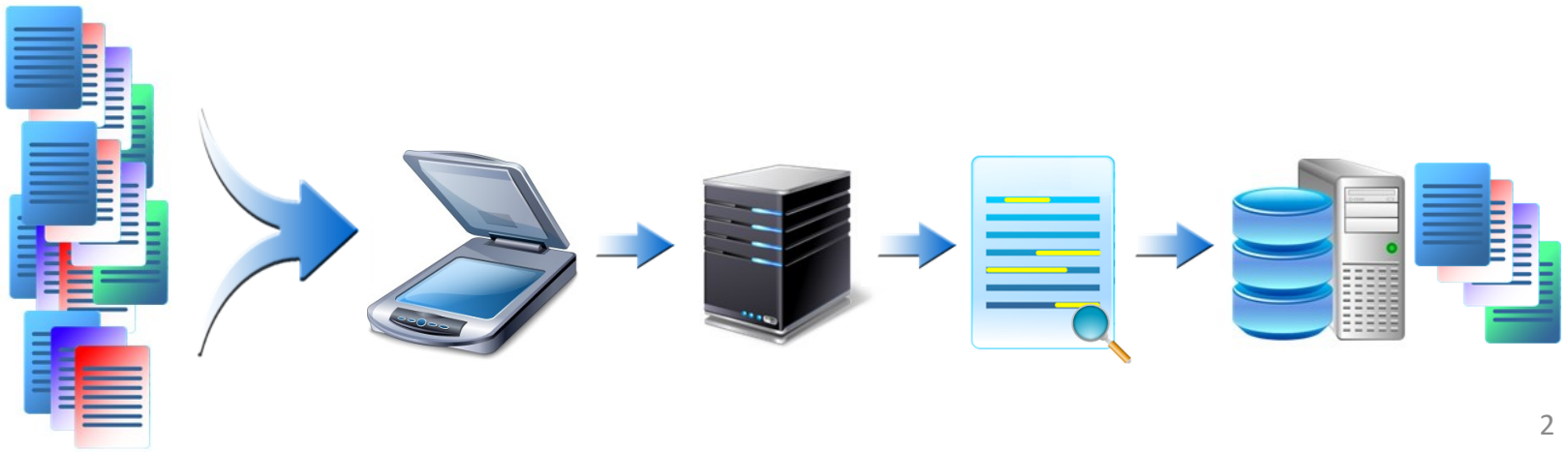
Evaluating and Improving OCR Efficiency

Dr. **Hrvoje Stančić**, full professor
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
hstancic@ffzg.hr

Željko Trbušić, mag. inf.
Croatian Academy of Sciences and Arts, Zagreb, Croatia
ztrbusic@hazu.hr

Contents

1. Introduction
2. Methodology
    - 2.1. Dataset extraction
    - 2.2. GT and stopwords preparation
    - 2.3. Accuracy measurement method selection
    - 2.4. Results comparison and analysis
3. Research results
4. Discussion
5. Conclusion

