Co-funded by the
Creative Europe Programme
of the European Union

REPORT

Project Acronym:	**EODOPEN**
Grant Agreement Number:	**607666-CREA-1-2019-1-AT-CULT-COOP2**
Project Title:	**EODOPEN | eBooks-On-Demand-Network** **Opening Publications for European Netizens**
Project Website:	**https://eodopen.eu/**

A12 Evaluating Delivery Formats and Solutions

D12b&c Report on Trial Implementations for Mobile Devices and Print-Disabled Users

Author(s):

Andreja Hari

Alenka Kavčič Čolić

Project Co-funded by the Creative Europe Programme of the European Union		
Dissemination Level		
P	**Public**	**X**
C	**Confidential, only for members of the project partners and the Commission Services**	

Co-funded by the
Creative Europe Programme
of the European Union

DOCUMENT INFORMATION

Activity number:	A12
Activity title:	Evaluating Delivery Formats and Solutions
Contractual date of activity:	31.10.23
Actual date of activity:	01.01.22–31.03.23
Author(s):	Andreja Hari, Alenka Kavčič Čolić (NUK)
Contributor(s):	Constantin Lehenmeier (UREG), Tina Glavič (NUK)
Participant(s):	*EODOPEN-project members*
Working group:	WG4
Working group title:	Delivery Formats of Digitised Material for Special Needs
Working group leader:	Alenka Kavčič Čolić
Dissemination Level:	P

HISTORY OF VERSIONS

Version	Date	Status	Author (organisation)	Description/Approval Level
1	27.07.22–09.08.22	Draft	Andreja Hari (NUK) Alenka Kavčič Čolić (NUK) [Tina Glavič (NUK) co-author of the methodological approach]	First draft
2	21.11.22	Draft	Andreja Hari (NUK) Alenka Kavčič Čolić (NUK)	Sent to partners for additional changes. Corrections and suggestions received from Constantin Lehenmeier (UREG).
3	09.03.23	Draft	Andreja Hari (NUK) Alenka Kavčič Čolić (NUK)	Final draft
4	07.06.23	Draft	Proofreading completed	Final draft
5	19.06.23	Draft	Delivered for peer review to project partners	Final draft
6	29.09.23	Final	Edited after review and finalised	Completed

EODOPEN PROJECT SUMMARY

Libraries all over Europe face the difficult challenge of managing tremendous amounts of 20th and 21st century textual material that has not yet been digitised due to the complex copyright situation. These works cannot be accessed by the general public and are hidden deep in library stacks, as they are often out of print or have never even been printed at all, while reprints or facsimiles are out of sight.

The **EODOPEN** project focuses on making 20th and 21st century library collections digitally visible by **directly engaging with communities** in the selection, digitisation and dissemination processes. As a leading partner, the University Library of Innsbruck, joined by 14 European libraries from 11 nations, has set itself the goal of making 15,000 pieces of textual material digitally available, and of reaching more than one million people in Europe by 2024.

Among other goals, such as building a common portal to display the project outcomes, EODOPEN aims to stimulate interest in and improve access to 20th and 21st century textual material, including grey and scientific literature. EODOPEN continuously carries out social media campaigns in order to attract new audiences. Furthermore, the participating libraries establish contact with commemorative institutions all over Europe, as well as with researchers and doctoral study boards, history associations and local publishing houses, in order to obtain suggestions from a broad audience.

In collaboration with local institutions, all of the project partners select hidden library treasures, **deal with rights clearance questions** and put new content online, while dissemination activities display the digital content via international channels.

In addition, EODOPEN aims to provide alternative delivery formats suitable **for blind or visually impaired users.** An international survey gathers data from a broad European public about the use of e-books. By evaluating this data, the project broadens its scope to alternative delivery formats in order to fulfil the needs of **blind or visually impaired users.**

In order to promote best practice in rights clearance among the library community, EODOPEN provides handouts and tools to make 20th and 21st century books available beyond the project's lifetime. In this regard, the project partners cooperate closely to develop an online tool for the documentation of rights clearance, especially suited for out-of-print and orphan works. Interactive workshops investigate needs related to **dealing with rights clearance** questions in order to implement the requirements of the international community in establishing the online tool.

ABSTRACT

The aim of the Report on Trial Implementations for Mobile Devices and Print-Disabled Users (hereinafter: the Report) is to help libraries and other cultural organisations to make digitised content available to a broader community. The Report is based on EODOPEN partners' digitisation experiences at their organisations and complements the EODOPEN Project Deliverable 11: *Guidelines and Recommendations for the Provision of Alternative and Special Formats*, which addresses delivery formats and criteria for increasing the quality of digitisation results for users of mobile devices and blind and partially sighted users. The Report presents the results of a trial implementation among EODOPEN partners on their digitisation workflows, the delivery file formats used and, consequently, the quality of optical character recognition (OCR) results, depending on file format type and accessibility criteria.

Statement of originality:
This report contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

DOCUMENT INFORMATION **2**
HISTORY OF VERSIONS **3**
EODOPEN PROJECT SUMMARY **4**
ABSTRACT **5**
TABLE OF CONTENTS **6**
LIST OF TABLES **8**
1. Introduction **9**
1.1. Purpose *9*
1.2 Description of the Report *9*
1.3 Explanation of the key concepts *9*
1.4 Context description: User needs for delivery formats *10*
2 Evaluation of delivery formats: Trial implementation **12**
2.1 Background *12*
2.2 Methodological approach *13*
3 Test results **21**
3.1 Results by sample scan number *25*
3.2 Results according to criteria *36*
4 Test findings **45**
5 Possible solutions and recommendations **47**
5.1 Solutions for mobile devices *47*
5.2 Solutions for print-disabled users *48*
11. Summary **50**
12. References **52**
13. Vocabulary **53**
14. Used acronyms **54**
15. Annexes **55**
Annex 1. Testing samples *55*
Scan 1 56
Scan 2 57
Scan 3 58
Scan 4 59
Scan 5 60
Scan 6 61
Scan 7 62
Scan 8 63
Scan 9 64
Scan 10 65

Scan 11 66
Scan 12 67
Scan 13 68
Scan 14 69
Scan 15 70
Scan 16 71
Annex 2: Testing report questionnaire 72
Annex 3. Testing report questionnaires by partners 74
P1 - UIBK, University of Innsbruck – RTF format 75
P1 - UIBK, University of Innsbruck – ODM workflow – PDF and RTF format....77
P2 - UT, University of Tartu....79
P3 - NUK, National and University Library – usual workflow 81
P3 - NUK, National and University Library – PDF edited with Adobe Acrobat Pro....83
P3 - NUK, National and University Library – PDF made with Abbyy FineReader 15 85
P3 - NUK, National and University Library – PDF and ePUB made from Word 87
P4 - MZK, Moravian Library – small edited....89
P4 - MZK, Moravian Library – edited 91
P5 - UG, University of Greifswald....93
P6 - NLS, National Library of Sweden....96
P7 - NCU, Nicolaus Copernicus University in Torun....98
P9 - VKOL, Research Library Olomouc 100
P10 - BNP, National Library of Portugal – PDF....102
P10 - BNP, National Library of Portugal – docx 104
P11 - NLE, National Library of Estonia 106
P12 - OSZK, National Széchényi Library 108
P13 - CVTI SR, Slovak Centre of Scientific and Technical Information....111
P14 - UREG, University of Regensburg....113

LIST OF TABLES

Table 1: The appearance of the criteria in the 16 scans of the test sample. 18
Table 2: Software overview for automatically generated outputs 21
Table 3: Software overview for outputs containing additional manual corrections 22
Table 4: Results of Scan 1 25
Table 5: Results of Scan 2 26
Table 6: Results of Scan 3 26
Table 7: Results of Scan 4 27
Table 8: Results of Scan 5 27
Table 9: Results of Scan 6 28
Table 10: Results of Scan 7 29
Table 11: Results of Scan 8 30
Table 12: Results of Scan 9 31
Table 13: Results of Scan 10 31
Table 14: Results of Scan 11 32
Table 15: Results of Scan 12 33
Table 16: Results of Scan 13 33
Table 17: Results of Scan 14 34
Table 18: Results of Scan 15 34
Table 19: Results of Scan 16 34
Table 20: Results according to criteria for the automatically generated outputs 36
Table 21: Results according to criteria for the outputs with additional manual corrections 37

1. Introduction

1.1. Purpose

The aim of the Report on Trial Implementations for Mobile Devices and Print-Disabled Users (hereinafter: the Report) is to help libraries and other cultural organisations in the field of culture to make digitised content available to a broader communities. The Report is based on EODOPEN partners' digitisation experiences at their organisations and complements the EODOPEN Project Deliverable 11: Guidelines and Recommendations for the Provision of Alternative and Special Formats, which is based on a survey on the special needs of users, and technical requirements concerning regarding delivery formats and criteria for increasing the quality of digitisation results for users of mobile devices, as well as for blind and partially sighted users. The Report gathers experiences from all EODOPEN consortia partners.

1.2 Description of the Report

The Report presents the results of a trial implementation among EODOPEN partners. It comprises a brief introduction followed by a description of the methodology and the test results, and it concludes with the findings and some recommendations.

The introductory chapter (Chapter 1) describes the Report's purpose, scope and key concepts used in the text, and defines the user needs for mobile users, as well as for blind and partially sighted users. Chapter 2 provides the background, the methodological approach, a description of the test sample, and a definition of the evaluation criteria of delivery formats. Chapter 3 presents partners' test results. This is followed by a discussion of the findings (Chapter 4) and some recommendations (Chapter 5). The Report is accompanied by a list of literature sources and recommended references, as well as a definition of the terms used and a list of acronyms. All of the samples and test report questionnaires are attached to this document in the annexes.

