Improving accessibility of digitization outputs: EODOPEN project research findings

Alenka Kavčič Čolić and Andreja Hari
Research Department, National and University Library of Slovenia, Ljubljana, Slovenia

Received 18 September 2023
Revised 29 November 2023
Accepted 22 January 2024

Abstract

Purpose – The current predominant delivery format resulting from digitization is PDF, which is not appropriate for the blind, partially sighted and people who read on mobile devices. To meet the needs of both communities, as well as broader ones, alternative file formats are required. With the findings of the eBooks-On-Demand-Network Opening Publications for European Netizens project research, this study aims to improve access to digitized content for these communities.

Design/methodology/approach – In 2022, the authors conducted research on the digitization experiences of 13 EODOPEN partners at their organizations. The authors distributed the same sample of scans in English with different characteristics, and in accordance with Web content accessibility guidelines, the authors created 24 criteria to analyze their digitization workflows, output formats and optical character recognition (OCR) quality.

Findings – In this contribution, the authors present the results of a trial implementation among EODOPEN partners regarding their digitization workflows, used delivery file formats and the resulting quality of OCR results, depending on the type of digitization output file format. It was shown that partners using the OCR tool ABBYY FineReader Professional and producing scanning outputs in tagged PDF and PDF/UA formats achieved better results according to set criteria.

Research limitations/implications – The trial implementations were limited to 13 project partners' organizations only.

Originality/value – This research paper can be a valuable contribution to the field of massive digitization practices, particularly in terms of improving the accessibility of the output delivery file formats.

Keywords Digitization, Accessibility, Blind and partially sighted, Mobile technology users

Paper type Research paper

1. Introduction

In recent years, the way we access information has undergone significant transformation. Digital access to information is no longer an alternative but has evolved into a standard requirement for most library users. Consequently, digitization has become a routine practice in numerous libraries. Depositary libraries, including national, regional and state libraries,

This work has received funding with support from the European Commission. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use that may be made of the information contained therein.

Digital Library Perspectives
Vol. 40 No. 2, 2024
pp. 187-211
Emerald Publishing Limited
2059-5816
DOI 10.1108/DLP-09-2023-0080

have incorporated the digitization of their collections into their strategies to facilitate access to and preserve their national cultural heritage.

However, the current output file formats of digitization fall short of meeting the needs of the blind, partially sighted communities and users of mobile devices. The majority of digitized collections are available in PDF or image file formats, posing accessibility challenges for these communities to access.

It is noteworthy to mention the European Accessibility Act, [1] which is founded on an EU Directive [2] aiming to enhance the functioning of the internal market for accessible products and services by eliminating barriers arising from divergent rules in Member States. According to this Act, individuals with disabilities and the elderly are entitled to benefit from more accessible products and services in the market, including e-books, e-readers or other devices facilitating digital reading [3].

Compliance with the same Act is assessed based on the Harmonized European Standard: Accessibility requirements for ICT products and services EN 301 549 [4], which, at the very least for Web pages and documents, is grounded in the Web Content Accessibility Guidelines 2.1 (WCAG 2.1) [5], [6]. The forthcoming version, WCAG 3, currently in draft, will also be applicable to Web content, apps, tools, publishing and emerging technologies on the Web [7]. These and similar guidelines, already in existence for the accessibility of digital content, could also be applied to digitized content, particularly in the final stages of workflow processes.

In the European project EODOPEN [8], we analyzed this problem and sought to identify optimal delivery formats that could provide better access to digitized content. To address this, we conducted several surveys on the use of mobile technologies and the specific needs of users, as well as the technical requirements. These surveys formed the basis for the development of the "Guidelines and Recommendations for the Provision of Alternative and Special Formats" (2022).

In adherence to these guidelines and in collaboration with EODOPEN partners, we carried our trial implementations for mobile devices aimed at users with print disabilities. The objective of these trials was to analyze our partners' digitization workflows, the file formats used for delivery and, consequently, the quality of optical character recognition [9] (OCR) results. This analysis took into account the file format type and adherence to accessibility criteria outlined in the WCAG [10]. The insights gained from our analysis can assist libraries in enhancing the quality of their digitized content. This, in turn, enables to provide improved access to all their users.

2. Related research

Increased digitization in libraries and other cultural organizations has had a significant impact on the development of scanning technologies and image-processing tools. This progress has also been reflected in the advancement of OCR software and automated scanning processes.

Several noteworthy European projects have concentrated on enhancing OCR accuracy, with one such initiative being the Computational History and Transformation of Public Discourse in Finland, 1640–1910 (COMHIS) [11] co-funded by the Academy of Finland. In this project, the National Library of Finland reprocessed a representative portion of its collection of digitized journals and newspapers collection from the 1990s using new OCR tools. The result was a 25% reduction in errors compared to the older OCR's text, achieving an accuracy of over 80% (Kettunen *et al.*, 2020).

Another European project is Improving Access to Text [12] (IMPACT) (2008–2011), which included ABBYY production, the developer of the ABBYY FineReader OCR tool, as a

partner. IMPACT's primary goal was to automate OCR production and enhance digital access to historical printed texts through the development and use of innovative OCR software and linguistic technologies. Their finding was validated through the comparison of automatically OCR-ed old texts with their manually corrected versions, demonstrating an automatic achievement of over 95% OCR accuracy (ground truth).

The tranScriptorium project (2013–2015) [13] aimed to develop efficient and cost-effective solutions for indexing, searching and transcribing historical handwritten document images using modern handwritten text recognition (HTR) technology.

Building on the findings of tranScriptorium, the recognition and enrichment of archival documents (READ) project [14] (2016–2020) sought to establish a service platform for automated recognition, transcription and searching of historical documents [15]. The READ project leveraged machine learning technologies and set new standards in HTR, keyword spotting, layout analysis, automatic writer identification and related fields.

Also noteworthy is the Czech national project PERO [16] by the Brno University of Technology and Moravian Library. Its objective was to develop technology and tools that:

> [. . .] would improve the accessibility of digitized historic documents. These tools, based on state-of-the-art methods from computer vision, machine learning, and language modeling, [. . .] enable existing digital archives and libraries to provide full-text search and content extraction for low-quality historic printed and all handwritten documents, which cannot be automatically processed by the currently available tools.

While the results of HTR in tranScriptorium, READ and PERO are very promising, the text processing remains time-consuming. However, due to the higher sensitivity of this technology, it could yield better results when applied to print texts.

There is limited literature focusing on delivery formats as the output of digitization. Avyodri *et al.* (2022) provide an insightful literature review on OCR for text recognition and its postprocessing methods. Their findings reveal that most OCR literature concentrates on image preprocessing (how the quality of the image and scanning can influence better OCR results), text segmentation (e.g. detection of text, nontext, title, edge etc.), feature extraction (e.g. zoning or projection profile), text recognition and postprocessing (OCR correction through lexical-based postprocessing, using a static dictionary or spelling check and lexical-based methods). They conclude that while 100% OCR accuracy is not guaranteed, some research indicates the possibility of achieving over 90% accuracy. The quality also depends on the language used and the amount of data used for training the OCR tool.

In scanning projects, the majority of OCR errors arise from incorrect or insufficient layout analysis of scanned texts. This involves tasks such as page division and segmentation, integrating images into text blocks and lines, marking semantic tags (e.g. headings) and maintaining the correct reading order. Some of these criteria are also outlined in the WCAG as well. Research in this field focuses on identifying potential layout errors and using various automatic methods to reduce them (Shafait, 2008; Erkilinc *et al.*, 2011; Forczmański *et al.*, 2020; Xu *et al.*, 2020; Zhong *et al.*, 2019).

Jääskeläinen *et al.* (2023) demonstrated that pre-OCR image manipulation does not significantly impact OCR results. Reisswig *et al.* (2020) developed an end-to-end trainable OCR system for printed documents based on character instance segmentation, aiming to enhance OCR text accuracy.

The International Conference on Document Analysis and Recognition (ICDAR) also addresses topics related to optimal OCR quality. Annually, it organizes a competition on software development for the automatic detection and quality improvement of OCR results. Reports on each competition have been published in their conference proceedings. In 2023, a competition focusing on accurately segmented page layouts across various document styles

and domains was held (Auer *et al.*, 2023). In 2021, the competition centered on mathematical formula detection (Zhong *et al.*, 2021). In 2019, the competition concentrated on methods for table detection and recognition (Gao *et al.*, 2019). In the same year, ICDAR organized a competition on post-OCR text correction (Rigaud *et al.*, 2019). These competitions have spurred the development of tools and algorithms for the automatic detection of specific elements in OCR-ed texts. Additionally, participants had access to an extensive sample of OCR-ed documents and resulting lexicons, with competitors deciding on the methods of analysis.

Trbušić (2022) delved into the optimization of OCR for long-term preservation in archives and its integration into the scanning and OCR tool processes. The author primarily investigates various aspects, including the typography of printed texts, image file formats, compression standards, text file formats and text encoding standards. The analytical tools for OCR evaluation, developed at The Information Science Research Institute at the University of Nevada, Las Vegas, are used to analyze digitization output delivery formats. The study encompasses the implementation of four OCR tools: ABBYY FineReader 15, Google Cloud Vision API, Tesseract 4 and Asprise OCR 15 (refer to Table 1). ABBYY FineReader supports the majority of the listed file formats, whereas Tesseract 4 supports PDF, ALTO, hOCR, TSV and BOX. Asprise OCR 15, on the other hand, supports only PDF, RTF, TXT and XML, with none of these formats being supported by Google Cloud Vision API. The research findings indicate that the quality of OCR across different tools is not contingent on the scanning resolution.

