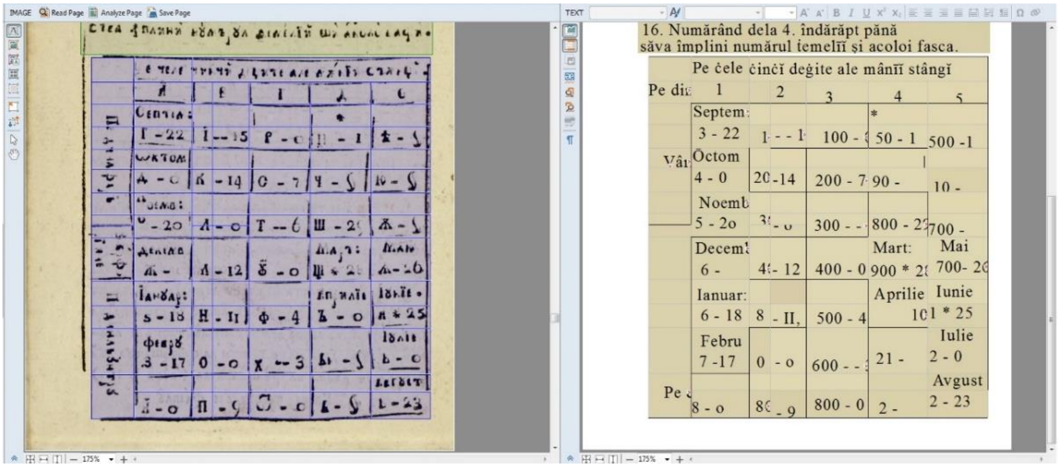
Pe ċe: ċinċĭ deġite amâinĭĭ stângĭ.					Mâna anului:
1	2	3	4	5	
Mart.	Aprilīe	Aprilīe	Aprilīe	Aprilīe	
25 3 - 22	1 1 - 15	ǫ 100 - -	15 900 1	22 Ѣ - -	7
26 4 - -	2 20 - 14	9 200 -7	16 90 - -	23 Ю - -	6
27 5 - 2o	3 30 - -	10 300 - 6	17 Ш - 29	24 500 - -	5
28 Ж - -	4 40 - 12	11 400 - -	18 Ш - 28	25 A- 26 Mart	4
29 6 - - 18	5 50 - 11	. . 500 - 4	19 Ъ - -	1 22 25	3
3o 7 - - 17	6 0- -	13 600 - -	2o ӡы - -	23 2 --	2
31 И - -	7 80 - - 9	14 Ō - -	21 2 - -	24 2 - 23	1

„Cum dară găsim ċele maĭ de sus ċerute?

Figure 3. Example of numbers and letters that cannot be found in the standard correspondences' numbers-letters for the Cyrillic alphabet (BRV 1487)[12]

[12] Antonie Monahul, *Mâna lui Damaschin* (Iași: Tipografia Mitropoliei, 1830), 20.

Sometimes, in the case of digitized texts, Arabic numbers are combined with Cyrillic letters to indicate bigger or decimal numbers. This aspect may lead researchers to erroneous transliterations:



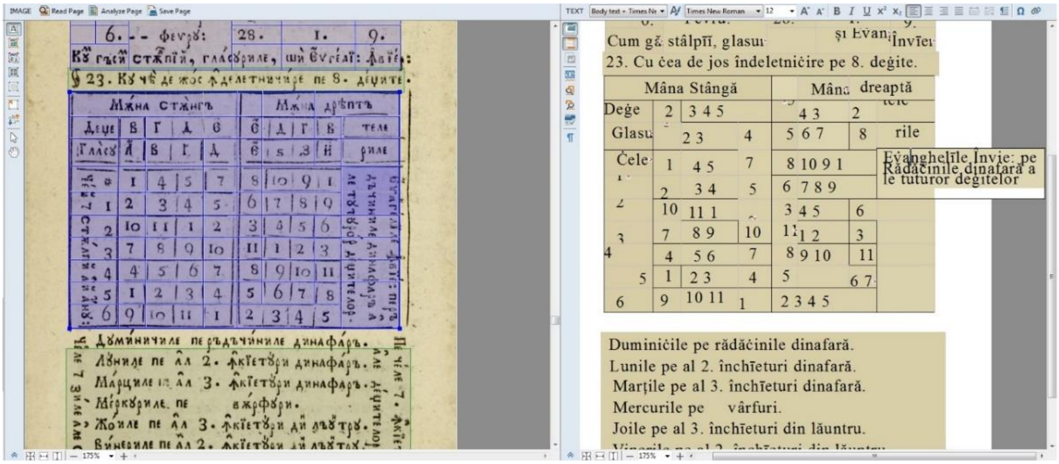
16. Numărând dela 4. îndărăpt pănă
săva împlini numărul ṫemelīī şi acoloi fasca.