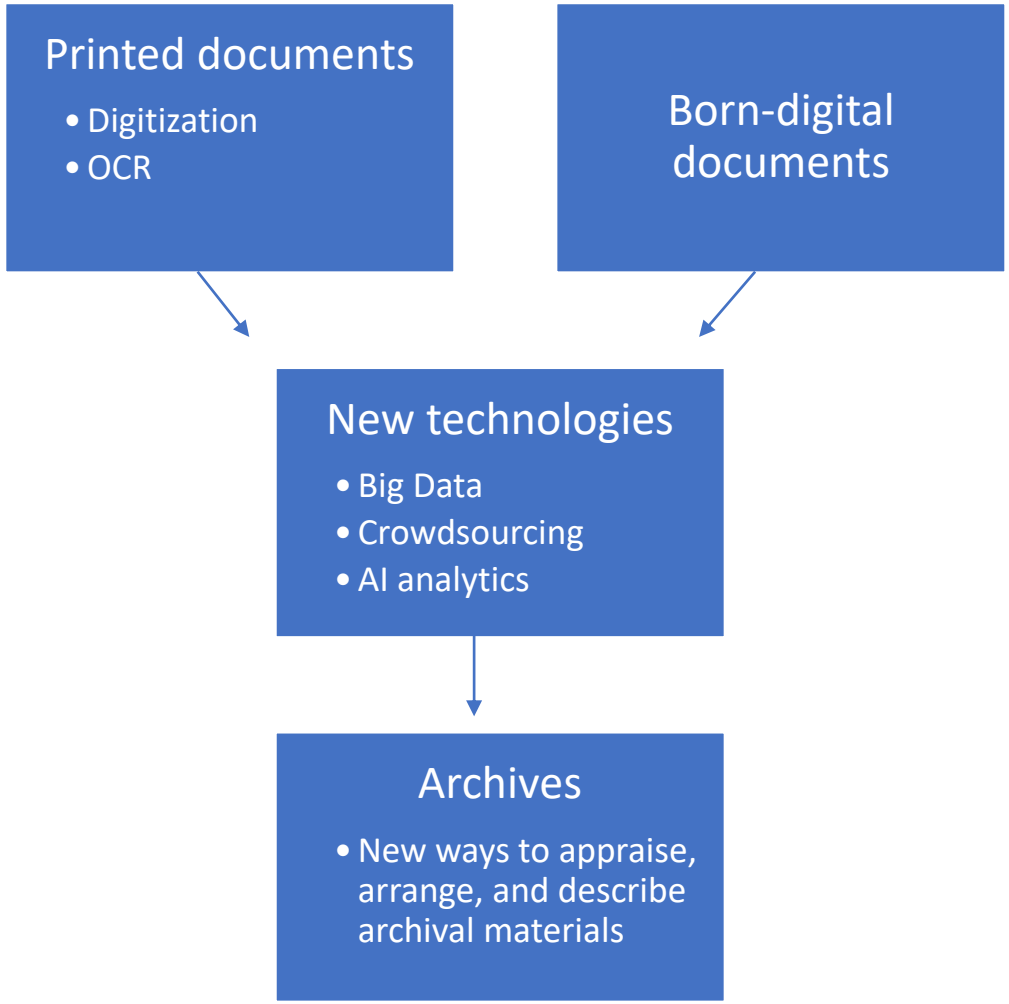


1. Introduction

- Archives = large amounts of (mostly textual) data
- Digitization enables the use of computing power to control this (big) data
- Optical Character Recognition (OCR) helps automating the process of data collection
- Research goal: optimization of OCR process implementation with minimal loss of data

Optical Character Recognition (OCR)

- Technology that enables transfer of printed text into computer encoded format
- OCR can be applied on any images that contain text, e.g. traffic and street signs, license plates, etc.
- Use of OCR in archives – production of formatted documents (TXT, DOC, PDF, etc.) that can be digitally edited, copied and searched **without the need for manual transcription**



Implementing OCR in archival processes

1. Image capture and pre-processing

- acquiring digital reproductions (optical scanner or digital camera)
- image-editing tools (commercial or open-source)
- **H1: The higher the resolution the better the recognition rate of the OCR**

2. OCR

- choosing the OCR engine (commercial or open-source)
- **H2: Commercial software offers higher accuracy than open-source OCR engines**