The literature review underscores a significant ongoing concern and emphasis on developments aimed at improving the quality of texts generated by OCR tools. High-quality OCR has the potential not only to enhance accessibility for mobile devices and users with visual impairments but also to contribute to improved data mining and text analyses (Jääskeläinen *et al.*, 2023; Inbasekaran *et al.*, 2021).

In practical terms, digitized contents are typically processed with OCR tools, which, although unable to achieve 100% accuracy, often make materials accessible to users in PDF file formats. Consequently, when absolute accuracy is essential, additional manual corrections become necessary. Nevertheless, emerging technologies based on artificial intelligence, such as ChatGPT, are evolving, and there is an expectation that they will play a role in automatically improving OCR-ed texts and reducing the error percentage. It is

Table 1.
Selected OCR tools and file formats supported

OCR tool	File format supported	File format partly supported	File format not supported
ABBYY Fine Reader 15	PDF, DOC, XLS, PPT, HTML, RTF, TXT, CSV, ODT, FB2, EPUB, DjVu	ALTO	XML, hOCR, TSV, BOX
Google Cloud Vision API	Not appliable	Not appliable	Not appliable
Tesseract 4	PDF, TXT, ALTO, hOCR, TSV, BOX	None	DOC, XLS, PPT, HTML, RTF, CSV, ODT, FB2, EPUB, DjVu, XML
Asprise OCR 15	PDF, RTF, TXT, XML	None	DOC, XLS, PPT, HTML, CSV, ODT, FB2, EPUB, DjVu, ALTO, hOCR, TSV, BOX

Source: Trbušić (2022), Table 2, p. 48

anticipated that, with machine learning technologies drawing from larger collections/samples of texts, these systems will become more accessible to librarians.

Although exploring the procedures of digitization and their impact on enhancing the accessibility of digitized materials for the blind and partially sighted, few extensive studies have been identified. More attention is directed toward researching the accessibility of born-digital content and its creation based on standards and recommendations. Recommendations for born-digital content can also be applied to digitized works and their conversion process for accessibility purposes:

> The conversion process involves adapting material formats into a form that can be used by users with reading impairments. Conversion is carried out according to technical guidelines for creating accessible formats (Kodrič-Dačić *et al.*, 2014, p. 31).

Gunn (2016, p. 4), in the Accessible eBook Guidelines for self-publishing authors, emphasizes that:

> [. . .] the essence of eBook accessibility relates to supporting flexible ways for people to engage in the eBook content based on their personal needs [. . .] and one of the strengths of eBook technologies is to allow users to quickly and easily customise the way the content is presented to suit their requirements.

Individuals with special needs, specifically those who are blind and partially sighted, use the same devices as individuals without disabilities but additionally use assistive technology to access the digital content. This may include a braille display, speech synthesis or software that facilitates content enlargement on the screen. Such technology presents the content to users in a linear manner, from top to bottom, making the order of elements a crucial factor for them to comprehend the content. Moreover, it is vital to consider the variability in a person's vision, affected by factors such as the remaining sight, variations from day to day, light conditions, tiredness and stress. Therefore, it is essential to empower users to adapt the visual presentation of text to suit their individual needs. Common challenges faced by individuals, especially those who are partially sighted, include difficulties in focusing on text, reduced contrast sensitivity, a narrowed field of vision, sensitivity to movement and visual fatigue. For these individuals, making adjustments such as changing font size, font type, color themes and modifying margins and spacing can be highly beneficial. Equally crucial is the provision of options to access the full text (with preferably checked OCR) and to facilitate the use of the assistive technologies.

Femc (2018, pp. 42–45), in her study on e-book users, identifies the advantages and drawbacks of e-books. Among the benefits are text searching, dictionary use, font enlargement, internet access, screen brightness, portability, quick access to content and affordability. She also notes the following drawbacks: dissatisfaction related to the physical and functional capabilities of the device, screen brightness, lack of a paper-like feel, device battery performance, frustration with limited or disabled e-book transfers to personal devices and various e-book formats that are not universally compatible with all devices.

We believe that the blind and visually impaired face similar challenges, with their difficulties additionally influenced by their residual vision. In the user study by Završek *et al.* (2013, pp. 155–156), respondents with special needs mentioned various aids, including a computer, electronic magnifier, telescopic glasses, MP3 player, scanner, prescription glasses, voice recorder, Braille display, screen reader and a sound amplifier. Razpet (2017, p. 74) states that:

> [. . .] e-book technology offers a significant advantage over printed books, allowing users to adjust font size and simultaneously adapt text to the screen. This is especially convenient for older users or those with various forms of visual impairment.

A similar study on e-book reading among users with dyslexia (Rello and Baeza-Yates, 2017, p. 30) notes that recommendations for the blind and partially sighted are very similar to those for individuals with dyslexia. They also emphasize the importance of text size, character spacing, reading black text on a white background and vice versa.

In the guidelines, provided by The DAISY consortium on how to create accessible Word documents, it is stated that "there are globally accepted standards and best practices for creating accessible digital content" and "that some of the most adopted standards are: WCAG, Section 508, EPUB Accessibility and PDF/UA." (Creating accessible Word documents, 2023). According to these guidelines, all the standards have the same goals for:

- "Creating a structured and navigable document – It should be possible for all readers to easily identify and move to any position in the document [. . .]
- Provision of text descriptions for graphical content such as pictures, flow charts and maps so that visually impaired readers do not miss out on important aspects of understanding the document [. . .]
- Providing an adaptable format that is marked-up semantically – It should be possible for readers to adapt the visual presentation of the document to suit their reading needs [. . .]."

From these goals, it is evident that design and structure are the most crucial elements in creating accessible digital and digitized materials. In a Japanese study, Ishihara *et al.* (2012, p. 93) emphasized that the digitization and conversion process into an accessible format is a time-consuming procedure, outlining different steps in adapting a digitized publication into digitally accessible content [17].

The WCAG mentioned earlier consists of a set of success criteria, primarily for the Web, with a history dating back to 1999 when the first version was available. The latest version, WCAG 2.2 (Campbell *et al.*, 2023), states that these guidelines cover:

> [. . .] a wide range of recommendations for making Web content more accessible. Following these guidelines will make content more accessible to a wider range of people with disabilities, including accommodations for blindness and low vision, deafness and hearing loss, limited movement, speech disabilities, photosensitivity, and combinations of these, and some accommodation for learning disabilities and cognitive limitations; but will not address every user need for people with these disabilities.

Given our focus on mobile devices, it is particularly important to adhere to these guidelines when assessing the accessibility of delivery formats, as WCAG 2.2 explicitly mentions that they "address accessibility of web content on desktops, laptops, tablets, and mobile devices." The guidelines are built on four principles: perceivable, operable, understandable and robust, with 86 success criteria categorized under conformance levels A, AA and AAA. According to the EU Directive on the accessibility requirements for products and services, digital content must conform at least to criteria labeled with levels A and AA. For digitized content, not all success criteria are relevant, but some are crucial for mobile devices and blind and partially sighted users, such as 1.1.1 nontext content (level A), 1.3.2 meaningful sequence (level A), 1.4.1 use of color (level A), 1.4.10 reflow (level AA), 2.1.1 keyboard (level A), 2.4.6 headings and labels (level AA) and others.

User studies on information access preferences by blind and partially sighted individuals have shown a preference for the DAISY file format (Calvert *et al.*, 2019). For accessible full text, WCAG should be applied. File formats like PDF/universal accessibility (UA), EPUB 3 and HTML meet most of the WCAG criteria

(Mulliken and Falloon, 2017). Tagged PDF Best Practice Guide: Syntax (2019) and the Section 508 Guide: Tagging PDF's in Adobe Acrobat Pro (2018), both guidelines focus on how to make PDF documents more compliant with WCAG. Authors can thus produce more accessible tagged PDFs and PDF/UA. On the other hand, Ganner *et al.* (2023) provided a useful framework for publishing e-books for individuals with print disabilities, incorporating some of the WCAG criteria analyzed in our work. Their guidelines align with the outputs of our research.

3. Methodological approach

Our research focused on the digital transformation of modern printed library materials. The goal was to identify optimal delivery file formats resulting from digitization, recognizing that various formats offer users different experiences. Additionally, we aimed to explore whether there was any dependence between the quality of digitization results and the scanning and recognition workflows in EODOPEN partner libraries.

Despite numerous developments in OCR quality for scanned text images, many are still in the experimental stage and are not widely deployed. Libraries predominantly operate with officially acquired licenses. Avyodri *et al.* (2022) outlined digitization workflows comprising different phases: image preprocessing, text segmentation, feature extraction, text recognition and postprocessing. Notably, Jääskeläinen *et al.* (2023) found that pre-OCR image manipulation does not influence OCR results significantly. Therefore, in our research, we opted to skip this phase and concentrate on subsequent processing phases.

Through the literature review and chosen sample, we identified 24 criteria, primarily based on WCAG, suitable for analyzing digitization results concerning text segmentation, feature extraction and text recognition. These criteria primarily focus on accessibility for the blind and partially sighted. The selected criteria include alt-text picture, alt-text picture (chem. formula), caption, footnotes, heading 1, heading 2, heading 3, initial, language segments, math (simple), math (advanced), OCR errors [18], page rotation, pagination, pagination–double, picture, picture (chem. formula), primary language, special character, stamp removal, table, table header, table rows and text order [19].