	Pe ċele ċinċĭ deġite ale mânīī stângĭ				
Pe di:	1	2	3	4	5
	Septem:			*	
	3 - 22	1 - - 1	100 -	50 - 1	500 -1
Vâı	Ōctom				
	4 - 0	20-14	200 - 7	90 -	10 -
	Noemb				
	5 - 2o	3 - o	300 - -	800 - 2	700 -
	Decem			Mart:	Mai
	6 -	4- 12	400 - 0	900 * 2	700- 2
	Ianuar:			Aprilie	Iunie
	6 - 18	8 - II,	500 - 4	10	1 * 25
	Febru				Iulie
	7 -17	0 - o	600 - -	21 -	2 - 0
Pe					Avgust
	8 - o	8 - 9	800 - 0	2 -	2 - 23

Figure 4. Example of table with numbers and "letter-numbers" (BRV 1487)[13]

In each case, the team chose one of the following options: a. maintaining the writing from the original text; b. deduction from the context of the correct number and correct transliteration. The second option is extremely important because it underlines the essential contribution of the researcher specialized in digitization. Although the IT instruments are essential and may lead to an increased volume of the covered resources, enhancing ample research, they are not sufficient without the contribution of the human factor. Also, there is not enough to have an elementary human intervention, but a special knowledge is required together with a good orientation in the characteristics and thematic content of the covered resources. These aspects are important in establishing the equilibrium, the ideal relationship between the use of the information instruments and the contribution of the specialists. We also found tables where some cells were completed vertically. Considering that ABBYY FineReader does not allow the orientation of the text as word processors, we decided to complete the text with normal orientation to provide full content, even if the aspect is different:

[13] Antonie, *Mâna lui Damaschin*, 12.



Cum gă stâlpiī, glasu şi Evan Învie

23. Cu ċea de jos îndeletniċire pe 8. deġite.

Mâna Stângă				Mâna dreaptă		
Deġe	2	3 4 5		4 3	2	tele
Glasu		2 3	4	5 6 7	8	rile
Ċele 1	1	4 5	7	8 10 9 1		Evanghelīīle Învie: pe Rădăċinile dinafară a le tuturor deġitelor
	2	3 4	5	6 7 8 9		
2	10	11 1		3 4 5	6	
3	7	8 9	10	11 1 2	3	
4	4	5 6	7	8 9 10	11	
5	1	2 3	4	5	6 7	
6	9	10 11	1	2 3 4 5		

Duminiċile pe rădăċinile dinafară.
Lunile pe al 2. închīeturi dinafară.
Marţile pe al 3. închīeturi dinafară.
Mercurile pe vârfuri.
Joile pe al 3. închīeturi din lăuntru.

Figure 5. Example of table with numbers and letters, the underlined aspect being the horizontal transliteration as opposed to the original with vertical orientation (BRV 1487)[14]

The proofreading precisely respected the original content, even though some adjustments were made; these adjustments were related to the aspect required by the technical limitations of the program used:

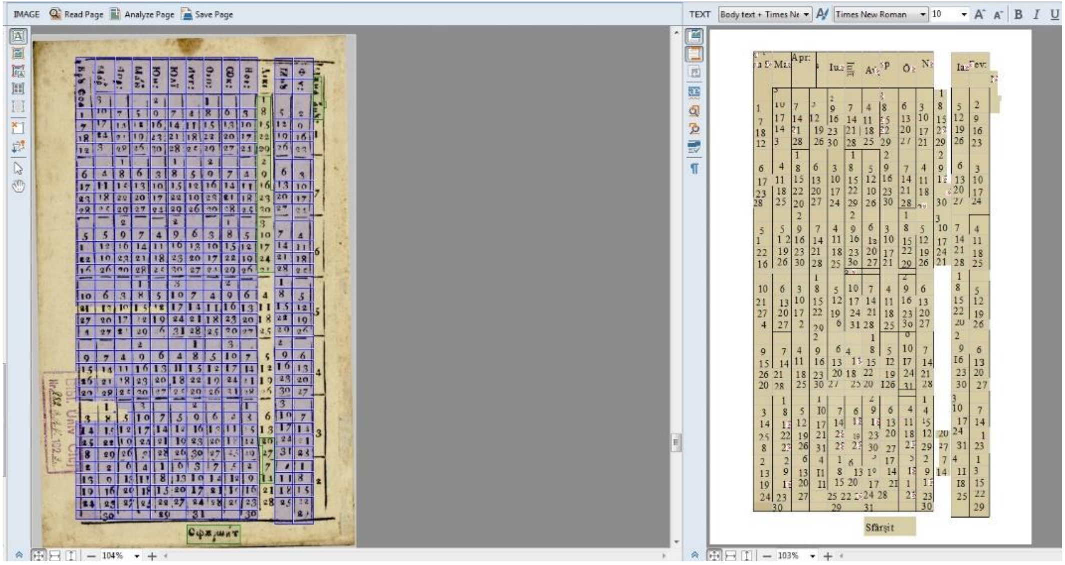
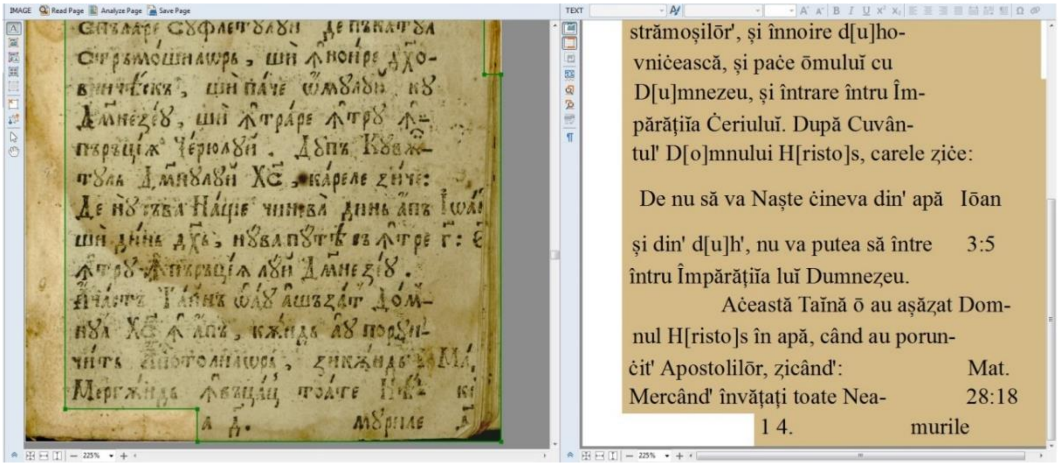


Figure 6. Example of table with corrected numbers, where in some cell's transliteration is given horizontally, but the content is given precisely (BRV 1487)[15]

[14] Antonie, *Mâna lui Damaschin*, 12.

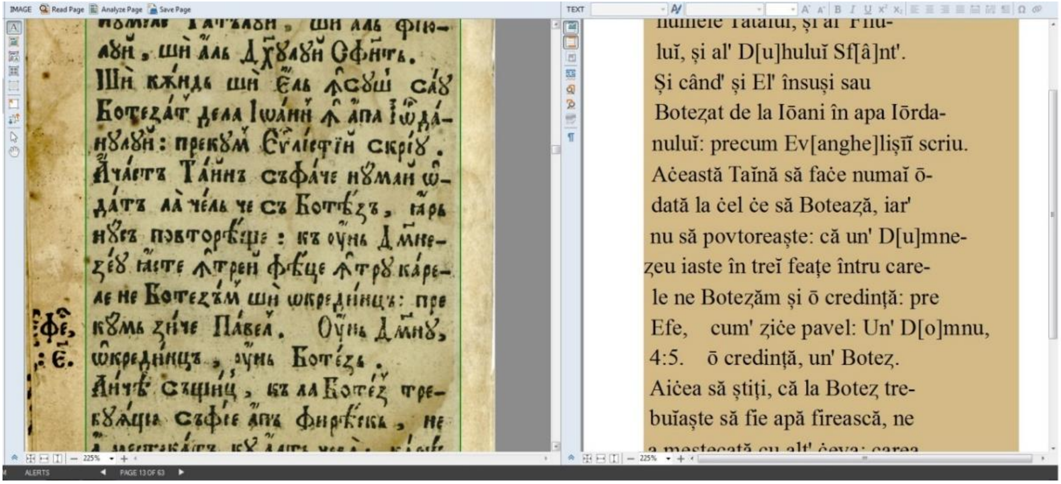
The contribution of the specialists was also essential in cases where scanned and OCR image did not offer the complete content of the resource; a study of the context or even additional research from other resources was necessary. For such steps, a specialized intervention was strictly necessary.