1.3 Explanation of the key concepts

In the Report, mobile devices are defined as smartphones, notebooks and tablet computers, as well as e-readers. In accordance with the recommendation of the European Blind Union (EBU), the term blind and partially sighted users is used instead of the term blind and visually impaired users. Although the term print disability covers a wide range of disabilities or problems related to reading text, the Report focuses solely on the blind and partially sighted, which is one of the project's primary groups. Digitisation means digital conversion of information on analogue carriers. Target communities are people who access digitised content in libraries and other cultural organisations. The term eBook usually refers to born-digital publications, but in this document it is used refer to digital publications produced as a result of digital conversion, including formats for special needs (audiobooks), which is one of the objectives of the EODOPEN project. However, this term does not exclude born-digital publications, as the delivery format is the same or has the same purpose or functions. eBooks can be accessible through e-readers or can simply be read on personal computers (PCs) or mobile devices such as smartphones, tablets or notebooks.

1.4 Context description: User needs for delivery formats

Mobile devices are integral tools of the global information society and smartphone technology is already part of our everyday life, enabling constant interconnection with other tools and people through various networks and social media. The development of mobile devices also has a strong impact on the development of their operating systems and tools, which is something that the service sector, including libraries and cultural organisations, should always bear in mind. It is therefore very important to plan and publish content in file formats that are and will continue to be supported by these devices.

eBooks can be read on different kinds of mobile devices, such as e-readers (Kindle, Kobo, Midia Inkbook, NOOK, etc.), smartphones, tablets and portable computers (notebooks). The selection of file format delivery and/or access depends on the type of device (size of screen, visual presentation) and the existing platform (Microsoft, Android and iOS are the most commonly used). Although there are no problems associated with accessing PDF files through devices with bigger screens, they not a recommended for smaller devices like smartphones or e-readers because PDF it is not a responsive file format. The most recommended formats for smaller devices are ePUB 3, AZW/MOBI, HTML, Microsoft Office Word documents (RTF, docx, etc.) or audio books (mp3, DAISY). Some platforms only support certain types of formats. For instance, Kindle e-readers did not originally support ePUB format, so users had to convert ePUB files to AZW/MOBI format before uploading them to their devices. Recently, however, due to new developments at Amazon, file format conversion to ePUB has been made available to users, while the obsolete MOBI format is no longer supported (Amazon, s.a.).

With regard to blind and partially sighted users, as well as for other print-disabled users, it is important to consider the degree to which the user can use his/her sight and its variation from day to day or due to light conditions, tiredness or stress levels, etc. It is therefore important to consider how to provide users with the possibility to adapt the visual presentation of the text to fit their needs. Some of the common challenges faced primarily by partially sighted people are difficulty in focusing on the text, reduced contrast sensitivity, reduced field of vision, sensitivity to movement, visual fatigue and similar. For the users, the most useful adjustments are adjustments in font size, font type, colour themes, margins and spacing. The option to access the full text (where optical character recognition (OCR) is preferably checked) is also important, as it enables the use of assistive technologies (such as braille display or screen readers). Although blind and partially sighted users mostly access documents through bigger screens, they are also avid users of smaller mobile devices. The most recommended formats for this group of users are Microsoft Office Word documents (RTF, docx, etc.), audio books (mp3, DAISY), HTML, ePUB 3 and AZW/MOBI, but tagged PDF format is also suitable: “PDF tags are the key to accessing a PDF document’s content with assistive technologies such as screen readers. When a tagged PDF is created, each page element in the document is ‘tagged’. Each tag identifies the type of content and stores some attributes about it. They also arrange the document content into a hierarchical architecture (or a ‘tag tree’). The tag tree forms the logical structure of the document (reading order).” (Accessible document solutions, n.d.)

For a more comprehensive overview of user needs for delivery formats, see Deliverable 11: *Guidelines and Recommendations for the Provision of Alternative and Special Formats*, which is based on a survey on the special needs of users and technical requirements.

2 Evaluation of delivery formats: Trial implementation

2.1 Background

A questionnaire survey conducted among EODOPEN project partners in 2020 revealed that libraries use various devices and software tools in the digitisation process. Digital conversion is automated or carried out in different phases and depends on financial resources as well as on adequately trained personnel. To ensure the best possible results, it combines a variety of technological and software solutions, resulting in a diverse range of digitised material. This material is available to users via digital libraries in various delivery formats, which also differ from each other in terms of the functionalities provided.

In the *Guidelines and Recommendations for the Provision of Alternative and Special Formats* (Deliverable D11), which were prepared within the framework of project's working group 4,[1] special emphasis is placed on the possibilities and ways of adapting digitised material to make it available in formats that ensure accessibility to blind and partially sighted users.

According to research, the appropriate structuring of a text and its elements is crucial for reading digital material, as it enables navigation through the text. Text navigation, recognition of text and graphic elements, and the ability to personalise settings are even more important for blind and partially sighted people, who use assistive technology and dedicated software in order to read. Optical character recognition (OCR) tools and their software modifications enable optical recognition of characters – letters, numbers, punctuation marks – as well as text structures. Machine learning technology has advanced to the point where errors in OCR are negligible. OCR recognition errors mainly occur when reading special characters, such as chemical formulas, mathematical operations and equations, although errors occur also in identifying headings, sub-headings and graphical elements in the text (images, graphs). Furthermore, in cases of more than one text column, the text flow is often not recognised correctly, with the linear sequence of the text appearing instead. If we assume the position of a blind person who uses speech synthesis to read, a text is unreadable without the proper interpretation of special characters, the specific sequence of the text, and the graphic elements with their corresponding descriptions.

Perception, operability, understanding and robustness – defined by the World Wide Web Consortium (W3C) through the Web Accessibility Initiative (WAI), as part of the Web Content Accessibility Guidelines (WCAG) – are the umbrella criteria for making websites and digital material accessible to blind and partially sighted users, as well as other groups of users with disabilities. Within the framework of the aforementioned working group, we sought to

[1] The working group is led by the National and University Library (Slovenia).

approach these accessibility criteria, which also apply to born-digital material or e-books. The objectives of the working group were:

- to develop a common test sample (a selection of scans), including as many different textual and graphic elements as possible;
- to test the sample in the further digitisation process by all partner institutions;
- to create representative samples based on the test sample, using various tools and attempting to meet the criteria of the Web Content Accessibility Guidelines (WCAG) in one case;
- to compare all of the received results based on set accessibility criteria and thus identify the most appropriate solutions;
- to obtain more detailed information on digitisation workflows in partner institutions;
- to identify workflows and digitisation phases that allow segmentation and identification of textual and graphical elements with all of their properties.

The purpose of the testing was to identify the best solutions in the digitisation process, and to determine whether there is any further room for improvement in the provision of digitised materials that meet accessibility criteria. This would allow libraries to review and, depending on the resources provided, improve digitisation workflows and user services.

As mentioned above, sighted people do not need such precise processing of texts to be able to access the content of digitised works. Nonetheless, responsive technologies are also based on accurate OCR and sighted people also use screen readers that enable text to be read aloud to them. The entire testing was therefore based on criteria that are essential for the blind and partially sighted, thus following the principle of universal design (for everyone). In Chapter 5 of this document, possible solutions and recommendations are presented both for mobile device users and print-disabled users, in case libraries want to focus on just one group of users. The use of solutions for print-disabled users is, however, recommended.

2.2 Methodological approach

The test phase was conducted between February and July 2022 at all of the partner institutions. EODOPEN partners received a test sample (see Annex 1) and a blank test report questionnaire (see Annex 2) on which they reported the work done with the test sample.

The aim of the testing was to find out which scanning and recognition workflows are optimal for achieving the best results in OCR, as well as to determine which file formats can be generated, as different file formats can provide users with different user experiences. For this purpose, it was decided that all of the partners would test the same samples containing English text, as none of the EODOPEN partners are located in regions where English is the native language. Using the same scans with English text would facilitate the comparative analysis of the results. In addition, some OCR tools are better adapted to majority languages (e.g., German), and we wanted to avoid discrimination of minority languages such as Slovenian, Estonian or Slovak. Partners could subsequently conduct the same analysis on scans in their own language for additional testing of their systems.

The test sample consisted of 16 scans in TIFF format (see Annex 1), comprising both textual and non-textual elements, such as plain text, chapters and sub-chapters, columns, tables, footnotes, flowcharts, images and text accompanying images (captions). The special examples in the texts were chemical formulas, mathematical equations and special characters (£, °C, etc.). Two of the scans contained two pages on one: in the first scan, the title and the name of the author were spread across both pages, while the second scan contained the chemical periodical table spread across both pages. Most of the scans had a complicated structure with elements that could disturb the text order (e.g., captions) or create problems with element recognition (e.g., tables). Only three of the scans (8, 9, 13) had basic layout with text in one column and a picture, which would not be expected to cause difficulties with regard to reading order.

The partners used the test sample in their usual digitisation workflows, conducting the process from scan processing to the creation of the most common delivery format available in their digital library. The results were returned to the testing team at the National and University Library (Slovenia).

The test report questionnaire (see Annex 2) consisted of 14 questions enabling the project partners to record the work processes, software tools and solutions used when testing the sample. Reviewing the reports enabled us to learn more about the different stages of the digitisation workflow: scan import, image processing, OCR options (multilevel document analysis and recognition of elements), any additional processing, and exporting the final delivery format. As digitisation processes are diverse, the questionnaire provides a framework enabling us to gain an insight into the workflows of the individual institutions, especially with regard to the stages and levels of the digitisation process that lead to meeting the WCAG criteria, or that bring better digitisation outputs for the end users.