In our qualitative analysis, we prioritized the quality of achieving OCR accuracy concerning scanning criteria elements and the impact of the delivery file format. Our goal was to ensure that the resulting text could be understood by the blind and partially sighted, aided by tools such as speech synthesis or other reading software. Consequently, the percentage analysis of achievement was not a focal point of our research.

We opted to select scanning samples that contained one or more features illustrating how OCR tools handle various elements. The testing sample comprised 16 scans in TIFF format (refer to Figure 1) [20] with varying complexity regarding textual and nontextual elements (e.g. pagination, footnotes, table structures, pictures, language segments, alternative text, etc.). To maintain consistency, all partners decided to test the same sample containing English text. This choice was driven by the fact that English was not the native language of any EODOPEN partner. Using the same scans with English text facilitated a comparative analysis of the results. Additionally, some OCR tools are better adapted to major languages (such as German), and our intention was to avoid discrimination against minor languages like Slovenian, Estonian or Slovak.

The results were evaluated based on the level of achievement of the aforementioned 24 criteria for optimal document accessibility and other best practice guidelines. Three levels were used to assess the set criteria:

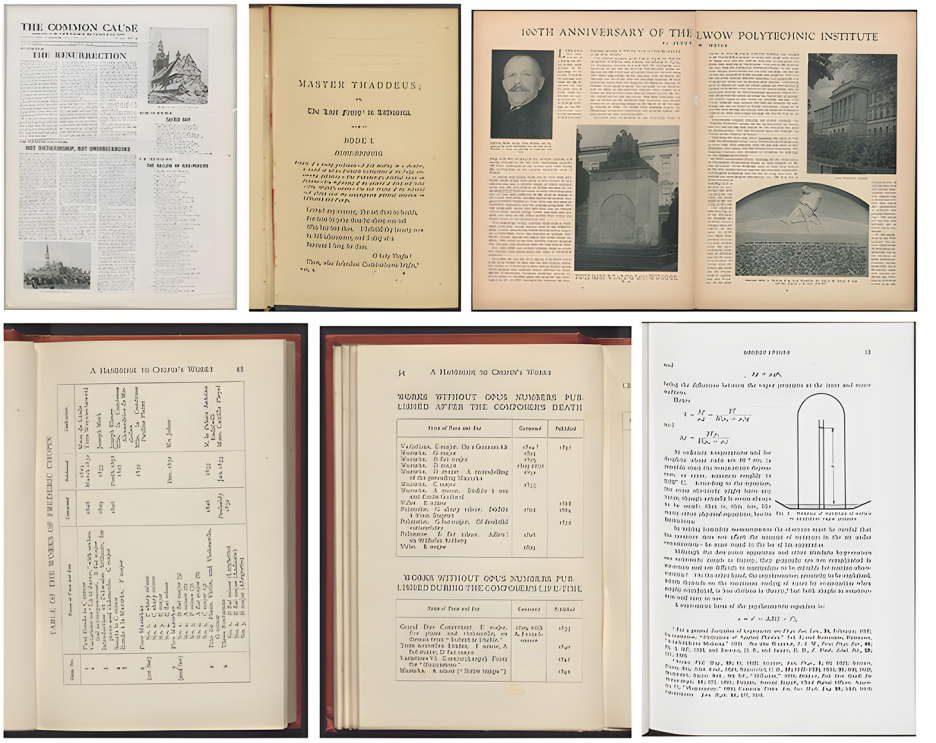


Figure 1.
Six examples of scans from the testing sample showcasing both simple and complex structures, including tables, varying orientations of content on the scan, titles spread over two pages and mathematical equations

Note: Refer to scans no. 6, 9, 7, 10, 11 and 12 in appendix 1
Source: All images (scans from books) included are in public domain

(1) *Criterion was fully achieved (A)*: This label was used when both the technical and content aspects of the criterion were met. For example, table rows were both technically and content-wise correct, with each row containing the accurate number of rows and the correct content.

(2) *Criterion was partly achieved (B)*: This label was used when either the technical or content part of the criterion was met, but not both or when there was a very minor mistake in the criterion. For example, the alt-text is technically correct, but the content is either the text of the caption or other surrounding text. Another example would be a minor mistake in the text order.

(3) *Criterion was not achieved (blank cell)*: This label was used when neither the technical nor the content part of the criterion was met. For instance, pagination was present but was not the first element on the page.

The testing phase among all partner institutions took place between February 2 and July 13, 2022. EODOPEN partners participating in the survey represented 13 national, academic and special libraries across 10 countries [21]. Each partner received a testing sample (refer to Appendix 1) and an empty testing report questionnaire (refer to Appendix 3) in which they documented the work carried out with the testing sample. The testing report questionnaire consisted of 14 questions, enabling project partners to

record the work processes, software tools and solutions used during the testing of the sample.

Audiobook production was not analyzed, as none of the EODOPEN partners were involved in the production of audiobooks.

4. Research findings

We received testing results from 23 submissions across 13 partner institutions. These include automatically generated results (17), as well as results that contain additional manual corrections (6).

Based on the information gathered from the testing report questionnaires, the delivery formats were diverse: PDF (15) – inclusive of both tagged or untagged PDFs, with one instance conforming to PDF/UA standard. Other formats included XML with TXT (3), ePUB (2), RTF (2) and DOCX (1). The results are shown in Table 2, where the best results achieved for each criteria are in italic.

ABBYY FineReader, in various versions, was the most widely used software for testing the samples. Additional software used included ScanGate by Treventus Mechatronics, Adobe Acrobat Pro, IRIS OCR, LIMB processing, Microsoft Office Word, Scan Tailor Advanced, Tesseract, Photoshop and Project PERO OCR.

In results involving additional manual corrections (refer to Table 3), Microsoft Office Word was predominantly used to rectify OCR errors and incorporate structural elements. Other software, such as Adobe InDesign and Adobe Acrobat Pro, played a key role in editing structural and navigational elements. In a unique instance, the WordToEpub tool was used to convert a manually edited Word file into an EPUB format.

The reported workflows by project partners exhibited similarities, as all adhered to digitization good practices [22]. From these reports, it was observed that while some partners engaged in preprocessing tasks [23], the majority performed automatic layout segmentation, with only one OCR engine using machine learning (PERO project). The quality of the results did not show a significant correlation with the specific digitization workflow processes used.

Here are observations and findings derived from the evaluation of the 24 criteria mentioned above.

The test results indicate that the best automatically generated outputs were achieved using PDF/UA or tagged PDF as the delivery format. These file formats outperformed others in all criteria. However, they lacked the flexibility to visually adapt content to specific needs, such as reflow on smaller screens, changing text size and font, color themes, margins and spacings.

The alt-text criterion presented challenges, as it currently requires human input to provide blind users with additional value from the images in the publications. In some cases, alt-text was present in the testing outputs, but the content was incorrect, including text from the caption or other surrounding text. This issue originated from layout analysis.

Different scans posed problems for some partners, as the system did not accept scans of varying sizes or different systems for monographs and newspapers used by the partners. Some partners resolved this issue by importing each scan separately.

The complexity of the structure of elements on the scans also influenced the results. Scans with a simple one-column structure had fewer errors compared to those with a complex structure involving multiple columns and other elements. In complex structured scans, the order of text was often problematic, occurring during automatic

Table 2.
Automatically generated output results based on criteria

File formats	Ref. no.	PDF 1.4	PDF 1.4	PDF 1.4	PDF 1.4	PDF 1.4	PDF 1.4	PDF 1.5	PDF 1.5	PDF 1.5	PDF 1.5/UA	PDF 1.6	PDF 1.7	XML and TXT	XML and TXT	XML and TXT	RTF 1.5–1.6 (ODM)	ePUB 2.0
EODOPEN Partners		NUK	NLE	NLS	CVRTI SR	UREG	UIBK*	OSZK	UIBK*	UG*	NCU*	UT*	BNP	VKOL	MZK ED.	MZK ed.	UIBK	UG
Alt-text picture	18										***17B***							
Alt-text chemical formula	2																	
Caption	19									5A 9B	***9A*** ***3B***	1A 6B						
Footnotes	1																	
Heading 1	7									***4A***	2A 3B							
Heading 2	10									1A	***5A***							
Heading 3	1									***1A***								
Initial	1	***1A***	***1A***	***1A***	***1A***	***1A***		***1A***				***1A***	***1A***					
Language segment	6																	
Math. (simple)	3	***3A***	2A	***3A***	2A	***3A***	2A	2A	2A	2A	***3A***	2A	***3A***	1A		1A	2A	
Math. (adv.)	4																***1B***	
OCR errors	1			***1A***				***1A***				***1A***	***1A***					
Page rotation	1		***1A***		***1A***	***1A***							1B			***1A***		
Pagination	12	***9A***	6A	5A	8A	8A	7A	6A	7A	***9A***	8A	8A	6A	7A	7A	8A	2A	1A
Pagination double	2	1A	***2A***	***2A***	1A	***2A***		***2A***		***2A***	***2A***	1A	***2A***		***2A***	***2A***		***2A***
Picture	18						6A		6A	***15A*** ***2B***	***16A*** ***1B***	13A 2B						14A 1B
Picture chem. formula	2									***1A***								***1A***
Primary language	1										***1A***						***1A***	
Special character	3	***2A***	***2A***	***2A***	***2A***	***2A***	1A	***2A***	1A	***2A***	***2A***	***2A***	1A	***2A***	***2A***	***2A***	1A	***2A***
Stamp removal	1							***1A***							***1A***	***1A***		
Table	4						3A	2A 1B	3A	3A	***4A***	3A					3A	2A 1B
Table header	4										***3A 1B***							
Table rows	4						2A 1B		2A 1B	1A 2B	***2A*** ***2B***	2B					2A 1B	3B
Text order	16	7A	2A 2B	5A 3B	9A 1B	10A 1B	3A 3B	5A	3A 3B	9A 1B	***12A***	10A	8A 2B	6A 5B	7A 3B	6A 2B	3A 3B	6A 1B