b. Identified Biblical texts in the Holy Scripture due to the incomplete or absent biblical reference in the processed MCVRO resource



strămoșilōr', și înnoire d[u]ho-
vnićească, și paće ōmuluĭ cu
D[u]mnezeu, și întrare întru Îm-
părățiĭa Ćeriuluĭ. După Cuvân-
tul' D[o]mnului H[risto]s, carele ẓiće:
De nu să va Naște ćineva din' apă   Iōan
și din' d[u]h', nu va putea să între   3:5
întru Împărățiĭa luĭ Dumneẓeu.
Aćeastă Taĭnă ō au așăẓat Dom-
nul H[risto]s în apă, când au porun-
ćit' Apostolilōr, ẓicând':   Mat.
Mercând' învățați toate Nea-   28:18
1 4.   murile

Figure 7. Biblical text searched for a complete transliteration (BRV 185)[16]

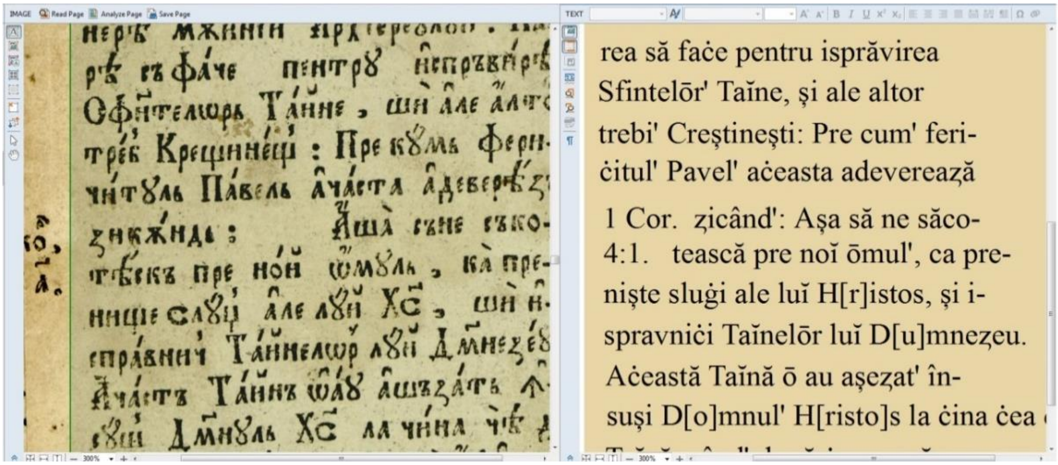


numele Tatăluĭ, și al' Fiiu-
luĭ, și al' D[u]huluĭ Sf[â]nt'.
Și când' și El' însuși sau
Boteẓat de la Iōani în apa Iōrda-
nuluĭ: precum Ev[anghe]lișĭĭ scriu.
Aćeastă Taĭnă să faće numaĭ ō-
dată la ćel će să Boteaẓă, iar'
nu să povtoreaște: că un' D[u]mne-
ẓeu iaste în treĭ feațe întru care-
le ne Boteẓăm și ō credință: pre
Efe,   cum' ẓiće pavel: Un' D[o]mnu,
4:5.   ō credință, un' Boteẓ.
Aićea să știți, că la Boteẓ tre-
buĭaște să fie apă firească, ne
a mestecată cu alt' ćeva: carea

Figure 8. Biblical text searched for a complete transliteration (BRV 185)[17]

[15] Antonie, *Mâna lui Damaschin*, 22.

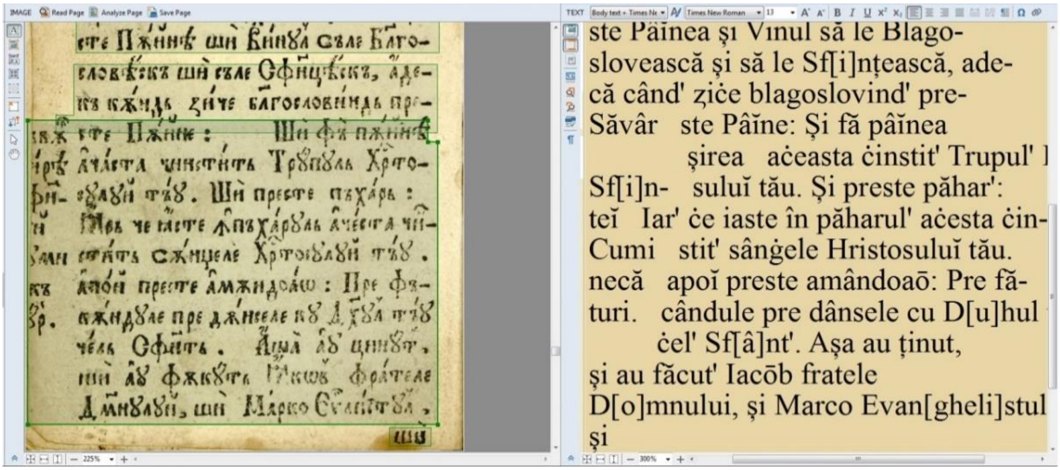
[16] Damaschin, *Învățătura despre șapte Taine*, 11.