For the evaluation of the outputs, **24 criteria** were prepared based on WCAG for the optimal accessibility of the documents and other best practice guidelines, focusing primarily on accessibility for blind and partially sighted. The criteria were established separately for each scan, as they were not all applicable to all of the scans. Moreover, some of the criteria were specific to individual scans, as they can produce different results during the OCR process (e.g., page rotation and pagination – double). The criteria used to evaluate each output of the digitisation process were:

1. **ALT-TEXT PICTURE** – Alt-text or alternative text for pictures provides a textual description for non-text 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.
2. **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 (see Scan 2 in Annex 1).
3. **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.[2] This criterion is primarily important for the blind and partially sighted.
4. **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.[3] This criterion is mainly important for the blind and partially sighted.
5. **HEADING 1** – Mainly for navigational purposes, the headings of the chapters should be marked and structured in depth (Heading 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.
6. **HEADING 2** – See criterion Heading 1
7. **HEADING 3** – See criterion Heading 1
8. **INITIAL** – A larger first letter at the beginning of a chapter is often not recognised or not recognised correctly (see Scan 7 in Annex 1). This criterion is important for all users.
9. **LANGUAGE SEGMENTS** – See 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

[2] 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.

[3] Good results can be achieved, for example, in Microsoft Word, HTML or ePUB by providing two-way hyperlinks.

proper pronunciation.[4] This criterion is primarily important for the blind and partially sighted, but could also be useful to sighted users using speech synthesis.

10. **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.
11. **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 recognised 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.
12. **OCR ERRORS (TEXT IN PICTURE 4)** – One image showed text written on a tombstone (see Scan 7 in Annex 1). Ideally, text of this kind would not be recognised, but the goal was to see what kind of results would be obtained. This criterion is important for all users.
13. **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.
14. **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.
15. **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.
16. **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.

[4] For example, a German text that is read aloud with an English voice sounds strange.

17. **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.
18. **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.
19. **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 recognising the first two characters. In the scan with decimal numbers, the numbers are written with an apostrophe (‘), which the English vocabulary fails to recognise 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.
20. **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.
21. **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.[5] This criterion is primarily important for the blind and partially sighted, but could also be useful to sighted users using speech synthesis.
22. **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.
23. **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 recognise 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.

[5] 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.

24. **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 recognised elements.[6] 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.

Table 1 provides an overview of the appearance of these 24 criteria in the test samples by scan number.

Table 1: The appearance of the criteria in the 16 scans of the test sample.

Criteria\Scan	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**=**
alt-text picture	1	1	1	1	3	2	4					1	1	3			18
alt-text picture (chem. formula)		2															2
caption	1	1	1	1	3	2	4			1	2	1	1			1	19
footnotes												1					1
heading 1	1				1	1	1		1				1		1		7
heading 2	1	1	2			4			1							1	10
heading 3									1								1
initial							1										1
language segments						1	1			1	1	1	1				6
math (simple)	1	1										1					3
math (advanced)	1	1			1							1					4
OCR errors (text in picture 4)							1										1
page rotation										1							1
pagination	1	1		1	1		1	1	1	1	1	1	1			1	12
pagination–double				1			1										2
picture	1	1	1	1	3	2	4					1	1	3			18
picture (chem. formula)		2															2
primary language	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16

[6] The most frequently used software for OCR – Abbyy FineReader desktop version – has this option during processing the digitised content. For post-processing, an example of software of this kind is Adobe Acrobat Pro.

Criteria\Scan	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	=
special character	1											1	1				3
stamp removal			1														1
table										1	2					1	4
table header										1	2					1	4
table rows										1	2					1	4
text order	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	16
=	11	13	8	7	14	14	20	3	6	9	12	11	9	8	3	8	

Three levels were used to evaluate the set criteria:

- *criterion was fully achieved* (A): used if the technical and content part of the criterion was achieved. Example: table rows were technically correct (each row contained the right number of rows and the correct content).
- *criterion was partly achieved* (B): used if either the technical or the content part of the criterion was achieved, but not both, or if there was a very minor mistake in the criterion. Example 1: alt-text is technically correct, but the content is either the text of the caption or other surrounding text. Example 2: there was a minor mistake in the text order.
- *criterion was not achieved* (blank cell): used if neither the technical nor the content part of the criterion was achieved. Example: pagination was present, but was not the first element on the page.

The evaluation was undertaken using various software and tools according to different output formats:

- DROID – with this software, the versions of the format (PDF 1.4 or RTF 1.9) were determined;
- Adobe Acrobat Reader Pro – with this software, the content of PDFs was checked, as well as the reading order and structure (tags) when the PDF was tagged;
- PDF Accessibility Checker 2021 – with this software, we checked what kind of errors were found in the PDF file according to the standards and whether the language of the document was set; when the PDF was tagged, the reading order and structure (tags) were also checked;
- Thorium Reader – with this software, we checked the content of an ePUB file and determined which options it enables with regard to visual adjustments and navigation (if a table of contents was available within the software);
- Sigil – with this software, the content of an ePUB file was checked, as well as the reading order and structure (HTML tags);
- Microsoft Office Word – with this software, the content of docx and RTF files was checked;
- Notepad – with this software, the content of TXT files was checked;

- Windows Narrator – this speech synthesis software was used only in special cases to check how the content is provided to the user.

3 Test results

A total of 23 test outputs were received from 13 partner institutions. These include automatically generated outputs (17), as well as outputs containing additional manual corrections (6). The software packages used for testing the samples were: ABBYY FineReader, ABBYY FineReader 11, ABBY Recognition server 4, ABBY Recognition server 14, ScanGate by Treventus Mechatronics, ABBYY FineReader PDF 15 Standard, Abbyy Finereader 15 desktop version, Adobe Acrobat Pro, IRIS OCR, LIMB processing, Microsoft Office Word, Scan Tailor Advanced v1.01.16, Tesseract 5.0.0-beta-20210815-22-g386dd, Photoshop 23.2.2., Project PERO OCR and WordToEpub (refer to Table 2).

Table 2: Software overview for automatically generated outputs

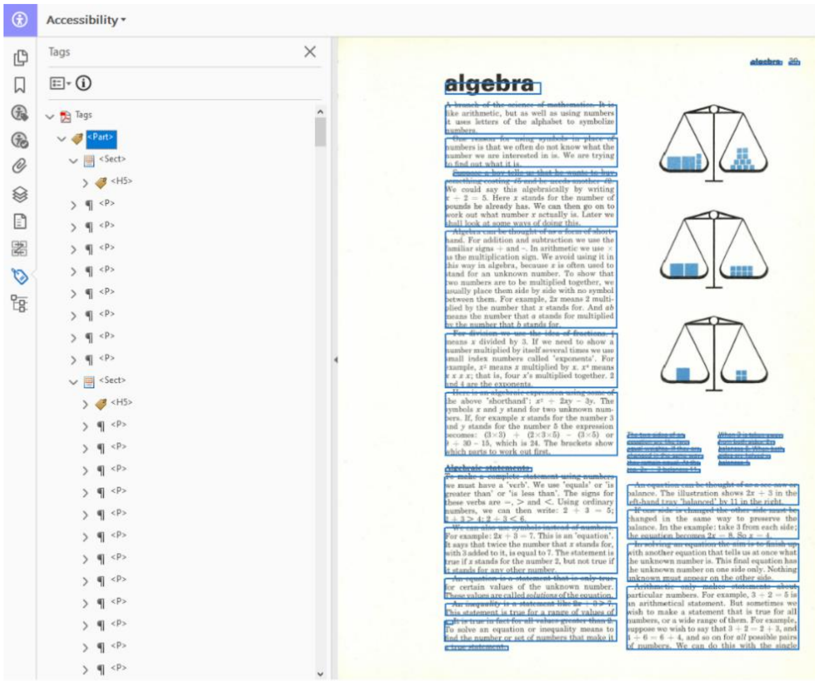
No.	PARTNER	SOFTWARE USED	GENERATED FORMATS
1	UIBK	ODM - Abbyy FineReader recognition server 4	PDF
2	UIBK	ODM - Abbyy FineReader recognition server 4	PDF/A
3	UIBK	ODM - Abbyy FineReader recognition server 4	RTF
4	UT	ABBYY FineReader PDF 15 Standard; ABBYY FineReader Server 14	PDF
5	NUK	Abbyy FineReader	PDF
6	MZK – small edited	Project Pero OCR	Page and Alto format (+TXT with plain text) – TXT tested
7	MZK - edited	Project Pero OCR	Page and Alto format (+TXT with plain text) – TXT tested
8	UG	Abbyy FineReader	PDF
9	UG	Abbyy FineReader	EPUB
10	NLS	Abbyy FineReader 11, Limb Processing	PDF
11	NCU	Abbyy FineReader Server 14.0	PDF/UA
12	VKOL	ScanTailor Advanced v1.01.16, Tesseract 5.0.0-beta-20210815-22-g386dd	xml + PDF (no OCR) and txt link shared to digital library – tested TXT
13	BNP	LIMB Processing, IRIS OCR	PDF
14	NLE	For books files: Abbyy FineReader 11, Abbyy Recognition Server 4. For newspaper/periodicals: Abbyy FineReader 11, CCS docWorks 7.1.0.90, Abbyy FineReader 12 OCR-engine	PDF/A
15	OSZK	Scans: 1-6, 8-11, 14-15: ScanTailor Advanced (1.0.16), Photoshop (v 23.2.2), Abbyy Recognition Server 4.0 Scans: 7, 13, 16: Photoshop (v 23.2.2), Abbyy Recognition Server 4.0 Scan 12: ScanTailor Advanced (1.0.16), Abbyy Recognition Server 4.0	PDF

No.	PARTNER	SOFTWARE USED	GENERATED FORMATS
16	CVTI SR	ScanGate by Treventus Mechatronics, Abbyy Recognition Server 4.0	PDF
17	UREG	Abbyy Recognition Server 4.0	PDF