Notes: Used codes: A: fully achieved criterion; B: partly achieved criterion; empty cell: criterion was not achieved; *: tagged PDF; The best-achieved results for each criterion are in bold italics
Source: Created by authors

File formats	Ref. no.	PDF 1.6 ADOBE ACROBAT PRO	PDF 1.5/UA ABBYY 15	PDF 1.7 WORD	RTF 1.9	DOCX 2007-	ePUB 3.0 WORD
EODOPEN partners		NUK*	NUK*	NUK*	UIBK	BNP	NUK
Alt-text picture	18			***18A***	13B	***18A***	***18A***
Alt-text chemical formula	2			***2A***	2B	***2A***	***2A***
Caption	19		***11A 1B***	19B	14B		19B
Footnotes	1	1B		1B	***1A***		***1A***
Heading 1	7	***7A***	3A	***7A***		***7A***	***7A***
Heading 2	10	***10A***	1A	***10A***		8A	***10A***
Heading 3	1	***1A***		***1A***			***1A***
Initial	1	***1A***	***1A***	***1A***	***1A***	***1A***	***1A***
Language segment	6	2A 1B		***5A*** ***1B***			***6A***
Math. (simple)	3	***3A***	***3A***	***3A***	***3A***	***3A***	***3A***
Math. (adv.)	4			***4A***	***4A***	***4A***	***4A***
OCR errors	1	***1A***	***1A***	***1A***	***1A***	***1A***	***1A***
Page rotation	1		***1A***	***1A***	***1A***	***1A***	***1A***
Pagination	12	11A	8A 1B	***12A***	10A 1B	9A	***12A***
Pagination double	2	1A	1A	***2A***	***2A***	1A	1A
Picture	18	***18A***	***18A***	***18A***		***18A***	***18A***
Picture chem. formula	2	***2A***	***2A***	***2A***		***2A***	***2A***
Primary language	1	***1A***	***1A***	***1A***	***1A***	***1A***	***1A***
Special character	3	***2A***	***2A***	***2A***	***2A***	***2A***	***2A***
Stamp removal	1			***1A***	***1A***	***1A***	***1A***
Table	4		***4A***	***4A***	***4A***	***4A***	***4A***
Table header	4		***4A***	***4A***		***4A***	***4A***
Table rows	4		***4A***	***4A***	***4A***	***4A***	***4A***
Text order	16	14A	15A	***15A*** ***1B***	13A 1B	***15A*** ***1B***	***15A*** ***1B***

Notes: Used codes: A: fully achieved criterion; B: partly achieved criterion; empty cell: criterion was not achieved; *: tagged PDF; The best-achieved results for each criterion are in bold italics
Source: Created by authors

Table 3.
Results of outputs with additional manual corrections based on criteria

layout analysis. Some software enables correcting the automatic detection of the element type (text, picture, table, etc.) and its order of appearance on the page, but this requires more work and time.

No connections were found between format versions or standards, as none yielded better results, except for PDF/UA, which excelled compared to other versions of PDF.

Page rotation resulted in improved OCR results in the case of horizontally placed tables. Due to the performed rotation, the text in the table was better recognized (refer to Figure 2).

When elements were spread over two pages (e.g. title and author of an article in a newspaper), leaving them as a double page yielded better results (refer to Figure 3).

Library stamps posed a problem in most testing results. When the stamp was removed before the text was recognized, it yielded better results (refer to Figure 4).

In terms of OCR quality, manual corrections produced better results and significantly improved the quality, but this process was time-consuming in the case of mass digitization. The focus was on achieving the best automatic results before resorting to manual corrections.

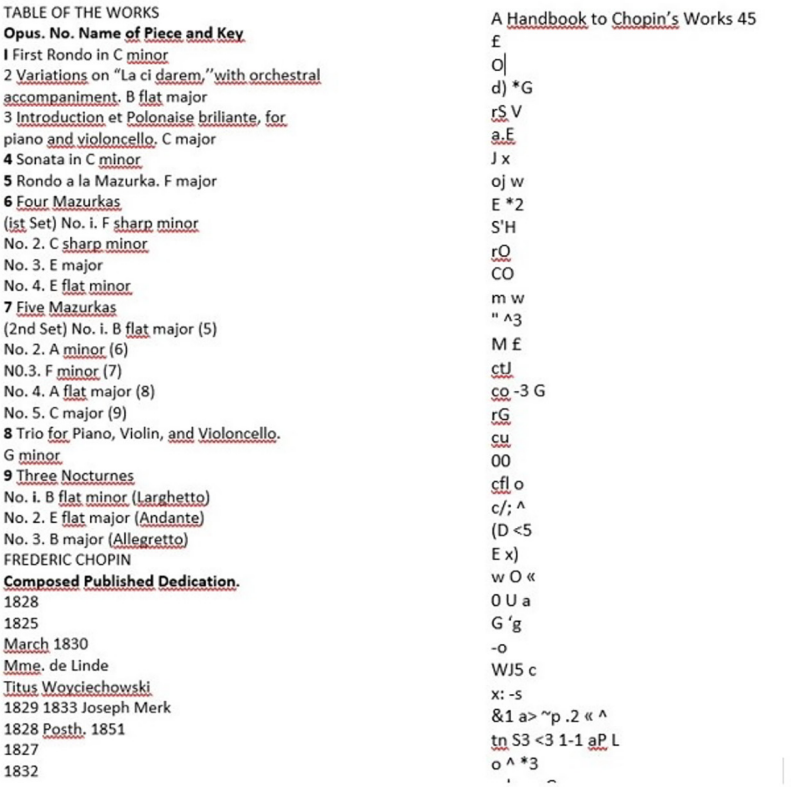
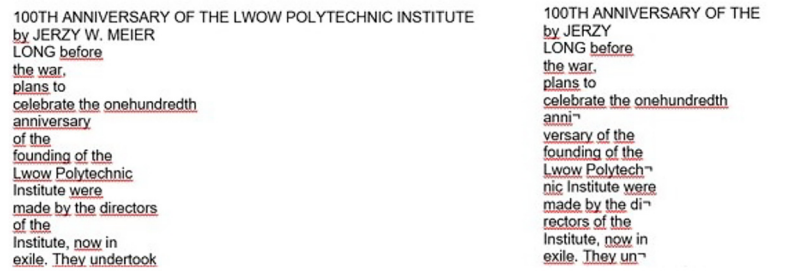


TABLE OF THE WORKS
Opus. No. Name of Piece and Key
I First Rondo in C minor
2 Variations on "La ci darem," with orchestral accompaniment. B flat major
3 Introduction et Polonaise briliante, for piano and violoncello. C major
4 Sonata in C minor
5 Rondo a la Mazurka. F major
6 Four Mazurkas
(ist Set) No. i. F sharp minor
No. 2. C sharp minor
No. 3. E major
No. 4. E flat minor
7 Five Mazurkas
(2nd Set) No. i. B flat major (5)
No. 2. A minor (6)
N0.3. F minor (7)
No. 4. A flat major (8)
No. 5. C major (9)
8 Trio for Piano, Violin, and Violoncello.
G minor
9 Three Nocturnes
No. i. B flat minor (Larghetto)
No. 2. E flat major (Andante)
No. 3. B major (Allegretto)
FREDERIC CHOPIN
Composed Published Dedication.
1828
1825
March 1830
Mme. de Linde
Titus Woyciechowski
1829 1833 Joseph Merk
1828 Posth. 1851
1827
1832

A Handbook to Chopin's Works 45
£
O
d) *G
rS V
a.E
J x
oj w
E *2
S'H
rO
CO
m w
" ^3
M £
ctJ
co -3 G
rG
cu
00
cfl o
c/; ^
(D <5
E x)
w O «
0 U a
G 'g
-o
WJ5 c
x: -s
&1 a> ~p .2 « ^
tn S3 <3 1-1 aP L
o ^ *3

Note: The first example (left) is the output when the page was rotated, and the second example (right) is the output when the page was not rotated
Source: All images (scans from books) included are in public domain