rea să face pentru isprăvirea
Sfintelōr' Taĭne, şi ale altor
trebi' Creştineşti: Pre cum' feri-
ċitul' Pavel' aċeasta adevereaȥă
1 Cor. ȥicând': Aşa să ne săco-
4:1. tească pre noĭ ōmul', ca pre-
nişte sluġi ale luĭ H[r]istos, şi i-
spravniċi Taĭnelōr luĭ D[u]mneȥeu.
Aċeastă Taĭnă ō au aşeȥat' în-
suşi D[o]mnul' H[risto]s la ċina ċea

Figure 9. Biblical text searched for a complete transliteration (BRV 185)[18]

c. Missing texts or texts hard to decipher deducted from the context or from adjacent sources



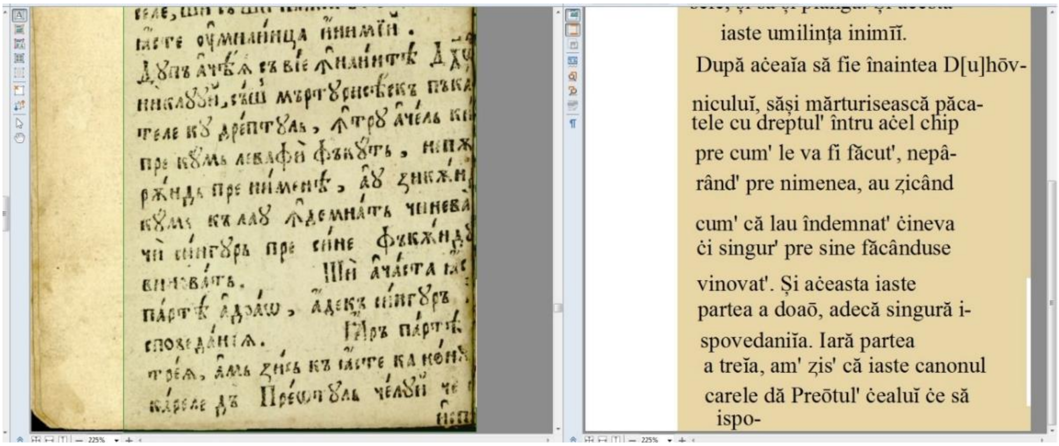
ste Pâinea şi Vinul să le Blago-
slovească şi să le Sf[i]nţească, ade-
că când' ȥiċe blagoslovind' pre-
Săvâr ste Pâine: Şi fă pâinea
şirea aċeasta ċinstit' Trupul'
Sf[i]n- suluĭ tău. Şi preste păhar':
teĭ Iar' ċe iaste în păharul' aċesta ċin-
Cumi stit' sânġele Hristosuluĭ tău.
necă apoĭ preste amândoaō: Pre fă-
turi. cândule pre dânsele cu D[u]hul
ċel' Sf[â]nt'. Aşa au ţinut,
şi au făcut' Iacōb fratele
D[o]mnului, şi Marco Evan[gheli]stul
şi

Figure 10. Missing text deducted from the context and/or with the help of adjacent sources (BRV 185)[19]

[17] Damaschin, *Învățătura despre șapte Taine*, 13.