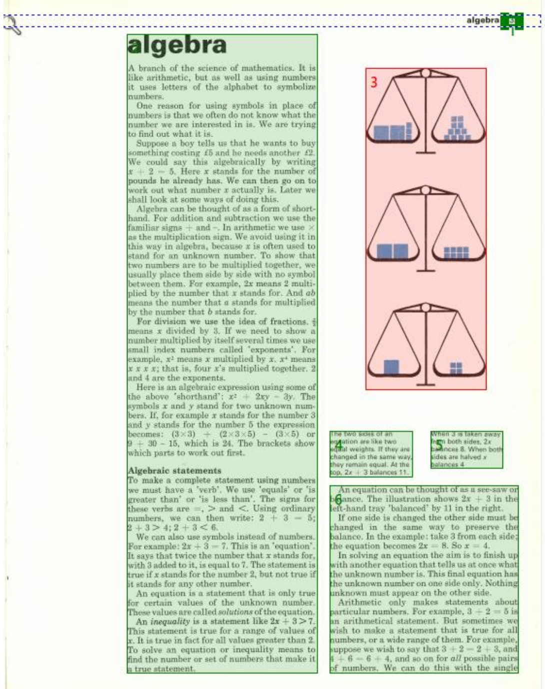
In the outputs with additional manual corrections, Microsoft Office Word (PDF 1.7, RTF 1.9, docx and ePUB 3.0) was mostly used for editing OCR errors and adding structural elements. In one test output, Adobe InDesign was used to edit headers, captions, original page numbers and footnotes. In another test output, the automatically generated PDF was additionally manually processed with Adobe Acrobat Pro, which tagged the content and edited the document's reading order (see Picture 1). In a received output processed with the latest desktop version of Abbyy FineReader 15, the page elements were additionally manually edited and the reading order was corrected (see Picture 2). Another received output used the WordToEpub tool to convert a manually edited Word file to an Epub file (refer to Table 3).

Table 3: Software overview for outputs containing additional manual corrections

No.	PARTNER	SOFTWARE USED	GENERATED FORMATS
1	UIBK	Abbyy FineReader 14, Adobe Indesign	RTF
2	NUK	Adobe Acrobat Pro	PDF
3	NUK	Abbyy FineReader 15, Adobe Acrobat Pro	PDF
4	NUK	Microsoft Office Word, Adobe Acrobat Pro	PDF
5	NUK	Microsoft Office Word, WordToEpub, Sigil	EPUB
6	BNP	LIMB Processing, IRIS OCR	DOCX



Picture 1: Screenshot of only automatically tagged content in Adobe Acrobat Pro before the tags were edited. On the left side, all of the tags are visible in the order they appear on the page, with each tag representing a specific box on the right side. At this point, heading levels are not yet fixed and the order has not been checked.



algebra

A branch of the science of mathematics. It is like arithmetic, but as well as using numbers it uses letters of the alphabet to symbolize numbers.

One reason for using symbols in place of numbers is that we often do not know what the number we are interested in is. We are trying to find out what it is.

Suppose a boy tells us that he wants to buy something costing £5 and he needs another £2. We could say this algebraically by writing $x + 2 = 5$. Here x stands for the number of pounds he already has. We can then go on to work out what number x actually is. Later we shall look at some ways of doing this.

Algebra can be thought of as a form of shorthand. For addition and subtraction we use the familiar signs + and −. In arithmetic we use × as the multiplication sign. We avoid using it in this way in algebra, because x is often used to stand for an unknown number. To show that two numbers are to be multiplied together, we usually place them side by side with no symbol between them. For example, $2x$ means 2 multiplied by the number that x stands for. And ab means the number that a stands for multiplied by the number that b stands for.

For division we use the idea of fractions. $\frac{x}{3}$ means x divided by 3. If we need to show a number multiplied by itself several times we use small index numbers called 'exponents'. For example, x^2 means x multiplied by x. x^4 means $x\ x\ x\ x$; that is, four x's multiplied together. 2 and 4 are the exponents.

Here is an algebraic expression using some of the above 'shorthand': $x^2 + 2xy - 3y$. The symbols x and y stand for two unknown numbers. If, for example x stands for the number 3 and y stands for the number 5 the expression becomes: $(3\times3) + (2\times3\times5) - (3\times5)$ or $9 + 30 - 15$, which is 24. The brackets show which parts to work out first.

Algebraic statements

To make a complete statement using numbers we must have a 'verb'. We use 'equals' or 'is greater than' or 'is less than'. The signs for these verbs are =, > and <. Using ordinary numbers, we can then write: $2 + 3 = 5$; $2 + 3 > 4$; $2 + 3 < 6$.

We can also use symbols instead of numbers. For example: $2x + 3 = 7$. This is an 'equation'. It says that twice the number that x stands for, with 3 added to it, is equal to 7. The statement is true if x stands for the number 2, but not true if it stands for any other number.

An equation is a statement that is only true for certain values of the unknown number. These values are called *solutions* of the equation.

An *inequality* is a statement like $2x + 3 > 7$. This statement is true for a range of values of x. It is true in fact for all values greater than 2. To solve an equation or inequality means to find the number or set of numbers that make it a true statement.

The two sides of an equation are like two equal weights. If they are changed in the same way, they remain equal. At the top, $2x + 3$ balances 11.

When 3 is taken away from both sides, $2x$ balances 8. When both sides are halved x balances 4

An equation can be thought of as a see-saw or balance. The illustration shows $2x + 3$ in the left-hand tray 'balanced' by 11 in the right.

If one side is changed the other side must be changed in the same way to preserve the balance. In the example: take 3 from each side; the equation becomes $2x = 8$. So $x = 4$.

In solving an equation the aim is to finish up with another equation that tells us at once what the unknown number is. This final equation has the unknown number on one side only. Nothing unknown must appear on the other side.

Arithmetic only makes statements about particular numbers. For example, $3 + 2 = 5$ is an arithmetical statement. But sometimes we wish to make a statement that is true for all numbers, or a wide range of them. For example, suppose we wish to say that $3 + 2 = 2 + 3$, and $4 + 6 = 6 + 4$, and so on for *all* possible pairs of numbers. We can do this with the single

Picture 2: Screenshot of edited elements on the page and fixed reading order in Abbyy FineReader 15. Elements are presented in colours: green for text and red for picture. The order numbers are visible on the top left of each element.

The formats of the provided outputs were:

- PDF (15):
  - Automatically generated outputs (12): 6 outputs were in version 1.4, 4 outputs were in version 1.5 (of which 1 was according to the PDF/UA standard), 1 output was in version 1.6 and 1 output was in version 1.7. Of these 12 PDF outputs, 5 were tagged PDFs.
  - Outputs with additional manual corrections (3): 1 output was in version 1.5 and was according to the PDF/UA standard, 1 output was in version 1.6 and 1 output was in version 1.7. All three of these outputs were tagged PDFs.
- XMLs with TXT (3): all 3 outputs were automatically generated. Evaluation was later done on TXT only.
- ePUB (2)
  - Automatically generated outputs (1): the output was in version 2.0.
  - Outputs with additional manual corrections (1): the output was in version 3.0.
- RTF (2)
  - Automatically generated output (1): the output was in version 1.5-1.6.
  - Outputs with additional manual corrections (1): the output was in version 1.9.
- DOCX (1) – the output was in the 2007 onwards version.

The following summary is based on a review of the completed test report questionnaires:

- For 8 outputs, partners reported that they made some changes before importing the scans into their system. These changes concerned changing the resolution to 300 dpi, converting 3 files because there were some problems with uploading, rotating and cropping certain images, and using Image frames and JPEG-Compression.
- When asked which image processing steps were used when working with the sample, partners replied that they used: automatic deskewing (in 9 examples), manual deskewing (in 2 examples), automatic and manual deskewing (in 4 examples), automatic cropping (in 3 examples), manual cropping (in 7 examples), automatic and manual cropping (in 4 examples), line straightening (in 2 examples), noise removal (in 1 example), contrast enhancement (in 2 examples), correction of geometric distortion (in 0 examples), binarization (in 1 example), removal of stamps and written notes (in 1 example), and equalising the dimensions (in 3 examples).
- With regard to OCR, partners mainly used the English language (13 examples), but they also used Latin (3 examples) and more than one language (4 examples). It was reported that machine learning was used for OCR in only 2 examples. For 14 examples, only automatic OCR recognition was used, while manual corrections were used for 5 examples.

- Regarding layout analysis, partners reported that they marked paragraphs (5 examples), columns (4 examples), headers (2 examples), images (3 examples), background images (2 examples) and tables (4 examples).
- Regarding reading order, partners reported that they worked on reading order for 4 examples, while no work was done on reading order in 15 examples.
- Regarding fixing OCR mistakes, it was reported that mistakes were corrected for 3 examples and were not corrected for 17 examples.

3.1 Results by sample scan number

The tables below show the test results by all partners regarding the different criteria in each of the sample scans. The test outputs have been classified into automatically generated outputs (17) and outputs that were additionally manually corrected (6). The results were fully achieved, partially achieved or not achieved. The additionally manually corrected test outputs were delivered in addition to the automatically generated test outputs.