Figure 2.
Comparison of two OCR outputs of scan no. 10

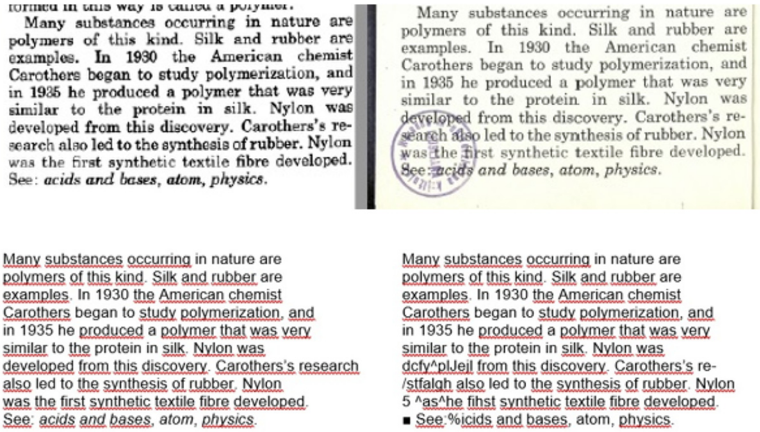


100TH ANNIVERSARY OF THE LWOW POLYTECHNIC INSTITUTE
by JERZY W. MEIER
LONG before
the war,
plans to
celebrate the onehundredth
anniversary
of the
founding of the
Lwow Polytechnic
Institute were
made by the directors
of the
Institute, now in
exile. They undertook

100TH ANNIVERSARY OF THE
by JERZY
LONG before
the war,
plans to
celebrate the onehundredth
anni¬
versary of the
founding of the
Lwow Polytech¬
nic Institute were
made by the di¬
rectors of the
Institute, now in
exile. They un¬

Notes: The first example (left) is the output when the double page was not split in two, resulting in the entire title and author at the top. The second example (right) is the output when the page was split in two, resulting in only part of the title and author appearing. Some other OCR differences are visible
Source: All images (scans from books) included are in public domain

Figure 3.
Comparison of two OCR outputs of scan no. 7



formed in this way is called a polymer.
Many substances occurring in nature are polymers of this kind. Silk and rubber are examples. In 1930 the American chemist Carothers began to study polymerization, and in 1935 he produced a polymer that was very similar to the protein in silk. Nylon was developed from this discovery. Carothers's research also led to the synthesis of rubber. Nylon was the first synthetic textile fibre developed. See: *acids and bases, atom, physics.*

Many substances occurring in nature are polymers of this kind. Silk and rubber are examples. In 1930 the American chemist Carothers began to study polymerization, and in 1935 he produced a polymer that was very similar to the protein in silk. Nylon was developed from this discovery. Carothers's research also led to the synthesis of rubber. Nylon was the first synthetic textile fibre developed. See: *acids and bases, atom, physics.*

Many substances occurring in nature are
polymers of this kind. Silk and rubber are
examples. In 1930 the American chemist
Carothers began to study polymerization, and
in 1935 he produced a polymer that was very
similar to the protein in silk. Nylon was
developed from this discovery. Carothers's research
also led to the synthesis of rubber. Nylon
was the first synthetic textile fibre developed.
See: *acids and bases*, *atom*, *physics*.

Many substances occurring in nature are
polymers of this kind. Silk and rubber are
examples. In 1930 the American chemist
Carothers began to study polymerization, and
in 1935 he produced a polymer that was very
similar to the protein in silk. Nylon was
dcfy^plJeil from this discovery. Carothers's re-
/stfalgh also led to the synthesis of rubber. Nylon
5 ^as^he fihst synthetic textile fibre developed.
■ See:%icids and bases, atom, physics.

Notes: On the left side are the outputs when the stamp was removed, resulting in a clean text. The second example on the right is the output when the stamp was not removed, which creates reading difficulties
Source: All images (scans from books) included are in public domain

Figure 4.
Comparison of two PDF and OCR outputs of scan no. 3

5. Discussion

The primary objective of our research was to ascertain the optimal methods for achieving high-quality OCR in digitized texts resulting from automatic mass scanning processes. An analysis of EODOPEN partners' testing reports indicated that the prescanning phase had limited relevance to the quality of the received OCR results, aligning with similar findings noted by Jääskeläinen *et al.* (2023). Our research concentrated on processes from the scanning phase onward, considering the quality of resulting OCR in relation to the scanning output file formats.

The choice of OCR tool significantly influences OCR quality. Newer versions of tools have demonstrated improved results, as illustrated by Kettunen *et al.* (2020). According to workflow reports, the majority of EODOPEN partners use the OCR software ABBYY Fine Reader. Trbušić (2022) found that this tool supports most file formats and, in our case, yielded superior results compared to other tools. To create texts accessible to blind and partly sighted users, manual correction was often necessary. Otherwise, the workflows of reporting partners' organizations were remarkably similar.

The results of the conducted testing revealed that most EODOPEN partners' digitization outputs were delivered in different versions of PDF file format. However, superior results were achieved when using PDFs that were tagged or in PDF/UA file format. The PDF/UA file format, following the ISO 14289-1 standard for universal accessibility, incorporates hidden markups that have the potential to enhance the reading experience for users of mobile devices and assistive technologies, such as the blind and partially sighted. This is particularly crucial for preserving the original structure of accessed contents. Furthermore, as highlighted by Razpet (2017) and Rello and Baeza-Yates (2017), for an improved reading experience among mobile device users and partly sighted individuals, it is essential that the text can adapt to the size of the screens. Unfortunately, PDF/UA does not provide this functionality.

As a consequence, we focused on two different aspects relevant to the accessibility with assistive technologies: criteria related to elements in the digitized printed material, such as

images, mathematical and chemical formula, foreign language text and tables, which are crucial for reading and understanding the scanned texts; and criteria related to the layout of scanned printed material. These 24 criteria proved to be most relevant in defining the OCR quality.

For blind and partially sighted users, visual elements pose significant accessibility challenges because these elements require descriptions that contribute additional value to the surrounding text, such as alt-text for images, graphs, etc. The testing revealed poor results unless the work was done manually, emphasizing that image descriptions still require human input.

Specifically for blind users, visual elements are often unnecessary, but alt-text is of utmost importance. It is recommended to use formats that do not contain visual elements and support assistive technology, such as TXT or variations of Microsoft Word documents. As mentioned before, alt-text still requires human input. For partially sighted individuals who can still use their remaining sight, it is recommended to use formats that are adaptable to screens and support modern functionalities, such as adding bookmarks and changing visual appearance. Variations of Microsoft Word documents and EPUB showed good results, and in the literature, for example, Ganner *et al.* (2023), HTML is also recommended as it contains semantically structured content that is adaptable to screen sizes.

Tables are a graphical presentation of data but can pose several difficulties regarding accessibility for users with special needs. Consistent with the literature, for example, Ganner *et al.* (2023), and other sources, our testing has also demonstrated that the best accessibility is achieved when tables are not presented as images but instead contain a structure from table headers to table rows and table cells.

Similarly, complex mathematical expressions present an accessibility challenge. In our testing, it was revealed that MathML was the most accessible solution, while some literature, for example, Ganner *et al.* (2023), suggests alternative use of LaTeX.

Implementation should take into account the specific context, user needs and available technologies. Small screens on mobile devices, particularly mobile phones, pose a challenge for accessing nonresponsive delivery file formats. These are formats that are adaptable to screens and allow for basic visual adaptations of the text to the personal needs of users, such as EPUB, MOBI, AZW, HTML or variations of Microsoft Word documents. Among our testing outputs, we obtained some results with formats like EPUB and Microsoft Word Documents, among which automatically generated results have not met many of the set accessibility criteria.

The testing has indicated that the navigation through the document is best achieved when the table of contents was included at the beginning of the content, which was mostly done with manual work, or when structural tags that mark the chapters in the publication are present.

The criteria "page rotation," "pagination double" and "stamp removal," directly connected with digitization workflows, have shown varied results. They can enhance visual appearance and OCR results when pages are rotated, double pages are not split and library stamps are removed before conducting OCR. However, decisions regarding these criteria should be made on a case-by-case basis, given that our testing sample had only one occurrence of each of these criteria.

Users of mobile devices may also use assistive technologies like speech synthesis. To achieve such results, findings for print-disabled users should be considered to ensure access for the widest possible group of users. Conversely, in regard to blind and partially sighted or other users with special needs, the results and conclusions are more complex and challenging to achieve with automatic processes, necessitating more human input.

Considering our findings based on testing results and reviewed literature, the specific needs of print-disabled users, available assistive technologies and even the publishing workflow require careful consideration. As mentioned above, solutions for the blind and partially sighted are more complex and demand additional work, time and expertise in this field. It is crucial not only to provide access to publications but also to ensure the effective

use of assistive technology, enabling a smooth flow of text despite complex elements and demanding page structures.

To ensure optimal access to digitized materials, it is crucial to perform OCR clean-up and verify the correct text order. Assistive technologies deliver text to users in a linear order from top to bottom, making unclean text or mixed order a potential source of difficulty for such users.

Digitized materials often exhibit insufficient color contrasts, particularly in scanned PDFs where low contrast between text and background can result from the paper color. Only two of the partner institutions used binarization to convert images to black and white, presumingly aiming for improved OCR accuracy, but this approach also enhanced reading contrast.

The testing revealed challenges in detecting document language and segments that are in different languages. This is especially critical for users of screen readers to ensure the correct application of speech synthesis (e.g. German text should not be spoken with an English voice).

6. Conclusion

In our research, we focused on delivery formats produced by EODOPEN project partners. Although their workflows were very similar and not highly relevant to the outputs, the significance became evident in cases where manual corrections were applied. We considered the relationships between used OCR tools and the quality of the produced output formats. The analysis was the qualitative approach, omitting a quantitative assessment of OCR accuracy, as the primary objective was to find optimal solutions for accessing digital content by users of mobile devices and the blind and partially sighted.