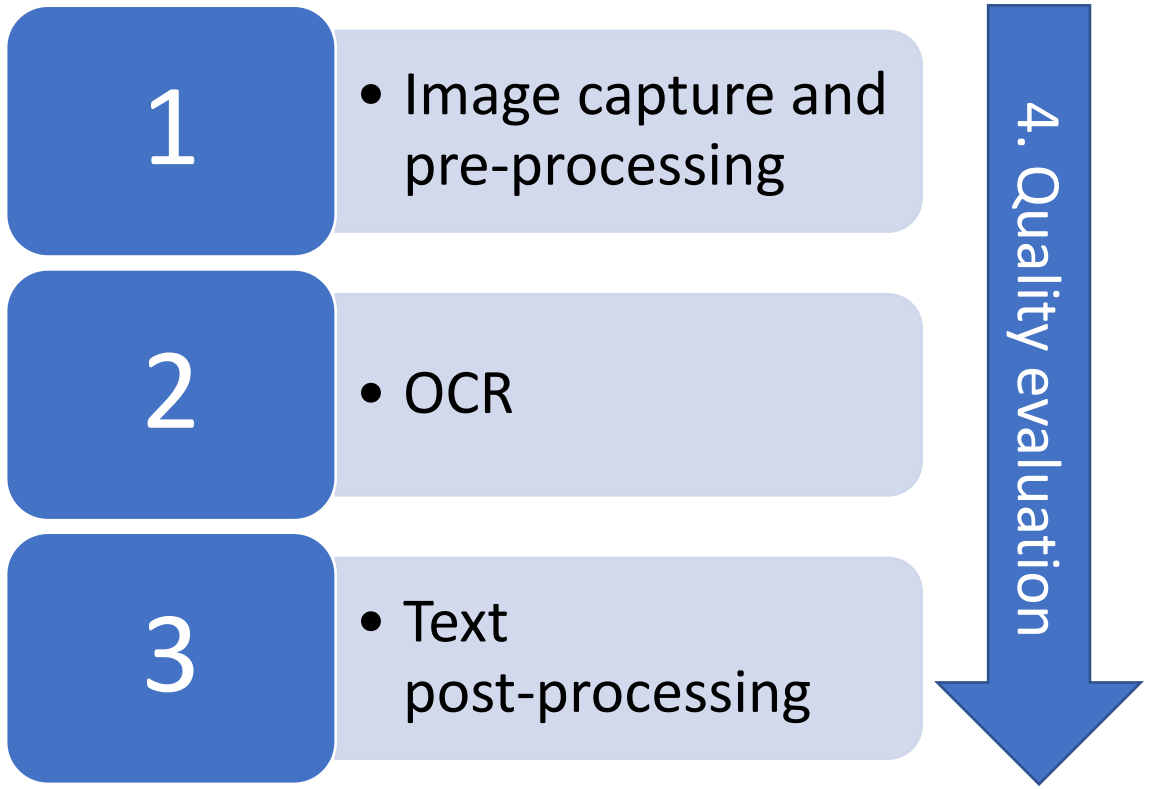
Implementing OCR in archival processes …

3. Text post-processing

- text clean-up and corrections (dictionary support, crowdsourcing)
- text formatting and arrangement (ALTO XML, hOCR, PDF, etc.)

4. Quality evaluation

- measuring OCR accuracy and effectiveness in the workflow
- can be applied to any part of the first three stages

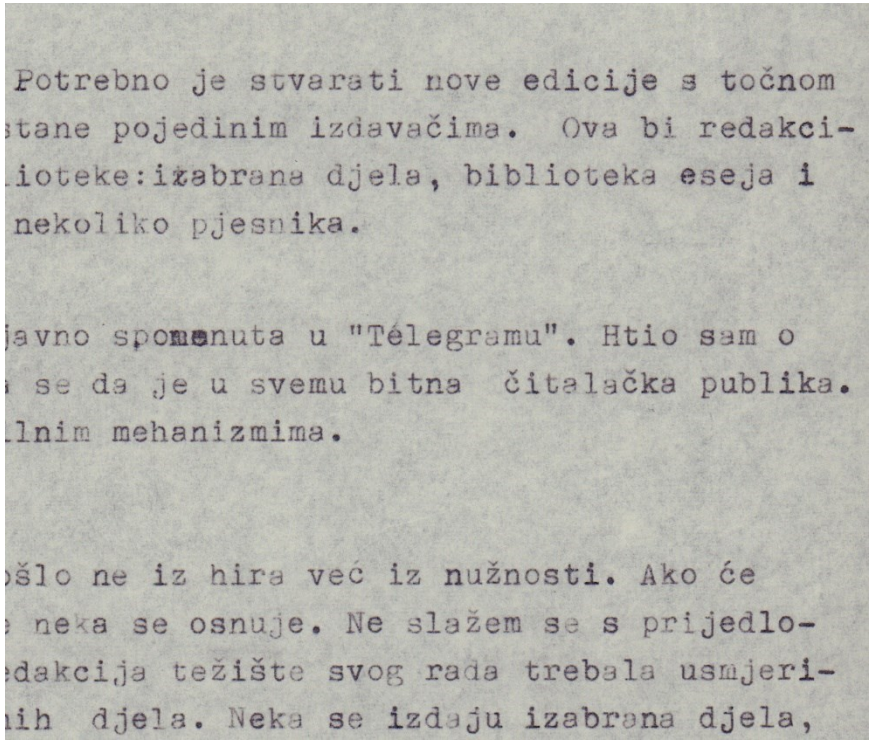


2. Methodology

- Research was conducted evaluating two different OCR engines' performance on a specific archival material
- Four stages:
    1. Dataset extraction
    2. GT and stopwords preparation
    3. Accuracy measurement method selection
    4. Results comparison and analysis