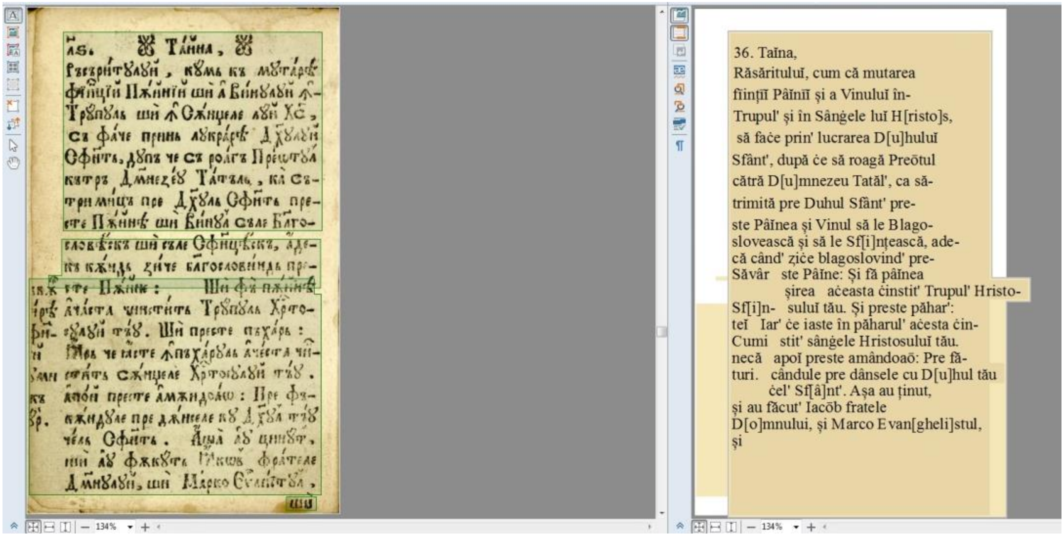
[18] Damaschin, *Învățătura despre șapte Taine*, 23.

[19] Damaschin, *Învățătura despre șapte Taine*, 41.



iaste umilinţa inimiĭ.
După aċeaĭa să fie înaintea D[u]hōv-
niculuĭ, săşi mărturisească păca-
tele cu dreptul' întru aċel chip
pre cum' le va fi făcut', nepâ-
rând' pre nimenea, au ȥicând
cum' că lau îndemnat' ċineva
ċi singur' pre sine făcânduse
vinovat'. Şi aċeasta iaste
partea a doaō, adecă singură i-
spovedaniĭa. Iară partea
a treĭa, am' ȥis' că iaste canonul
carele dă Preōtul' ċealuĭ ċe să
ispo-

Figure 11. Missing letters at the end of the rows, deducted from the context (BRV 185)[20]

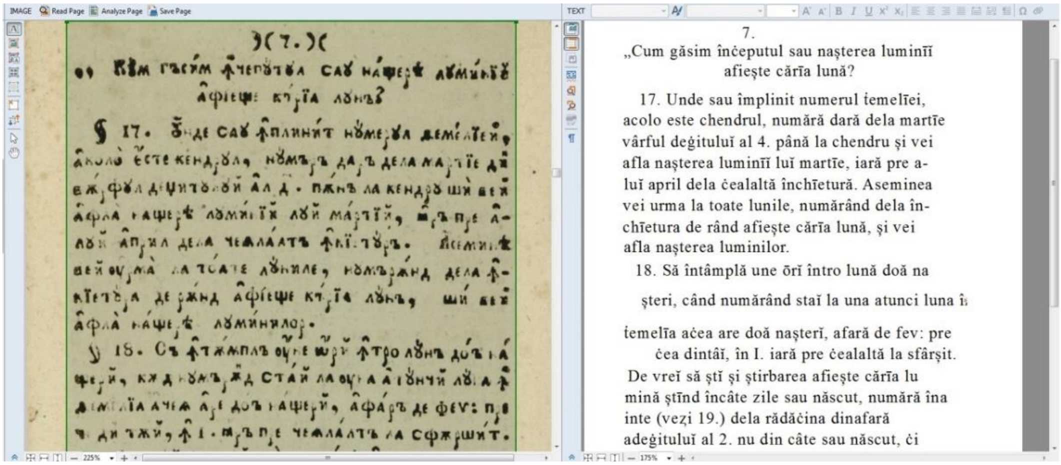


36. Taĭna,
Răsărituluĭ, cum că mutarea
fiinţiĭ Pâĭniĭ şi a Vinuluĭ în-
Trupul' şi în Sânġele luĭ H[risto]s,
să faċe prin' lucrarea D[u]huluĭ
Sfânt', după ċe să roagă Preōtul
către D[u]mnezeu Tatăl', ca să-
trimită pre Duhul Sfânt' pre-
ste Pâĭnea şi Vinul să le Blago-
slovească şi să le Sf[i]nţească, ade-
că când' ȥiċe blagoslovind' pre-
Săvâr ste Pâĭne: Şi fă pâĭnea
şirea aċeasta ċinstit' Trupul' Hristo-
Sf[i]n- suluĭ tău. Şi preste păhar':
teĭ Iar' ċe iaste în păharul' aċesta ċin-
Cumi stit' sânġele Hristosuluĭ tău.
necă apoĭ preste amândoaō: Pre fă-
turi. cândule pre dânsele cu D[u]hul tău
ċel' Sf[â]nt'. Aşa au ţinut,
şi au făcut' Iacōb fratele
D[o]mnului, şi Marco Evan[gheli]stul,
şi