Our findings indicate that preprocessing contributes minimally, with more significant improvements observed in the postprocessing phase. The key advantage demonstrated was the higher accessibility achieved through the use of tagged PDF or PDF/UA as digitization delivery file formats.

Most of the criteria set for the testing phase align with standards such as WCAG. We conducted tests using various software and one assistive technology. In future research, it would be beneficial to involve focus groups or test groups of blind and partially sighted individuals. Gathering feedback and incorporating observations from these groups into future workflows can significantly contribute to further development in this field.

In our research, we did not identify solutions for digitization that guarantee enhanced accessibility when processes are automated. Achieving accessibility for special needs, especially for the blind and partially sighted, involves complex workflows, requires human input and takes a longer time to adapt manually. These workflows are more intricate compared to those ensuring access to digitized content through mobile devices.

As a limitation of our research, we acknowledge the small sample of workflows tested, which involved only 13 EODOPEN partner institutions. For more comprehensive results, a larger number of organizations should be included to test their digitization outputs and evaluate them according to the established criteria. In future research, it would be beneficial to incorporate criteria related to adaptation to different screen sizes. Additionally, the testing sample, consisting of 16 scans, might yield different results with a larger and more homogenous sample. The current sample, derived from various publications, aimed to identify potential errors during OCR recognition. However, this approach posed challenges for some workflows within the partner institutions, as content from a single publication may be handled differently than diverse scans with varying printing fonts and styles from different publications. These are some of the limitations we encountered, and addressing them differently could be considered in future or subsequent research in this field.

A similar study focusing on conversion services, enabling the transformation of delivery formats (e.g. PDF to DOCX or DOCX to MP3), is underway. The evaluation of these

conversion services will be presented as a technical report on the implementation of special formats and conversion services, which will be made publicly available.

One of the outcomes of our research will be the development of training materials and courses designed for librarians and other cultural heritage institution workers involved in digitization. These resources will specifically emphasize delivery formats suitable for readers on mobile devices and users who are blind or partially sighted.

Notes

1. More information on the European Accessibility Act at website: https://ec.europa.eu/social/main.jsp?catId=1202
2. Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services. More information available at website: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L0882
3. Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (Text with EEA relevance) PE/81/2018/REV/1, available on 17 July 2023 at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0882
4. Whole standard is available at: www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf. Refer also to Campbell *et al.*, 2023.
5. Web Content Accessibility Guidelines 2.1 Web page: www.w3.org/TR/WCAG21/
6. Web Content Accessibility Guidelines 2.2 was officially released on 5th October 2023 but the mentioned EU directive obligates the liable party that their products conform to WCAG 2.1. Professionals suggest to follow the currently newest version.
7. World Wide Web Consortium. (16. 5. 2023). *WCAG 3 Introduction*. www.w3.org/WAI/standards-guidelines/wcag/wcag3-intro/
8. eBooks-On-Demand-Network Opening Publications for European Netizens. The project website is available at: https://eodopen.eu/
9. "Optical character recognition (OCR) is a method of detecting and recognizing typed, handwritten, printed or captured text into machine encoded text [. . .]" (Refer to Majumder, Mahmud, Jahan, and Alam, 2019).
10. More about the WCAG is explained in the chapter: related research.
11. Project COMHIS website: https://blogs.helsinki.fi/natlibfi-bulletin/?page_id=757
12. Project IMPACT website: www.ukoln.ac.uk/projects/impact/index.html
13. Project tranScriptorium website: www.transkriptorium.com/
14. Project READ website: https://readcoop.eu/
15. Platform transkribus website: http://transkribus.eu
16. Project PERO website: https://pero.fit.vutbr.cz/about
17. According to (Ishihara *et al.* 2012, p. 93), these steps include: (1) text correction of the OCR results; (2) reading order compensation; (3) adding structures for blocks of content: headings, table structures, paragraphs and footnotes; (4) adding structures to support navigation: table of contents, references and physical page numbers; (5) adding descriptions or structures for nontext objects: alternative text for graphics, associated labels for graphic objects and convert images to vector graphics; (6) adding information for complex text block: mathematical formulas, ruby (pronunciation hints for Japanese characters and sections in other languages.

18. Refer to image 4 on scan no. 7 at Appendix 1.
19. To read more about these criteria, refer to Appendix 2 for further details.
20. In Appendix 1 is the complete list of used scans for testing purposes.
21. University of Innsbruck, Austria (UIBK), Moravian Library in Brno (MZK) Czech Academy of Sciences Library (KNAV) and Research Library Olomouc (VKOL), Czech Republic, National Library of Estonia (NLE) and Tartu University Library (UT), Estonia, University of Greifswald Library (UG) and University of Regensburg (UREG), Germany, National Széchényi Library, Hungary (OSZK), Nicolaus Copernicus University in Torun, Poland (NCU), National Library of Portugal, Portugal (BNP), Slovak Centre of Scientific and Technical Information, Slovakia (CVTI SR), National and University Library, Slovenia (NUK) and National Library of Sweden, Sweden (NLS).
22. Several guidelines and recommendations for digitization have been produced through various European projects. An illustrative example is the document created during the European Travel project by Štular Sotošek (2011): *Best practice examples in library digitisation*, which can be accessed at: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaTravel/Deliverables/D2.2%20Best%20practice%20examples%20in%20library%20digitisation.pdf
23. The most frequently used techniques were deskewing and cropping, executed either automatically or manually. Two partners used equalizing the dimensions of the scans before conducting OCR, while three partners implemented lines straightening (dewarping). Additionally, two partners used binarization and the removal of stamps and written notes, and three different partners implemented contrast enhancement. Noise removal (denoising) was carried out by two partners, and one partner focused on the correction of geometric distortion.
24. Captions can be inserted technically. In tagged PDFs, for example, a specific tag can be added in Adobe Acrobat Pro. When working in Microsoft Word, the "insert caption" option can be used.
25. Good results can be achieved, for example, in Microsoft Word, HTML or ePUB by providing two-way hyperlinks.
26. For example, a German text that is read aloud with an English voice sounds strange.
27. The structure of the table can be created technically. For example, in tagged PDFs, tags appear for table, table header, table rows and table data, much like in HTML formatting. Microsoft Word, for instance, also has the option to set a table header.
28. The most frequently used software for OCR – ABBYY FineReader desktop version – has this option during processing the digitized content. For postprocessing, an example of software of this kind is Adobe Acrobat Pro.

References

Auer, C., Nassar, A., Lysak, M., Dolfi, M., Livanthinos, N. and Staar, P. (2023), "ICDAR 2023 competition on robust layout segmentation in corporate documents", ICDAR, 24 May 2023. Available in arXiv:2305.14962v1.

Avyodri, R., Lukas, A. and Tjahyadi, H. (2022), "Optical character recognition (OCR) for text recognition and its post-processing method: a literature review", *1st International Conference on Technology Innovation and Its Applications (ICTIIA)*, doi: 10.1109/ICTIIA54654.2022.9935961.

Calvert, P., Creaser, C. and Pigott, C. (2019), "Information access preferences and behaviour of blind foundation library clients", *Journal of Librarianship and Information Science*, Vol. 51 No. 1, pp. 162-170.

Campbell, A., Adams, C., Bradley Montgomery, R., Cooper, M. and Kirkpatrick, A. (Eds) (2023), "Web content accessibility guidelines (WCAG) 2.2. W3C recommendation", 5 October 2023, available at: www.w3.org/TR/WCAG22/#input-modalities (accessed 22 November 2023).

Creating accessible Word documents (2023), "The daisy consortium", available at: https://daisy.org/guidance/info-help/guidance-training/daisy-tools/creating-accessible-word-documents/

Erkilinc, M.S., Jaber, M., Saber, E., Bauer, P. and Depalov, D. (2011), „"Page layout analysis and classification for complex scanned documents", *Proceedings of SPIE – The International Society for Optical Engineering*, Vol. 8135, September 2011, Applications of Digital Image Processing XXXIV, p. 813507, doi: 10.1117/12.893909.

Femc, M. (2018), "E-knjige kot izziv: primer mestne knjižnice ljubljana", Master thesis, Univerza v Ljubljani, Filozofska fakulteta, available at: https://repozitorij.uni-lj.si/Dokument.php?id=116771&lang=slv

Forczmański, P., Smoliński, A., Nowosielski, A. and Małecki, K. (2020), "Segmentation of scanned documents using Deep-Learning approach", in Burduk R., Kurzynski M., Wozniak M. (Eds), *Progress in Computer Recognition Systems. Advances in Intelligent Systems and Computing*, Vol. 977 Springer, Cham, pp. 141-152.

Ganner, J., Mrva-Montoya, A., Park, M. and Duncan, K. (2023), "Books without barriers. A practical guide to inclusive publishing", Institute of Professional Editors, available at: www.iped-editors.org/wp-content/uploads/2023/04/Bookswithoutbarriers_Screen.pdf

Gao, L., Huang, Y., Déjean, H., Meunier, J.L., Yan, Q., Fang, Y., Kleber, F. and Lang, E. (2019), "ICDAR 2019 competition on table detection and recognition (cTDaR)", doi: 10.1109/ICDAR.2019.00243

Gunn, D. (2016), "Accessible eBook guidelines for self-publishing authors", Accessible Books Consortium; International Authors Forum, available at: https://internationalauthors.org/wp-content/uploads/2017/11/Accessible-eBook-Guidelines-for-Self-Publishing-Authors.pdf

Inbasekaran, A., Gnanasekaran, K.R. and Marciano, R. (2021), "Using transfer learning to contextually optimize optical character recognition (OCR) output and perform new feature extraction on a digitized cultural and historical dataset", *2021 IEEE International Conference on Big Data (Big Data)*, *Orlando, FL, USA*, pp. 2224-2230, doi: 10.1109/BigData52589.2021.9671586.