1. Dataset extraction

- 123 typewritten pages
  - transcripts of board meetings and plenary sessions of Croatian Writers' Society
  - 1966-1968
- Scanned at 6 different resolutions (100 - 600 dpi)
- Recognized using 2 different engines
  - Abbyy FineReader 15 release 4
  - Tesseract 4
- Total: 1,476 data units


Potrebno je stvarati nove edicije s točnom
stane pojedinim izdavačima. Ova bi redakci-
lioteke:izabrana djela, biblioteka eseja i
nekoliko pjesnika.

javno spomenuta u "Télegramu". Htio sam o
se da je u svemu bitna čitalačka publika.
lnim mehanizmima.

šlo ne iz hira već iz nužnosti. Ako će
neka se osnuje. Ne slažem se s prijedlo-
dakcija težište svog rada trebala usmjeri-
ih djela. Neka se izdaju izabrana djela,

A sample of typewritten text included in the dataset

2. GT and stopwords preparation

- Ground Truth (GT) – 100% correctness of the transcribed text
  - used to make comparison and accuracy calculations
  - manual input for every page in the dataset (123 pages)
  - illegible or overtyped characters were marked with a tilde (~) that serves as a wildcard
- Generating GT automatically (dictionary)
  - possibility of avoiding time-consuming manual input
  - it was not used - false positives and false negatives
  - can also be tested

2. GT and stopwords preparation

- Stopwords
  - contain less relevant data for searching and indexing
  - universal list does not exist
  - common stopwords in English include: *this*, *that*, *the*
  - Used list contained 1,198 unique Croatian stopwords

```
1  a ah aha aj ako al ali arh au
   avaj bar baš bez bi bih bijah
   bijahu bijaše bijasmo bijaste
   bila bile bili bilo bio biše
   bismo biste biva bivaju bivajući
   bivam bivamo bivaš bivate
   bivavši blizu brr buć budavši
   bude budem budemo budeš budete
   budi budimo budite budu budući bum
```

A sample of stopwords used in the accuracy measurement procedure

3. Accuracy measurement method selection

- The ISRI Analytic Tools for OCR Evaluation
  - developed in the Information Science Research Institute, Las Vegas, Nevada in the early 90's
  - open-source since 2005
  - most comprehensive OCR evaluation tools to date
  - 17 accuracy measurement and analysis programs
- This research used 4 different accuracy measurement methods
  - character accuracy
  - word accuracy
  - distinct non-stopwords precision
  - distinct non-stopwords recall

3. Accuracy measurement method selection …

- Character accuracy
  - provided by the ***accuracy*** program of The ISRI Tools
  - calculates the necessary deletions, substitutions and insertions (Levenshtein distance) of characters needed to correct the recognized text

- Word accuracy
  - ***wordacc*** program
  - calculates the percentage of misrecognized words
  - delivers the statistical data on stopwords, non-stopwords and **distinct non-stopwords**

```
UNLV-ISRI OCR Word Accuracy Report Version 6.1
------------------------------------------------

     301   Words
      17   Misrecognized
  94.35%   Accuracy

Stopwords
   Count    Missed    %Right    Length
      31         1     96.77         1
      41         2     95.12         2
      10         0    100.00         3
      22         2     90.91         4
      11         0    100.00         5
       2         0    100.00         7
       1         0    100.00         8
     118         5     95.76     Total

Non-stopwords
   Count    Missed    %Right    Length
       9         4     55.56         1
       3         2     33.33         2
       7         1     85.71         3
       8         0    100.00         4
      23         0    100.00         5
      17         0    100.00         6
      14         0    100.00         7
      22         2     90.91         8
      37         0    100.00         9
      20         2     90.00        10
      12         0    100.00        11
       8         1     87.50        12
       1         0    100.00        13
       2         0    100.00        14
     183        12     93.44     Total

Distinct Non-stopwords
   Count    Missed    %Right    Occurs
     137         8     94.16         1
      10         1     90.00         2
       6         0    100.00         3
       1         0    100.00         8
     154         9     94.16     Total
```