Table 4: Results of Scan 1

CRITERIA	**AUTOMATICALLY generated outputs**			**Additional MANUAL correction**	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	10	0	7	6	
Text order	7	2	8	6	
Heading 1	2	0	16	5	
Heading 2	1	0	18	4	
Picture	4	0	14	5	
Alt-text picture	0	1	18	3	1
Caption	0	0	20		3
Math (simple)	10	0	7	6	
Math (advanced)	0	0	19	4	
Special character	11	0	6		

Additional observations:

- BNP (PDF) text order – pagination disturbs flow of text
- MZK (edited XML, TXT) text order – captions appear at the end of the whole text
- NLE (PDF) text order – recognised text from right to left, top to bottom
- OSZK (PDF) text order – trouble with recognition of columns – order from right to left, top to bottom
- UG (PDF) math simple – only one + is wrongly recognised
- VKOL (XML, TXT) text order – pagination interrupts the text order (it is placed before captions)

Table 5: Results of Scan 2

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	16		1	6	
Text order	11	3	3	6	
Heading 2	1		16	4	
Picture	4		13	5	
Alt-text picture		1	16	3	1
Alt-text picture (chem. formula)			17	3	1
Alt-text picture (chem. formula)			17	3	1
Picture (chem. Formula)	5		15	5	
Picture (chem. Formula)			17	5	
Caption	3			1	3
Math (simple)	15		2	6	
Math (advanced)		1	16	4	
Special character					

Additional observations:

- NLE (PDF) math – 0 appears instead of O
- OSZK (PDF) text order – trouble with recognition of columns – order from right to left, top to bottom
- OSZK (PDF) math – 0 appears instead of O
- UG (PDF) picture chem. – one picture is not recognised
- UIBK (ODM PDF) text order – not all text is recognised
- UIBK (ODM RTF) math advanced – some examples are done correctly, but not all
- UT (PDF) picture chem. – neither of the two chemistry pictures are recognised

Table 6: Results of Scan 3

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Text order	7	2	8	6	
Heading 2	1		16	4	
Heading 2	1		16	4	
Picture	5		12	5	
Alt-text picture	1		15	3	1
Caption		1			3
Stamp removal	3		14	4	
Math (simple)					

Additional observations:

- MZK (edited XML, TXT) text order – captions appear at the end of the whole text
- MZK (small edited XML, TXT) text order – the first paragraph appears at the end
- OSZK (PDF) text order – columns are not recognised, so the text flows in rows from left to right
- OSZK (PDF) text order – trouble with recognition of columns – order from right to left, top to bottom
- UG (PDF) caption – a tag is created, but it does not contain the right text
- UIBK (RTF) pagination – the original does not have pagination here
- UIBK (ODM PDF) text order – the order in the PDF is not correct – it flows from right to left
- UIBK (ODM RTF) text order – not all of the text is recognised
- VKOL (XML, TXT) text order – the caption columns are switched (the right column appears before the left one)

Table 7: Results of Scan 4

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination double	13		4	6	
Pagination	10		6	5	1
Text order	6		10	3	2
Picture			17	5	
Alt-text picture			17	3	
Caption			17		2

Additional observations:

- OSZK (PDF) text order – trouble with recognition of the columns – order from right to left, top to bottom
- UIBK (RTF) – it is unclear how the chemical elements were presented in the table

Table 8: Results of Scan 5

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	12		5	6	
Text order	6	5	6	6	
Heading 1	1	1	15	5	
Picture	6		11	5	
Picture	4	2	11	5	
Picture	6		11	5	
Alt-text picture		1		3	1

Caption	1	2	14	1	3
Caption	1	2	14	1	3
Caption		1			3
Math (advanced)			17	4	

Additional observations:

- NCU (PDF) heading 1 – the heading is marked, but as a level 3 heading
- NCU (PDF) caption – one caption is missing (it is marked as a paragraph)
- NLS (PDF) text order – minor mistake in text order
- OSZK (PDF) text order – trouble with recognition of the columns – order from right to left, top to bottom
- UG (PDF) text order – the first two captions are switched (the second caption appears before the first one)
- UG (PDF) caption – a tag is created, but the content is switched between the first two captions; the third caption is tagged, but the content is not correct
- UG (PDF) picture – the scheme is divided into five pictures
- UIBK (ODM PDF) text order – minor mistakes
- UIBK (ODM RTF) text order – missing text
- UT (PDF) text order – the first two captions are switched (the second caption appears before the first one)
- UT (PDF) caption – a tag is created, but the content is switched between the first two captions; the third caption is missing
- VKOL (XML, TXT) text order – the first two captions are switched (the second caption appears before the first one)

Table 9: Results of Scan 6

CRITERIA	**AUTOMATICALLY generated outputs**			**Additional MANUAL correction**	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Text order	1	5	11	5	1
Heading 1	2		15	5	
Heading 2	1		16	4	
Heading 2			17	4	
Heading 2			17	3	
Heading 2			17	3	
Picture	3		14	5	
Picture	3		14	5	
Alt-text picture		1	16	3	
Alt-text picture		1	16	3	
Caption	2		15		3
Caption	1	1	15		3
Language segments			17	2	1

Additional observations:

- BNP (PDF) text order – caption interrupts the flow of the text
- BNP (DOCX) text order – one picture is misplaced
- CVTI SR (PDF) text order – minor mistake in text order (the caption before the last line)
- CVTI SR (PDF) – the scan has better contrast due to the white background, which is better for OCR as well as for users
- MZK (edited XML, TXT) text order – minor mistake – the first caption is at the end of the whole text
- MZK (small edited XML, TXT) text order – the columns are not detected, the text follows in one straight line
- NCU (PDF) heading 2 – there are four occurrences of heading 2, but only one is marked (the first one)
- NLE (PDF) text order – minor mistake in text order (caption before the last line) and the author appears after the title
- OSZK (PDF) text order – mixture of text order, horizontal and vertical
- UG (PDF) heading 2 – the heading is wrongly tagged
- UG (PDF) caption – both captions are tagged, but one does not have the right text
- UG (EPUB) text order – incorrect text order (there is one column from the top to the bottom of the whole page, then a second column and a third column, etc.)
- UREG (PDF) text order – minor mistake in text order (the caption before the last line)
- VKOL (XML, TXT) text order – the caption interrupts the text order (it is placed before the third column)

Table 10: Results of Scan 7

CRITERIA	**AUTOMATICALLY generated outputs**			**Additional MANUAL correction**	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination double	10		7	2	
Pagination			17	5	
Text order		1	16	5	
Heading 1		1	16	4	
Picture	4		13	5	
Picture	4		13	5	
Picture	4		13	5	
Picture	1	2	14	5	
Alt-text picture		1	16	3	1
Alt-text picture		1	16	3	1
Alt-text picture		1	16	3	1
Alt-text picture		1	16	3	1
Caption	2	1	14	1	3
Caption	2		14	1	3

Caption	1	2	14	1	3
Caption		2	14	1	3
Initial	8		9	6	
Language segm.			17	3	
OCR errors (text in picture 4)	4		13	6	

Additional observations:

- MZK (small edited XML, TXT) text order – the main text is correct, but the picture captions interrupt the flow of the text and the text from the image is also captured
- NCU (PDF) caption – the fourth caption has the wrong text (the text is taken from the picture)
- NCU (PDF) heading 1 – the title and author are marked as heading 3 and heading 4, respectively
- NCU (PDF) picture – the fourth picture is only half recognised (probably because of the text in the picture)
- NCU (PDF) initial – the initial is marked as a picture with alt-text, which is "Figure without the caption"
- NLS (PDF) text order – not all of the text is OCR recognised (the text in the last two captions and the text in the last column is omitted)
- UG (PDF) caption – all four captions are tagged, but two do not have the right text
- UG (PDF) picture – the fourth picture is only half recognised (probably because of the text in the picture)
- UT (PDF) caption – 2 of 4 captions are tagged (and have the right text)
- VKOL (XML, TXT) text order – captions interrupt the text order, the main title is placed within the text

Table 11: Results of Scan 8

CRITERIA	**AUTOMATICALLY generated outputs**			**Additional MANUAL correction**	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	15		2	5	
Text order	16	1		6	

Additional observations:

- NCU (PDF) heading 1 – the heading is marked as heading 5
- NLE (PDF) text order – the page number and heading are not included

Table 12: Results of Scan 9

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination			17		
Text order	13		4	6	
Heading 1	1	1	15	4	
Heading 2	1		16	4	
Heading 3	1		16	3	

Additional observations:

- NCU (PDF) heading 1, 2, 3 – the parallel title is marked as heading 5
- UG (PDF) heading 1, 2, 3 – all three headings are tagged correctly, but there is an error due to a parallel title that should also be heading 1

Table 13: Results of Scan 10

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	5		12	5	1
Text order			17	5	
Caption		2	15		2
Table	2	1	14	5	
Table header	1		16	4	
Rows	1		16	5	
Language segments			17	2	
Page rotation	4	1	12	5	

Additional observations:

- BNP (PDF) page rotation – there is a remark that the page is not rotated visually, but OCR is rotated and correctly recognised
- MZK (edited XML, TXT) text order – heading 1 and page number are not recognised
- MZK (small edited XML, TXT) text order – the text is not correctly recognised (columns instead of rows)
- NCU (PDF) caption – the caption is tagged, but it does not have the right text
- NCU (PDF) table – **best result without manual corrections!**
- NLS (PDF) text order – only the table title is OCR recognised
- OSZK (PDF) table – a table is created, but without content
- UG (PDF) caption – the caption is tagged, but it does not have the right text

Table 14: Results of Scan 11

CRITERIA	**AUTOMATICALLY generated outputs**			**Additional MANUAL correction**	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	16		1	6	
Text order	5	6	6	5	
Caption			17	1	2
Table	8		9	5	
Table header	1		16	4	
Rows	3	4	10	5	
Caption			17	1	2
Table	8		9	5	
Table header	1		16	4	
Rows	3	4	10	5	
Language segments			17	1	1

Additional observations:

- CVTI SR (PDF) table rows – some minor errors in the recognised table rows (there is a problem with two or three lines in one row)
- MZK (edited XML, TXT) text order –the titles of the rows are recognised first, followed by the columns from left to right (not the rows)
- MZK (small edited XML, TXT) text order – the titles of the rows are recognised first, followed by the columns from left to right (not the rows)
- NCU (PDF) table rows – minimal error in row recognition
- UG (PDF) table rows – rows are tagged, but incorrectly (should be 13 rows but only 4 are tagged)
- UG (PDF) caption – the caption is tagged as heading 3
- UG (EPUB) table rows – table rows are incorrectly formulated/recognised
- UIBK (ODM PDF) text order – some minor errors in the recognised table rows (problem with two or three lines in one row)
- UT (PDF) table rows – minimal errors in row recognition
- VKOL (XML, TXT) table rows – some minor errors in the recognised table rows (there is a problem with two or three lines in one row); in the second table, the first row is placed at the end

Table 15: Results of Scan 12

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	15		2	6	
Text order	3	1	13	6	
Picture	6		11	5	
Alt-text picture		1	16	3	
Caption	2		15	1	3
Math (simple)	8		9	6	
Math (advanced)			17	4	
Special character	17			6	
Footnotes			17	2	2
Language segments			17	2	

Additional observations:

- MZK (edited XML, TXT) text order – the captions appear at the end of the whole text
- NCU (PDF) picture – the picture is divided into two parts
- UIBK (ODM RTF) pagination – the page number is in a text block that is not detectable by assistive technologies

Table 16: Results of Scan 13

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination			17	3	
Text order	17			6	
Heading 1			17	4	
Picture	6		11	5	
Alt-text picture		1	16	3	
Caption		1	16	1	3
Language segments			17	3	
Special character	2		15	2	1

Additional observations:

- BNP (PDF) text order – a figure interrupts the flow of the text
- CVTI SR (PDF) pagination – page number not recognised
- UT (PDF) caption – the caption is tagged, but the page number is also included in the text

Table 17: Results of Scan 14

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Text order	2		15	6	
Picture	4		13	5	
Picture	3	1	13	5	
Picture	3	1	13	5	
Alt-text picture		1	16	3	1
Alt-text picture		1	16	3	1
Alt-text picture		1	16	3	1

Additional observations:

- UG (EPUB) – the page is doubled
- UT (PDF) picture – two of three pictures are tagged; the second and third pictures are merged into one
- UT (PDF) text order – the third paragraph is marked as a caption

Table 18: Results of Scan 15

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Text order	5	1	11	4	1
Heading 1			17	4	

Additional observations:

- BNP (PDF) text order – minor mistakes in text order
- NCU (PDF) text order – some chapters are marked as a list
- UIBK (RTF) text order – we do not think this page should be represented as a table
- UT (PDF) text order – the chapters are marked as a list

Table 19: Results of Scan 16

CRITERIA	AUTOMATICALLY generated outputs			Additional MANUAL correction	
	FULLY achieved	PARTIALLY achieved	NOT achieved	FULLY achieved	PARTIALLY achieved
Pagination	13		4	6	
Text order	12	3	2	6	
Heading 2			17	4	
Table	5	1	11	5	
Table header		1	16	4	
Rows	2	4	11	5	
Caption		3	14		3

Additional observations:

- MZK (edited XML, TXT) text order – most of the numbers in the table are missing
- MZK (small edited XML, TXT) text order – some text from the table is missing
- NCU (PDF) table header – the table header is marked but has the wrong text
- NCU (PDF) caption – the caption is tagged but has the wrong text
- NLS (PDF) text order – numbers in the table and the table caption are not recognised
- UG (PDF) caption – the caption is tagged but does not have the right text
- UG (EPUB) text order – the top cells of the table are missing
- UIBK (ODM PDF) text order – the top rows of the table are missing
- UT (PDF) caption – the caption is tagged but does not have the right text

General observations:

- NCU (PDF) – all of the pictures have alt-text, but the content is correct (the text is the content of the caption)
- NCU (PDF) – most of the headings are marked, but the levels are not correct in some cases
- NCU (PDF) – Scans 13 and 15 are doubled: OCR and whole page layout picture
- NLE (PDF) – mixed text order in most scans
- NLE (PDF) – OCR works much better for newspaper than for monographs!
- UG (EPUB) – the file for Scans 10–12 have a table of contents
- UIBK (RTF) – heading 1 and 2 should be used; "titel mit abstand" was used as well as its copy for heading 2
- UIBK (RTF) – no pictures were included
- UIBK (RTF) – alt-text is included with the text instead of behind a picture (no pictures)
- UIBK (RTF) – the caption is marked, but not with the function ("insert caption")
- UT (PDF) – none of the scans are cropped, but we noticed that OCR recognised some characters from the next page
- VKOL (XML, TXT) – **the original pagination is marked in the top right corner of each scan in the digital library portals!**

3.2 Results according to criteria

Tables 20 and 21 show the results according to each of the established criteria. For easier understanding, the top results for each criterion (shown in bold) are further described.

Table 20: Results according to criteria for the automatically generated outputs

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**
EODOPEM Partners		NUK	NLE	NLS	CVTI 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

Used codes: A = fully achieved criterion; B = partly achieved criterion; empty cell = criterion was not achieved; * = tagged PDF

Table 21: Results according to criteria for the outputs with additional manual corrections

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**
EODOPEM 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**

Used codes: A = fully achieved criterion; B = partly achieved criterion; empty cell = criterion was not achieved; * = tagged PDF

- ALT-TEXT PICTURE (18)

The best result among the automatically generated outputs was achieved by the PDF/UA format from the Nicolaus Copernicus University in Torun (17B). The best result among outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (18A). The same result was also achieved by the PDF and ePUB formats created by Microsoft Word with manual corrections by the National and University Library (Slovenia).

- ALT-TEXT PICTURE (CHEM. FORMULA) (2)

None of the automatically generated outputs achieved this criterion, but the best result among the outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (2A). The same result was also achieved by the PDF and ePUB formats created by Microsoft Word with manual corrections by the National and University Library (Slovenia).

- CAPTION (19)

The best result among the automatically generated outputs was achieved by the PDF/UA format from the Nicolaus Copernicus University in Torun (9A 3B) and the PDF format from the University of Greifswald (5A 9B). The best result among the outputs with additional manual corrections was achieved by the PDF/UA format created by the latest desktop version of Abbyy FineReader from the National and University Library (Slovenia) (11A 1B).

- FOOTNOTES (1)

None of the automatically generated outputs achieved this criterion, but the best result among the outputs with additional manual corrections was achieved by the RTF format from the University of Innsbruck (1A). The same result was also achieved by the ePUB format created by Microsoft Word with manual corrections by the National and University Library (Slovenia).

- HEADING 1 (7)

The best result among the automatically generated outputs was achieved by the PDF format from the University of Greifswald (4A). The best result among the outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (7A). The same result was also achieved by the PDF format with manually corrected tags in Adobe Acrobat Pro, as well as by the PDF and ePUB formats created by Microsoft Word with manual corrections, all three of which were from the National and University Library (Slovenia).

- HEADING 2 (10)

The best result among the automatically generated outputs was achieved by the PDF/UA format from the Nicolaus Copernicus University in Torun (5A). The best result among the outputs with additional manual corrections was achieved by the PDF with manually corrected tags in Adobe Acrobat Pro, as well as by the PDF and ePUB formats created by Microsoft Word with manual corrections, all three of which were from the National and University Library (Slovenia) (10A).

- HEADING 3 (1)

The best result among the automatically generated outputs was by achieved the PDF format from the University of Greifswald (1A). The best result among the outputs with additional manual corrections was achieved by the PDF with manually corrected tags in Adobe Acrobat Pro, as well as by the PDF and ePUB formats created by Microsoft Word with manual corrections, all three of which were from the National and University Library (Slovenia) (1A).

- INITIAL (1)

Among the automatically generated outputs, 8 PDF outputs fully achieved this criterion: the National Library of Estonia, the National and University Library (Slovenia), the National Library of Sweden, Slovak Centre of Scientific and Technical Information, the University Library Regensburg, the National Széchényi Library, the National Library of Portugal and the University of Tartu Library. All six outputs with additional manual corrections fully achieved this criterion.

- LANGUAGE SEGMENTS (6)

None of the automatically generated outputs achieved this criterion, but the best result among those with additional manual corrections was achieved by the ePUB format created by Microsoft Word with manual corrections from the National and University Library (Slovenia) (6A), closely followed by the PDF format created by Microsoft Word with manual corrections, also from the National and University Library (Slovenia) (5A 1B).

- MATH (SIMPLE) (3)

The best possible result (3A) among the automatically generated outputs was achieved by five PDF outputs: the National Library of Sweden, the National and University Library (Slovenia), the University Library Regensburg, the National Library of Portugal and the Nicolaus Copernicus University in Torun. All six of the outputs with additional manual corrections fully achieved this criterion.

- MATH (ADVANCED) (4)

The best result among the automatically generated outputs was achieved the RTF format from the University of Innsbruck (1B). The best result among the outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (4A). The same result was also achieved by the RTF format from the University of Innsbruck, as well as by the PDF and ePUB formats created by Microsoft Word with manual corrections from the National and University Library (Slovenia).