Ishihara, T., Itoko, T., Sato, D., Tzadok, A.I. and Takagi, H. (2012), "Transforming Japanese archives into accessible digital books", *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 91-100, Association for Computing Machinery, doi: 10.1145/2232817.2232836.

Jääskeläinen, A., Lipsanen, M., Föhr, A. and Räisänen, T. (2023), "OCR quality: Key to enhanced data mining", *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 19-21 July 2023, *Tenerife, Canary Islands, Spain*, pp. 1-6.

Kettunen, K., Koistinen, M. and Kervinen, J. (2020), "Ground truth OCR sample data of Finnish historical newspapers and journals in data improvement validation of a re-OCRing process", *LIBER Quarterly*, Vol. 30 No. 1, pp. 1-20, doi: 10.18352/lq.10322.

Kodrič-Dačić, E., Vodeb, G., Bon, M., Poličnik-Čermelj, T. and In Vilar, P. (2014), "Vzpostavitev infrastrukture za zagotavljanje enakih možnosti dostopa do publikacij slepim in slabovidnim ter osebam z motnjami branja. Model knjižnice za slepe, slabovidne in osebe z motnjami branja, vključno z modelom zagotavljanja in koordinacije knjižničnih storitev za slepe, slabovidne in osebe z motnjami branja na področju celotne države", Narodna in univerzitetna knjižnica, available at: www.kss-ess.si/wp-content/uploads/2016/04/NUK-Izdelava-modela-KSS.pdf

Majumder, M.M.R., Mahmud, B.U., Jahan, B. and Alam, M.M. (2019), "Offline optical character recognition (OCR) method: an effective method for scanned documents", *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 18-20 December, doi: 10.1109/ICCIT48885.2019.9038593.

Mulliken, A. and Falloon, K. (2017), "Blind academic library users' experiences with obtaining full text and accessible full text of books and articles in the USA. A qualitative study", *Library Hi Tech*, Vol. 37 No. 3, pp. 456-479.

Razpet, M. (2017), "Biblos. In S. Zwitter in N. Bucik (ur.)", E-gradiva kot bližnjica do uspeha?, p. 71-76, Bralno društvo Slovenije, available at: www.bralno-drustvo.si/wp-content/uploads/2013/06/E-GRADIVA_KOT_BLIZNJICA_DO_USPEHA_BDS20171.pdf

Reisswig, C., Katti, A.R., Spinaci, M. and Höhne, J. (2020), "Chargid-OCR: End-to-end Trainable Optical Character Recognition for Printed Documents using Instance Segmentation", arXiv:1909.04469v4 [cs.CV], available at: https://doi.org/10.48550/arXiv.1909.04469

Rello, L. and Baeza-Yates, R. (2017), "How to present more readable text for people with dyslexia", *Universal Access in the Information Society*, Vol. 16 No. 1, pp. 29-49, doi: 10.1007/s10209-015-0438-8.

Rigaud, C., Doucet, A., Coustaty, M. and Moreux, J.P. (2019), "ICDAR 2019 competition on post-OCR text correction", HAL Id: hal-02304334, available at: https://hal.science/hal-02304334

Section 508 Guide: Tagging PDF's in Adobe Acrobat Pro (2018), "U.S. Department of health and human services", available at: www.hhs.gov/sites/default/files/pdf-tagging.pdf

Shafait, F. (2008), "Geometric layout analysis of scanned documents", Dissertation, Department of Computer Science, Technical University of Kaiserslautern, available at: https://d-nb.info/989481123/34 (accessed 12 July 2023).

Tagged PDF Best Practice Guide: Syntax (2019), "PDF association", available at: https://pdfa.org/resource/tagged-pdf-best-practice-guide-syntax/ (accessed 18 July 2023).

Trbušić, Ž. (2022), "Metode analize i optimizacije procesa optičkog prepoznavanja znakova u arhivskim informacijskim sustavima. Doktorski rad'. (= Methods for analysis and process optimisation of optical character recognition in archival information systems", Doctoral Thesis, Zagreb.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F. and Zhou, M. (2020), "LayoutLM: Pre-training of text and layout for document image understanding", arXiv:1912.13318v5 [cs.CL], available at: https://arxiv.org/pdf/1912.13318v5.pdf (accessed 16 Jun 2020).

Zaviršek, D., Kačič, M., Krstulović, G. and Sobočan, A.M. (2013), "Izvedbeni projekt zadovoljevanja potreb potencialnih deležnikov, uporabnikov in posebnih potreb slepih, slabovidnih in oseb z motnjami branja", Fakulteta za socialno delo, available at: www.kss-ess.si/2016/02/studija-fsd-zadovoljevanje-potreb-potencialnih-deleznikov-uporabnikov-in-posebnih-potreb-slepih-slabovidnih-in-oseb-z-motnjami-branja/#.XykioSgzaUl

Zhong, X., Tang, J. and Yepes, A.J. (2019), "PubLayNet: largest dataset ever for document layout analysis", arXiv:1908.07836v1 [cs.CL] 16 Aug 2019, available at: https://arxiv.org/pdf/1908.07836v1.pdf (accessed 12 July 2023).

Zhong, Y., Qi, X., Li, S., Gu, D., Chen, X., Ning, P. and Xiao, R. (2021), "1st place solution for ICDAR 2021 competition on mathematical formula detection", ICDAR, 12 July 2021. Available in arXiv:2107.05534v1.

Further reading

Drümmer, O. and Chang, B. (2013), "PDF/UA in a nutshell", Accessible documents with PDF, PDF Association, available at: www.pdfa.org/wp-content/until2016_uploads/2013/08/PDFUA-in-a-Nutshell-PDFUA.pdf (accessed 22 November).

Konicek, K., Hyzny, J. and Allegra, R. (2003), "Electronic reserves: the promise and challenge to increase accessibility", *Library Hi Tech*, Vol. 21 No. 1, pp. 102-108.

The DAISY Consortium (2023), "Creating accessible word documents", available at: https://daisy.org/info-help/guidance-training/daisy-tools/creating-accessible-word-documents/

Appendix 1. Sources of testing samples

- Scan no. 1.

Source: *The Lexicon Encyclopedia of Science* (p. 5 volumes). (1984). Lexicon Publications, Inc., vol. 1, p. 39.

- Scan no. 2.

Source: *The Lexicon Encyclopedia of Science* (p. 5 volumes). (1984). Lexicon Publications, Inc., vol. 1, p. 152.

- Scan no. 3.

Source: *The Lexicon Encyclopedia of Science* (p. 5 volumes). (1984). Lexicon Publications, Inc., vol. 1, p. 156.

- Scan no. 4.

Source: *The Lexicon Encyclopedia of Science* (p. 5 volumes). (1984). Lexicon Publications, Inc., vol. 1, pp. 154–155.

- Scan no. 5.

Source: *The Lexicon Encyclopedia of Science* (p. 5 volumes). (1984). Lexicon Publications, Inc., vol. 1, p. 132.

- Scan no. 6.

Source: *Sprawa: dwutygodnik Polskiego Instytutu "Miecz Ducha" 1945, R. 4 nr 4/5*. (1945). p. [5]. Available at: https://kpbc.umk.pl/dlibra/publication/228476/edition/227280/content

- Scan no. 7.

Source: *The Polish Review*. (1944). Vol. 4 no. 28, p. 4–5. Available at: https://kpbc.umk.pl/dlibra/publication/235389/edition/233504/content.

- Scan no. 8.

Source: Mitchell, Margaret. (1936). *Gone with the wind*, p. 4. Available at: https://kpbc.umk.pl/dlibra/publication/255991/edition/255385/content

- Scan no. 9.

Mickiewicz, Adam. (1885). *Master Thaddeus; or, the Last foray in Lithuania: an historical epic poem in twelve books. Vol. 1*., p. 1. Available at: https://kpbc.umk.pl/dlibra/publication/215648/edition/233290/content

- Scan no. 10.

Source: Jonson, G. C. Ashton. (1908). *A handbook to Chopin's works: giving a detailed account of all the compositions of Chopin, short analyses for the piano student, and critical quotations from the writings of well-known musical authors: the whole forming a complete guide for concert-goers, pianists and pianola-players, also a short biography, critical bibliography and a chronological list of works, etc., p. 45.* Available at: https://kpbc.umk.pl/dlibra/publication/225034/edition/234756/content

- Scan no. 11.

Source: Jonson, G. C. Ashton. (1908). *A handbook to Chopin's works: giving a detailed account of all the compositions of Chopin, short analyses for the piano student, and critical quotations from the writings of well-known musical authors: the whole forming a complete guide for concert-goers, pianists and pianola-players, also a short biography, critical bibliography and a chronological list of works, etc., p. 54.* Available at: https://kpbc.umk.pl/dlibra/publication/225034/edition/234756/content

- Scan no. 12.

Source: Humphreys, William Jackson. (1940). *Physics of the air*, p. 13. Available at: https://kpbc.umk.pl/dlibra/publication/194654/edition/209487/content

- Scan no. 13.