3. Accuracy measurement method selection …

- Distinct non-stopwords
  - relevant words whose occurrence in the text is low
  - counterargument to the statement that the full-text searchable documents are resilient to the OCR errors because of the redundancy of text

36	Distinct Non-stopwords			
37	Count	Missed	%Right	Occurs
38	137	8	94.16	1
39	10	1	90.00	2
40	6	0	100.00	3
41	1	0	100.00	8
42	154	9	94.16	Total

- Precision and Recall
  - distinct non-stopwords = relevant data (can be argued further)
  - retrieved data = total correctly recognized words
  - **Precision: percentage of retrieved data that is relevant to the query**
  - **Recall: percentage of successfully retrieved relevant data**

$$Precision = \frac{\text{correct distinct non-stopwords}}{\text{correct total words}} \; x\; 100\%$$

$$Recall = \frac{\text{correct distinct non-stopwords}}{\text{total distinct non-stopwords}} \; x\; 100\%$$

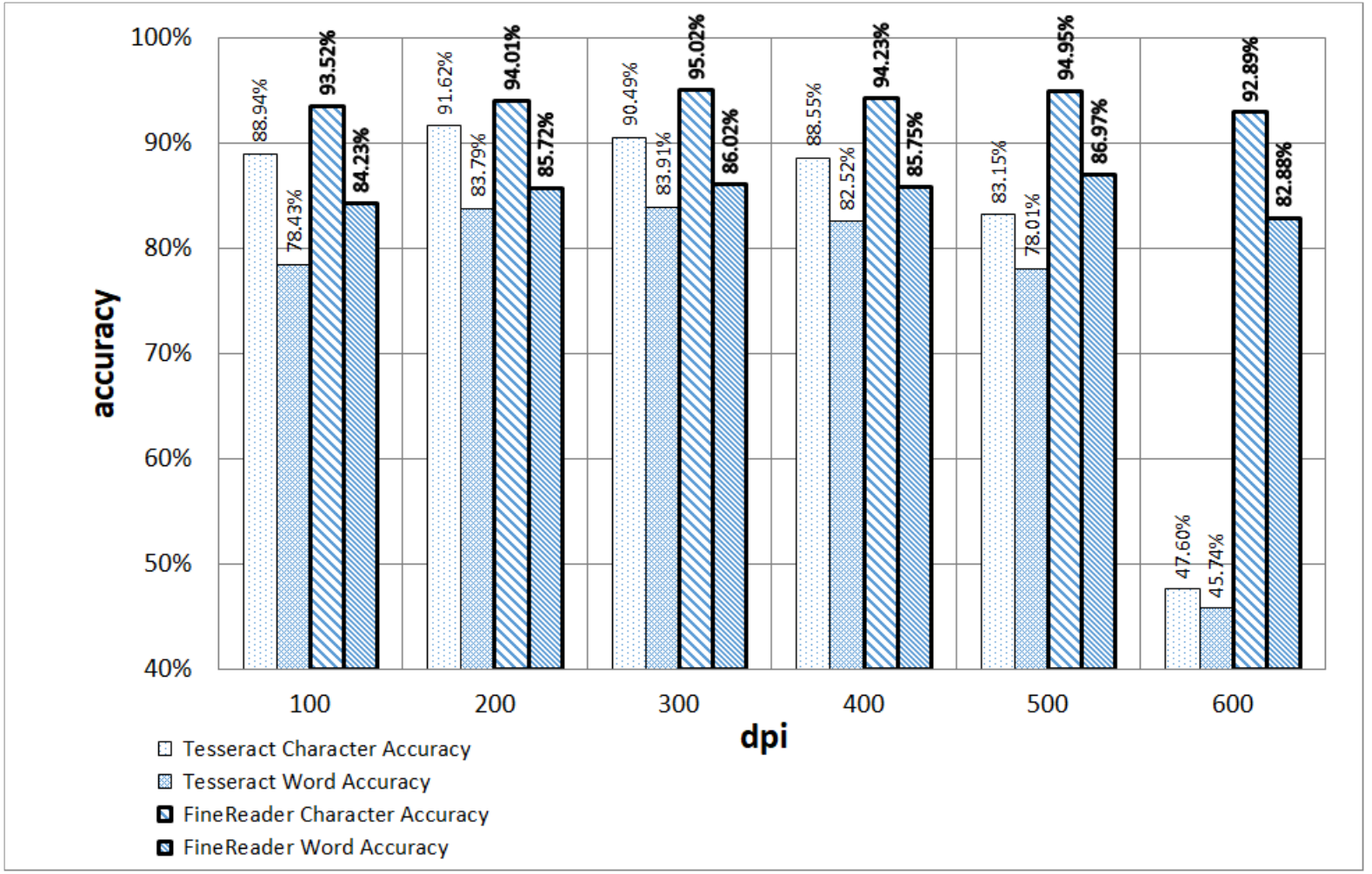
4. Results comparison and analysis

- Data collection
- Results division and comparison
- Development of new methods for OCR process on archival materials

3. Research results

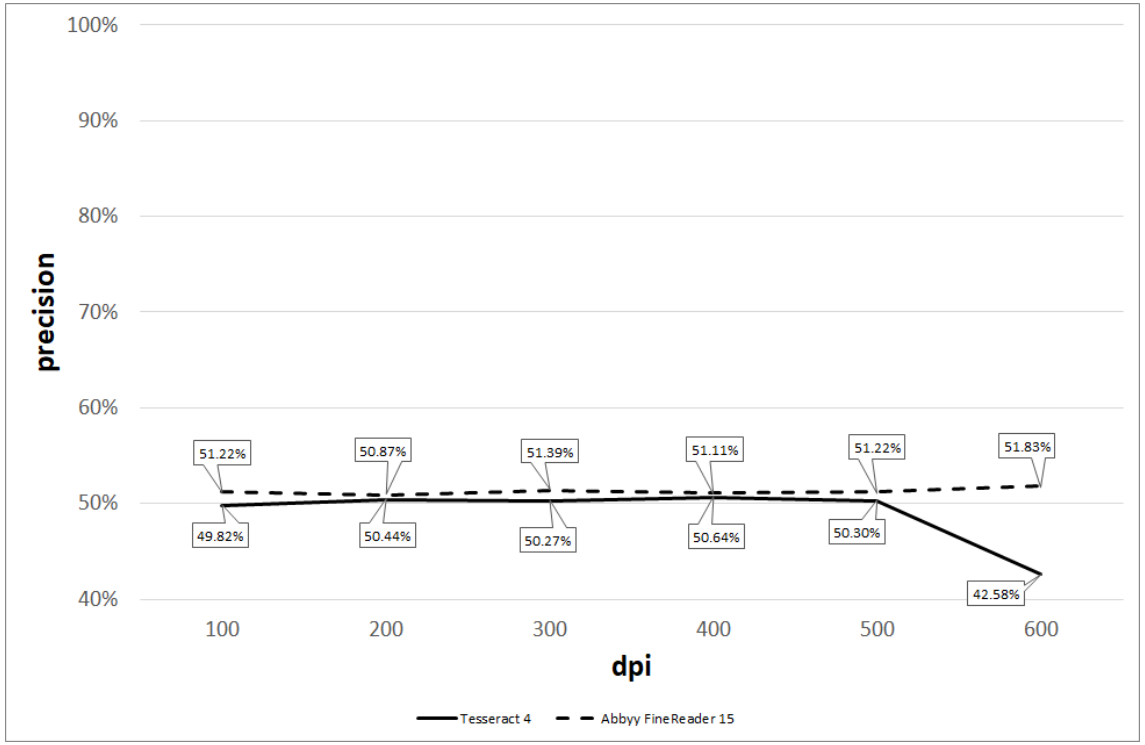
- Examining the usefulness of the testing for
  - faster and cheaper digitization process
  - constant quality level of digitized materials
  - stable OCR accuracy