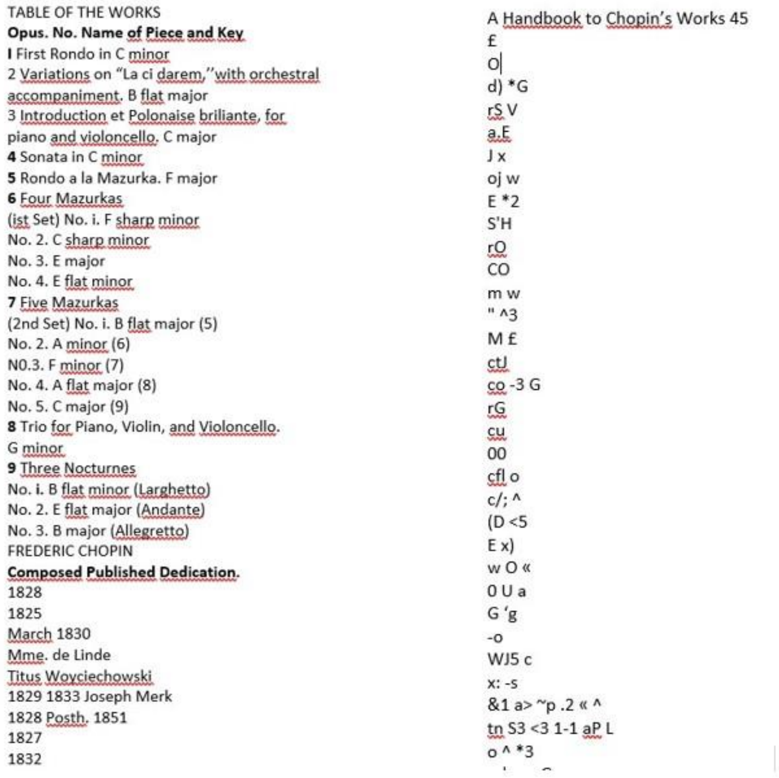
- OCR ERRORS (1)

The best possible result (1A) among the automatically generated outputs was achieved by four PDF formats from the National Library of Sweden, the National Széchényi Library, the

National Library of Portugal and the University of Tartu Library. All six of the outputs with additional manual corrections fully achieved this criterion.

- PAGE ROTATION (1)

Four of the automatically generated outputs used page rotation. The PDF outputs were: the National Library of Estonia, the Slovak Centre of Scientific and Technical Information and the University Library Regensburg. The same result was achieved by the edited XML and TXT format from the Moravian Library. All of the outputs with additional manual corrections fully achieved this criterion, except for the PDF format with manually corrected tags in Adobe Acrobat Pro from the National and University Library (Slovenia).
On verifying whether page rotation influenced any of the other criteria, it was observed that, at least in the presented outputs, this criterion did not influence the recognised table elements, as these PDF outputs were tagged PDFs. Nor did it influence the text order. It did, however, influence OCR recognition, as all four automatically generated outputs had good OCR, whereas the other examples were not always good (see Picture 3).



```
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
```

Picture 3: Comparison of two OCR outputs. 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 (poor OCR output).

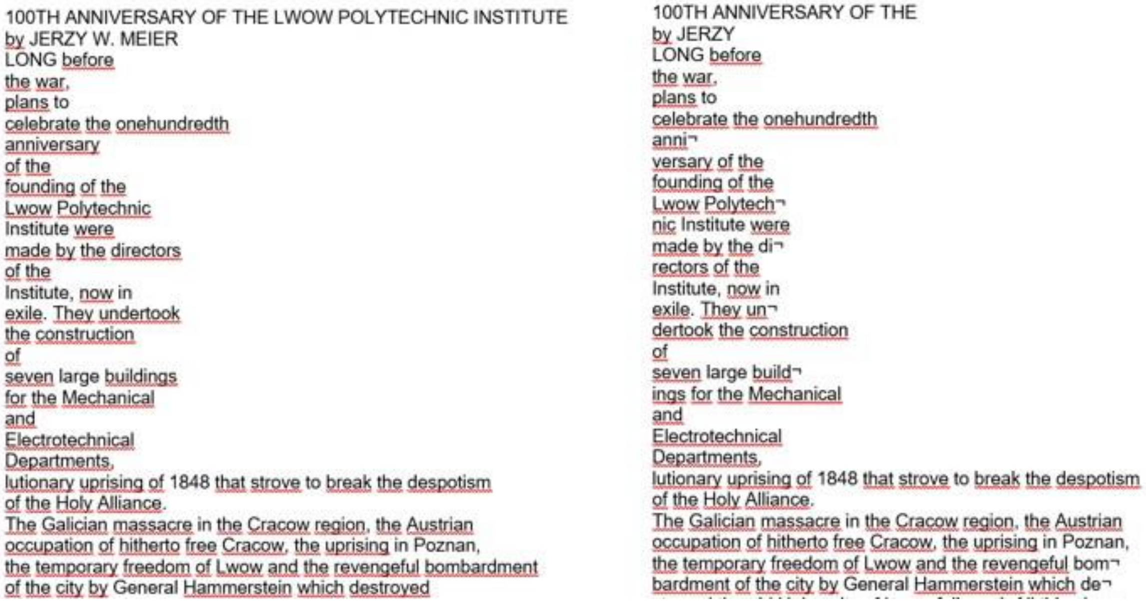
- PAGINATION (12)

The best result among the automatically generated outputs was achieved by the PDF format from the University of Greifswald and the National and University Library (Slovenia) (9A). The best result among the outputs with additional manual corrections was achieved by the PDF and ePUB formats created by Microsoft Word with manual corrections by the National and University Library (Slovenia).

- PAGINATION–DOUBLE (2)

Ten of the automatically generated outputs did not split double pages. The PDF outputs were: the National Library of Estonia, the National Library of Sweden, the National Széchényi Library, the University Library Regensburg, the University of Greifswald, the Nicolaus Copernicus University in Torun and the National Library of Portugal. The same result was achieved by the edited and small edited XML and TXT format from the Moravian Library and the ePUB format from the University of Greifswald. Two of the outputs with additional manual corrections also failed to split double pages: the RTF format from the University of Innsbruck and the PDF format created by Microsoft Word from the National and University Library (Slovenia).

We verified whether the pagination-double criterion influenced any of the other criteria and observed that, at least in the presented outputs, this criterion did not influence any other criteria. It did, however, influence OCR recognition, as some of the automatically generated outputs had better OCR with regard to the full title and author that were spread over a double page, but this did not work on all of the examples (see Picture 4).



```
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
the construction
of
seven large buildings
for the Mechanical
and
Electrotechnical
Departments,
lutionary uprising of 1848 that strove to break the despotism
of the Holy Alliance.
The Galician massacre in the Cracow region, the Austrian
occupation of hitherto free Cracow, the uprising in Poznan,
the temporary freedom of Lwow and the revengeful bombardment
of the city by General Hammerstein which destroyed
```

```
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¬
dertook the construction
of
seven large build¬
ings for the Mechanical
and
Electrotechnical
Departments,
lutionary uprising of 1848 that strove to break the despotism
of the Holy Alliance.
The Galician massacre in the Cracow region, the Austrian
occupation of hitherto free Cracow, the uprising in Poznan,
the temporary freedom of Lwow and the revengeful bom¬
bardment of the city by General Hammerstein which de¬
```

Picture 4: Comparison of two OCR outputs. The first example 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 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, but they not related to the criterion pagination-double.*

- PICTURE (18)

The best result among the automatically generated outputs was achieved by the PDF format from the Nicolaus Copernicus University in Torun (16A 1B) and the PDF format from the University of Greifswald (15 A 2 B). All of the outputs with additional manual corrections achieved the best result (18A), except for the RTF format from the University of Innsbruck.

- PICTURE (CHEM. FORMULA) (2)

The best result among the automatically generated outputs was achieved by the PDF and ePUB formats from the University of Greifswald (1A). All of the outputs with additional manual corrections achieved the best result (2A), except for the RTF format from the University of Innsbruck.

- PRIMARY LANGUAGE (1)

Two of the automatically generated outputs had primary language added to the document: the PDF format from the Nicolaus Copernicus University in Torun and the RTF format from the University of Innsbruck. All six of the outputs with additional manual corrections fully achieved this criterion.

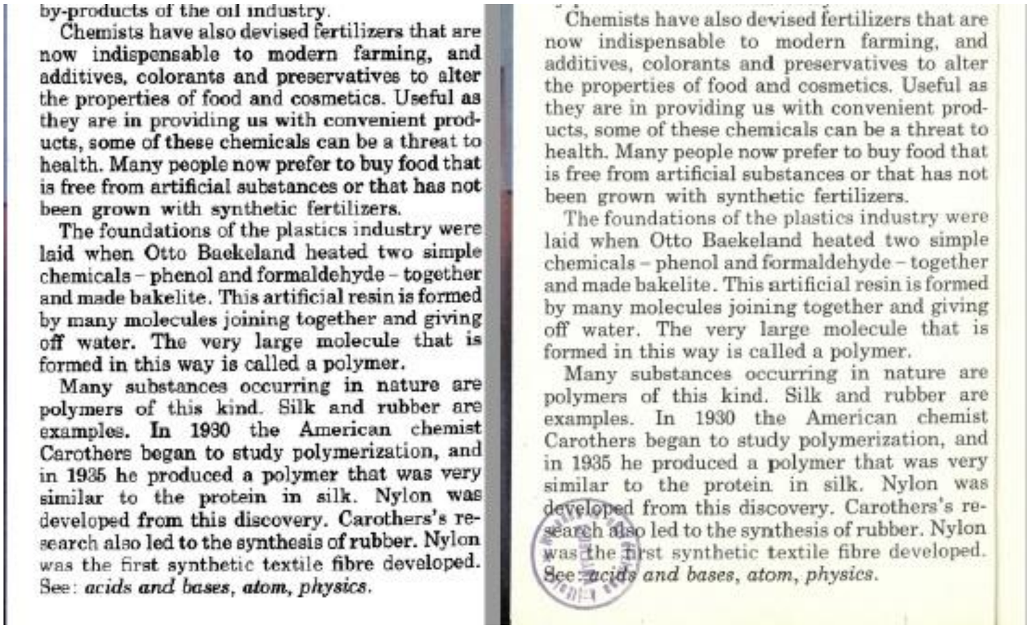
- SPECIAL CHARACTER (3)

The best result (2A) among the automatically generated outputs was achieved by the PDF format from the National and University Library (Slovenia), the National Library of Estonia, the National Library of Sweden, the Slovak Centre of Scientific and Technical Information, the University Library Regensburg, the National Széchényi Library, the University of Greifswald, the Nicolaus Copernicus University in Torun and the University of Tartu Library. The same result was achieved by the edited and small edited XML and TXT format from the Moravian Library, the XML and TXT format from the Olomouc Research Library and the ePUB format from the University of Greifswald. All six outputs with additional manual corrections achieved the same result at this criterion (2A).