Source: Dresser, Henry E. (1910). *Eggs of the birds of Europe*, Vol. 1, p. 27. Available at: https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr11969-2

- Scan no. 14.

Source: *German Democratic Republic: International Agricultural Exhibition in Cairo*. (1961). p. [22]. Available at: https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr10353-8

- Scan no. 15.

Source: Hawks, Francis L. and Perry, Matthew Calbraith. (1967). *Narrative of the expedition of an American squadron to the China Seas and Japan: Performed in the years 1852, 1853, and 1854, under the command of Commodore M. C. Perry, United Dates Navy, by order of the Government of the United States, p. [VII]*. Available at: https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr07179-4

- Scan no. 16.

Source: *Annual report of the State Bee Inspector/4. 1915*. (1916), p. 46. Available at: https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr15095-5

Appendix 2. Evaluation criteria

- *Alt-text picture*: Alt-text or alternative text for pictures provides a textual description for nontext content (pictures, graphics, diagrams, etc.). These are elements that enable mostly blind users, but also partially sighted users, to know the content of the graphic material so that they do not miss any information that the graphic material may be trying to convey. This criterion is primarily important to the blind and partially sighted but could also be useful to sighted users using speech synthesis.
- *Alt-text picture (chemical formula)*: Same as the criterion alt-text picture, but used for the two special images in the test sample that presented molecular reactions (refer to Scan no. 2 in Appendix 1).
- *Caption*: Some of the images and tables in the scans contained captions. In the document, it should be indicated that the text is a caption associated with a picture and not general paragraph text [24]. This criterion is primarily important for the blind and partially sighted.
- *Footnotes*: Footnotes are elements in a document that provide additional information related to the main text and should be technically separated from the main text, thus giving readers the option of skipping them. When creating or editing footnotes, the result should enable the reader to jump from the main text to the footnote and then back to the same area in the text [25]. This criterion is mainly important for the blind and partially sighted.
- *Heading 1*: Mainly for navigational purposes, the headings of the chapters should be marked and structured in-depth (Headings 1, 2, 3, etc.). Headings can also be used to form a table of contents. This enables users of assistive technologies to skip from chapter to chapter more easily and, thus, to navigate within the document instead of reading the whole document. This criterion is important for all users.
- *Heading 2*: Refer to criterion Heading 1.

- *Heading 3*: Refer to criterion Heading 1.
- *Initial*: A larger first letter at the beginning of a chapter is often not recognized or not recognized correctly (refer to Scan no. 7 in Appendix 1). This criterion is important for all users.
- *Language segments*: Refer to the criterion Primary Language. The Language Segments criterion was used on six different occasions in the test sample (Italian + Latin, Italian, French twice and German twice) where text appeared in a language other than English, which was the primary language. The language is important for users of screen reading technologies in which voice settings can be switched to the correct audio to provide proper pronunciation [26]. This criterion is primarily important for the blind and partially sighted but could also be useful to sighted users using speech synthesis.
- *Math (simple)*: The recognition of mathematical or chemical elements was divided into two criteria, as it is mainly simple mathematical elements that appear in one single line that create less problems for OCR [example from the test sample: $e = e' - AB(t - t')$] than advanced math which appears in more than one line. This criterion is primarily important for the blind and partially sighted.
- *Math (advanced)*: The second criterion for mathematical and chemical elements covers all expressions that appear in two or more lines. These elements are not usually recognized correctly during OCR. This criterion includes all elements with subscripts or superscripts (examples from the test sample: x^2, $2H_2O$, 10^{-4}, $C_6H_{12}O_6$), fractions (example from the test sample: $\frac{x}{3}$) or even more complicated expressions (examples from the test sample: $\Delta p = \rho_v gh$ or $\Delta p = \frac{2T\rho_v}{R(\rho_w - \rho_v)}$). The examples from the test sample contain various problematic elements (e.g. subscripts, superscripts, Greek letters and fractions). This criterion is primarily important for the blind and partially sighted.
- *OCR errors (text in picture 4 on Scan no. 7)*: One image showed text written on a tombstone (refer to Scan no. 7 in Appendix 1). Ideally, a text of this kind would not be recognized, but the goal was to see what kind of results would be obtained. This criterion is important for all users.
- *Page rotation*: This criterion was only used in one case where a table appeared horizontally on a page. For better OCR and structure results, the page could be turned so that the table would face the reader correctly. This criterion is important for all users.
- *Pagination*: This criterion was created for the purposes of the blind and partially sighted. Practice shows that blind and partially sighted users prefer the pagination to be the first information they receive when entering a page. When working on text order, the preference is for pagination to be the first information received, even if it actually appears at the bottom of the page. This criterion is primarily important for the blind and partially sighted but could also be useful to sighted users using speech synthesis or for easier navigation to the specific page in the document.
- *Pagination-double*: This criterion was used in two different cases when content appeared stretched across two pages. The first case involved an image of the periodic table of elements, while the second case concerned the title and author of the article, which were stretched across two pages. In both cases, better results would be obtained if the pages were not split. This criterion is important for all users.
- *Picture*: A graphic element that should be marked as a separate element and contain alt-text for users of assistive technologies. This criterion is primarily important for the blind and partially sighted.

- *Picture (chem. formula)*: Same as the criterion Picture. This was a separate criterion for two images that presented molecular reactions, which should also contain alt-text. This criterion is primarily important for the blind and partially sighted.
- *Primary language*: The primary language should be set for each document. This is important for users of screen reading technologies that provide sound in the correct language. The text in the test sample was in English, so the primary language should be set to English. This criterion is mainly important for the blind and partially sighted but could also be useful for sighted users using speech synthesis.
- *Special character*: This criterion appeared in three different cases (°C, £ and decimal numbers). The goal was to determine the number of examples in which there would be problems recognizing the first two characters. In the scan with decimal numbers, the numbers are written with an apostrophe (′), which the English vocabulary fails to recognize because full stops (.) are normally used for decimal numbers in English. The scan was tested to see whether we would receive any correct results. This criterion is important for all users.
- *Stamp removal*: Library stamps in books can affect the recognition of nearby characters. The goal was to determine whether removing the stamp from the scan would ensure clearer OCR in that area. In our example, the stamp was directly over the text, and we assumed that it would cause bad OCR results. This criterion is important for all users.
- *Table*: This is a structural element that should be technically marked and should not appear as an image only. Following the structure, the table header and table rows should also be present [27]. This criterion is primarily important for the blind and partially sighted but could also be useful to sighted users using speech synthesis.
- *Table header*: This is an element of a table that usually appears at the top of the table but can also be in the first column of the table. It provides the main information about the data in the rows following it, and it is important for users of assistive technologies for easier navigation and understanding of the table. This criterion is primarily important for the blind and partially sighted but could also be useful to sighted users using speech synthesis.
- *Table rows*: These are structural elements following the table header. For the test sample, which did not contain a grid to mark the lines in the table, it was interesting to see whether the rows had technical data inserted and how well the OCR tool could recognize the number of rows. This criterion is primarily important for the blind and partially sighted but could also be useful to sighted users using speech synthesis.
- *Text order*: This criterion establishes the flow of the text, especially when the structure on the page is more complicated (e.g. columns and additional graphical elements). When users copy text, convert the format or use assistive technology, it is important that the text is presented in the right order so as to prevent confusion (e.g. if a caption appears in the middle of a paragraph) or to avoid burdening users with the additional work of editing the content themselves. Some software tools for OCR also enable correcting the order of the recognized elements [28]. Furthermore, assistive technologies provide users with text linearly from top to bottom, so the text order is crucial for understanding and navigating the content. This criterion is important for all users.

Appendix 3. Testing report questionnaire

A12 TEST REPORT (SAMPLE)

Please, add detailed information! You can also add screenshots or record the testing process.

Partner organisation: _____
Which software for image processing and OCR did you use for this sample? ____

1. IMPORT OF SCANS IN TIFF FORMAT

Before uploading the sample files to your system, did you change anything, for instance resolution, scanning format etc.?
- yes
- no

If yes, what did you change? _____

2. IMAGE PROCESSING

Mark which image processing steps you used when working with the sample.

Deskewing:
- automatic
- manual
- automatic and manual

Cropping:
- automatic
- manual
- automatic and manual

Additional steps:
- lines straightening (dewarping)
- noise removal (denoising)
- contrast enhancement
- correction of geometric distortion
- binarisation
- removal of stamps, written notes
- equalising the dimensions of the scans (all same size after cropping)

Other notes: _____

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:
- English
- other (add): _____

Does OCR software use machine learning?
- yes
- no

OCR: page segmentation – recognition of different elements.
Layout segments are classified, either coarse (text, separator, image, table, ...) or fine-grained (paragraph, heading, ...).
- only automatic recognition
- additional manual corrections
- we do not use it

Other notes: _____

OCR: additional work on page segmentation – layout elements. We mark:
- marking paragraphs
- marking columns
- marking headers
- marking images
- marking background images
- marking table

Other notes: _____

(*continued*)

OCR: additional work on recognised text
- fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)
- no

Other notes: _____

4. ADDITIONAL PROCESSING
Did you use any other tools, software to enhance the quality of the results? For example: marking the final PDF with semantic tags or any other solutions. _____

5. EXPORT
Any additional comments? _____

Corresponding author
Alenka Kavčič Čolić can be contacted at: alenka.kavcic@nuk.uni-lj.si