Research results

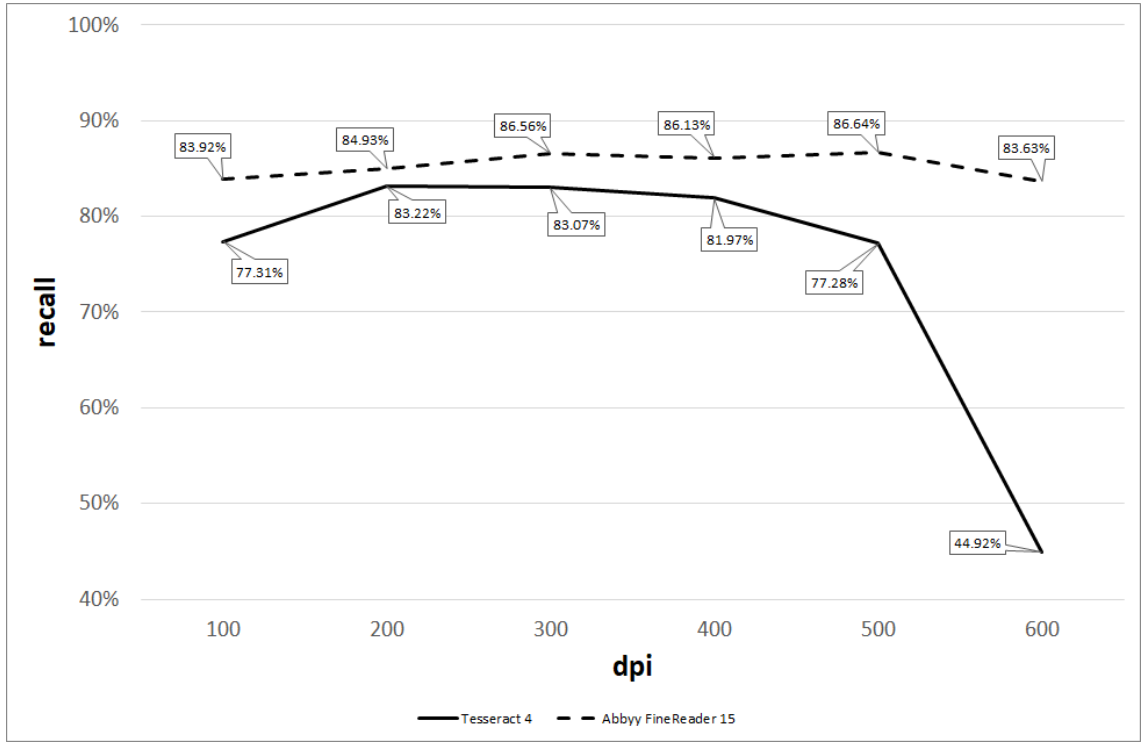


Graph 1. Accuracy of Tesseract 4 and Abbyy FineReader 15 on a 123-page sample of typewritten documents measured at six different quality levels.

Research results



Graph 2. Precision of distinct non-stopwords shown across different quality levels and OCR engines.



Graph 3. Recall of distinct non-stopwords shown across different quality levels and OCR engines.

4. Discussion

- **Precision**, as a distinct non-stopwords accuracy measure, didn't prove useful
  - it applies only to correctly recognized words
  - its usefulness is much higher for character accuracy – detection of noise in the OCR process

- **Recall** of distinct non-stopwords
  - probably the most useful measuring device for obtaining meaningful full-text search results
  - distinct non-stopwords = data that must not be lost during the OCR process
  - 90% non-stopwords accuracy level – 100% distinct non-stopwords accuracy – the document **will** be retrieved if any search query is conducted – the 10% of text data is redundant

Discussion …

- Accuracy rates in combination with expected quality levels of digitized document can **save resources**
- The research showed that it is not necessary to digitize documents using the best possible resolution to achieve the best accuracy rates **(H1 not confirmed)**

- Proprietary engine performs better and delivers more consistent results across all tested variables (**H2 confirmed**)
- Open-source engine can be trained further, and more implementation options are available

5. Conclusion

- The **goal** of OCR in archives is not 100% accurate transcriptions but **minimal loss of data**
- The process should be **automated** as much as possible
- **Recognition rate** information – signal the user to read through the retrieved document and not to rely solely on full-text search
- Evaluation of the process
  → setting up the workflow according to **optimal requirements** (accuracy, time, cost)
  → more digitized pages with **optimal recognition rates**
  → improved **search results** and **user experience**
  → developed **TRUST** in the archives and digital collections!

THANK YOU!

Evaluating and Improving OCR Efficiency

Dr. **Hrvoje Stančić**, full professor
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
hstancic@ffzg.hr

Željko Trbušić, mag. inf.
Croatian Academy of Sciences and Arts, Zagreb, Croatia
ztrbusic@hazu.hr