Figure 12. The letters in brackets are missing from the original text and can be deducted only by specialists (BRV 185)[21]

[20] Damaschin, *Învățătura despre șapte Taine*, 33.

[21] Damaschin, *Învățătura despre șapte Taine*, 41.



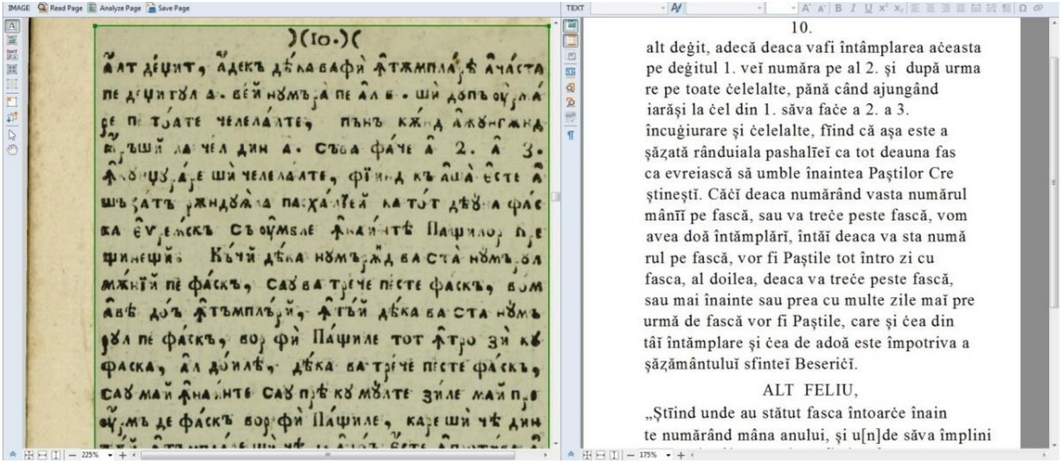
7.
„Cum găsim inċeputul sau naşterea luminiĭ
afieşte cărĭa lună?

17. Unde sau împlinit numerul ṫemelĭei,
acolo este chendrul, numără dară dela martĭe
vârful deġituluĭ al 4. până la chendru şi vei
afla naşterea luminiĭ luĭ martĭe, iară pre a-
luĭ april dela ċealaltă închĭetură. Aseminea
vei urma la toate lunile, numărând dela în-
chĭetura de rând afieşte cărĭa lună, şi vei
afla naşterea luminilor.

18. Să întâmplă une ōrĭ întro lună doă na
şteri, când numărând staĭ la una atunci luna î

ṫemelĭa aċea are doă naşterĭ, afară de fev: pre
ċea dintâĭ, în I. iară pre ċealaltă la sfârşit.
De vreĭ să ştĭ şi ştirbarea afieşte cărĭa lu
mină ştĭnd încâte zile sau născut, numără îna
inte (veẓi 19.) dela rădăċina dinafară
adeġituluĭ al 2. nu din câte sau născut, ċi

Figure 13. A text difficult to decipher, transliterated (BRV 1487)[22]



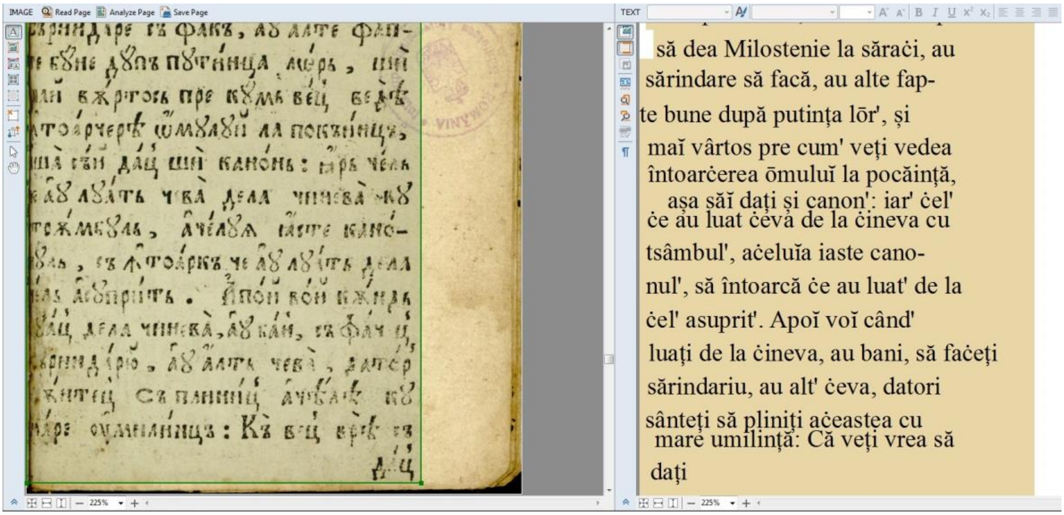
10.
alt deġit, adecă deaca vafi întâmplarea aċeasta
pe deġitul 1. veĭ număra pe al 2. şi după urma
re pe toate ċelelalte, pănă când ajungând
iarăşi la ċel din 1. săva faċe a 2. a 3.
încuġiurare şi ċelelalte, fĭind că aşa este a
şăẓată rânduiala pashalĭeĭ ca tot deauna fas
ca evreiască să umble înaintea Paştilor Cre
ştineştĭ. Căċĭ deaca numărând vasta numărul
mânĭĭ pe fască, sau va treċe peste fască, vom
avea doă întămplărĭ, întăĭ deaca va sta numă
rul pe fască, vor fi Paştile tot întro zi cu
fasca, al doilea, deaca va treċe peste fască,
sau mai înainte sau prea cu multe zile maĭ pre
urmă de fască vor fi Paştile, care şi ċea din
tâĭ întămplare şi ċea de adoă este împotriva a
şăẓământuluĭ sfinteĭ Biseriċĭ.

ALT FELIU,

„Ştĭind unde au stătut fasca întoarċe înain
te numărând mâna anului, şi u[n]de săva împlini

Figure 14. A text difficult to decipher, transliterated (BRV 1487)[23]

[22] Antonie, *Mâna lui Damaschin*, 13.
[23] Antonie, *Mâna lui Damaschin*, 16.



să dea Milostenie la săraċi, au
sărindare să facă, au alte fap-
te bune după putința lōr', și
maĭ vârtos pre cum' veți vedea
întoarċerea ōmuluĭ la pocăință,
așa săĭ dați și canon': iar' ċel'
ċe au luat ċevá de la ċineva cu
tsâmbul', aċeluĭa iaste cano-
nul', să întoarcă ċe au luat' de la
ċel' asuprit'. Apoĭ voĭ când'
luați de la ċineva, au bani, să faċeți
sărindariu, au alt' ċeva, datori
sânteți să pliniți aċeastea cu
mare umilință: Că veți vrea să
dați

Figure 15. Missing letters or almost illegible (BRV 185)[24]

Conclusions

A digitalization is needed to make old text searchable and accessible. In comparison to hard copy text, working with digitalized text offers a lot of great benefits in terms of editing, annotation and especially searching.

An OCR historical text can have various uses, including online access, re-print, and information retrieval at a more granular level.

To get the best quality in OCR processes, our recommendations include: clean and optimize the documents, images and text, remove dirt and flatten the paper, scan at best possible quality (for pre-process step), use high performance OCR application and clearly define regions for OCR (for process step) and use external resources (as dictionaries) to ensure more accurate recognition of documents and to simplify further verification of recognition results (for past process step).

A digitalized collection can be used in various ways. It can be used to study historical persons and their relationships to have an insight into the lives of people and the way society functioned in early times.[25]

[24] Antonie, *Mâna lui Damaschin*, 34.

[25] S. Torao-Pingi and R. Nayak, "Understanding People Relationship: Analysis of Digitised Historical Newspaper Articles, Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence," in *28th Australasian Joint Conference on Artificial Intelligence* (AI), J. Renz and B. Pfahringer (eds.) (Switzerland: Springer, 2015), 572-588.

Acknowledgment: This work was supported by UEFISCDI project ID PN-II-PT-PCCA-2013-4-2062, contract no. 311/2014, *The Identification and Correlation between Concepts in Old Romanian Books and Manuscripts, using a Computer Application to stimulate Social Dynamics through the Appraisal of the Cultural Production*.