- STAMP REMOVAL (1)

Three of the automatically generated outputs removed the stamp on a scan: the PDF format from the National Széchényi Library and the edited and small edited XML and TXT format from the Moravian Library. Four of the outputs with additional manual corrections also removed the stamp: the RTF format from the University of Innsbruck, the docx format from the National Library of Portugal, and the PDF and ePUB formats created by Microsoft Word from the National and University Library (Slovenia).

On verifying whether the removed stamp influenced OCR recognition in the area of the stamp, it was found that all three of the automatically generated outputs, as well as the outputs with manual corrections, had clean OCR with no mistakes in the paragraph concerned, in comparison to outputs in which the stamp had not been removed (see Pictures 5 and 6).



by-products of the oil industry.

Chemists have also devised fertilizers that are now indispensable to modern farming, and additives, colorants and preservatives to alter the properties of food and cosmetics. Useful as they are in providing us with convenient products, some of these chemicals can be a threat to health. Many people now prefer to buy food that is free from artificial substances or that has not been grown with synthetic fertilizers.

The foundations of the plastics industry were laid when Otto Baekeland heated two simple chemicals – phenol and formaldehyde – together and made bakelite. This artificial resin is formed by many molecules joining together and giving off water. The very large molecule that is formed in this way is called a polymer.

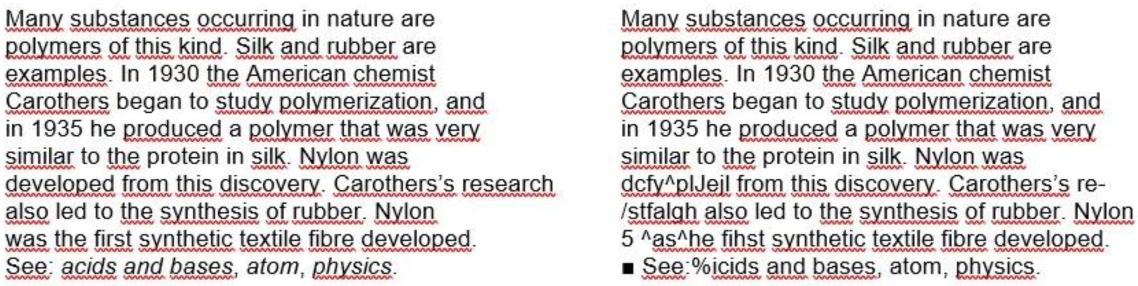
Many substances occurring in nature are polymers of this kind. Silk and rubber are examples. In 1930 the American chemist Carothers began to study polymerization, and in 1935 he produced a polymer that was very similar to the protein in silk. Nylon was developed from this discovery. Carothers's research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

Chemists have also devised fertilizers that are now indispensable to modern farming, and additives, colorants and preservatives to alter the properties of food and cosmetics. Useful as they are in providing us with convenient products, some of these chemicals can be a threat to health. Many people now prefer to buy food that is free from artificial substances or that has not been grown with synthetic fertilizers.

The foundations of the plastics industry were laid when Otto Baekeland heated two simple chemicals – phenol and formaldehyde – together and made bakelite. This artificial resin is formed by many molecules joining together and giving off water. The very large molecule that is 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.*

Picture 5: Comparison of two PDF outputs. The first example is the output when the stamp was removed, resulting in a clean text. The second example is the output when the stamp was not removed, which creates reading difficulties.



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^plJejl 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.

Picture 6: Additional comparison of the two OCR outputs. The first example is the output when the stamp was removed, resulting in a correct text with no mistakes. The second example is the output when the stamp was not removed, resulting in mistakes in the text that cause reading difficulties (especially with speech synthesis).

- TABLE (4)

The best result among the automatically generated outputs was achieved by the PDF format from the Nicolaus Copernicus University in Torun (4A), closely followed by two examples of the PDF format from the University of Innsbruck, the PDF format from the University of Greifswald and the University of Tartu Library, and the RTF format from the University of Innsbruck, all of which achieved the result 3A. All of the outputs with additional manual corrections achieved the best result, except for the PDF format with manually corrected tags in Adobe Acrobat Pro from the National and University Library (Slovenia).

- TABLE HEADER (4)

The best result among the automatically generated outputs was achieved by the PDF format from the Nicolaus Copernicus University in Torun (3A 1B). The best result among the outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (4A). The same result was also achieved by the PDF and ePUB formats created by Microsoft Word with manual corrections and the PDF/UA format created by the latest desktop version of Abbyy FineReader from the National and University Library (Slovenia).

- TABLE ROWS (4)

The best result among the automatically generated outputs was achieved by the PDF format from the Nicolaus Copernicus University in Torun (2A 2B). All of the outputs with additional manual corrections achieved the best result, except for the PDF format with manually corrected tags in Adobe Acrobat Pro from the National and University Library (Slovenia).

- TEXT ORDER (16)

The best result among the automatically generated outputs was achieved by the PDF format from the Nicolaus Copernicus University in Torun (12A). The best result among the outputs with additional manual corrections was achieved by the docx format from the National Library of Portugal (15A 1B). The same result was also achieved by the PDF and ePUB formats created by Microsoft Word with manual corrections from the National and University Library (Slovenia).

4 Test findings

Although manual corrections can significantly improve OCR quality, some exceptions were found, as described below. Since manual correction is time consuming, the objective is to achieve the best automatic outputs before any manual corrections are needed.

The test results show that the best outputs were achieved using PDF/UA as a delivery format or tagged PDF. These file formats dealt with all of the criteria better than the other file formats. The only shortcoming is the inability to visually adapt the content to specific needs, as described in section 1.2 Description of the report.

However, some criteria were only partially met: either they were almost met or they were technically adequate but the content was not related (e.g., the Alt-text field was assigned, but the content was not correct – the text corresponded to the caption, or the caption tag was assigned, but the text inside the element was not the text corresponding to the image). The alt-text criterion exemplified the most problems, as it is an element that currently requires human input.

For the blind and partially sighted, the most acceptable delivery formats are Microsoft Word files (RTF and doc) and ePUBs in annotated PDF format.

There are few criteria that are as important for mobile devices as they are for the blind and partially sighted, but output that is well adapted for the blind and partially sighted is also friendlier for mobile devices. The most useful formats for mobile devices are next delivery file formats: EPUBs and Microsoft Word files (RTF in docx).

Average scan qualities were deliberately used in order to test the partners' OCR tools to the greatest possible extent. Among other factors, the structure of the scan is very important. The results may also have been different if the test sample had comprised texts from a single publication. In this case, the texts would have had the same structure, or at least a similar one (e.g., the allocation of title tags is based on the size and font of the titles). Different scans also caused problems for some partners, as the system did not accept scans of different sizes or different systems were used for monographs and newspapers. Some partners solved this problem by importing each scan separately.

Some problems were due to specific elements in the scan, such as the table of contents in Scan 15 . The OCR output was plain text, although in the case of a whole book, internal links to individual chapters would be desirable.

The results were also influenced by the complexity of the structure of the elements in the scans. Scans with a simple (one column) structure had fewer errors than those with a complex structure (multiple columns and other elements).

In addition, the testing gave rise to the following findings:

- Except for the PDF/UA format, no links are evident between format versions or standards.

- Page rotation enabled better recognition of tables.
- In the case of double pages, scans with joint title and author's name spread over both pages were not recognised as a joint element: the texts of the title and author's name were affected when the double pages were split in two.
- There is a need to find a method for analysing partners' workflows and assessing the potential impact on results, and for determining how additional manual work affects the results (BNP, UIBK, NUK).

We should take in consideration the fact that delivery formats that meet the needs of the blind and partially sighted also enable a better experience for users of mobile devices.

None of the EODOPEN partners produce audiobooks, so we were unable to analyse this aspect.

5 Possible solutions and recommendations

Some recommendations and solutions concerning delivery formats for mobile devices as well as for print-disabled users are presented below. There are two possible paths that libraries and other institutions can choose, depending on the users' needs. In order to achieve the best possible user experience, we also offer a third option, i.e., the integration of both models, thereby increasing accessibility for everyone.

5.1 Solutions for mobile devices

The biggest problem with mobile devices, especially mobile phones, is the small screen size, for which non-responsive delivery file formats are not recommended. In addition, it is more difficult to search for parts of a publication on mobile devices. It is therefore recommended to enable easier navigation through the work, at least by the main chapters or by the original page numbers or other specified landmarks. It is also important to allow a format that enables at least basic visual adaptations of the text to the personal needs of the users (text background, font, text size, etc.). For better use of publications on mobile devices, we suggest minimal manual interventions in the publications themselves.

Based on the analysis of the test results, a survey among users and the reviewed literature in D11: Guidelines and Recommendations…, we recommend the following:

1. Delivery file formats that are adaptable to screens should be used (EPUB, MOBI, AZW, HTML and variations of Microsoft Word documents). These formats also enable additional functionalities, such as adding bookmarks, changing the visual appearance, etc.
2. Among the above-mentioned formats, Microsoft Word variations, EPUB and HTML are open and not proprietary file formats compared to MOBI and AZW. The proprietary file formats should be used only when we are aware that the user has appropriate software to access the content.
3. When using PDF as a delivery file format, we recommend selecting tagged PDF or PDF/UA.
4. We suggest enabling a table of contents or structural tags that mark the headings in the publication, thus allowing navigation within the applications/programs for reading on the devices themselves. Alternatively, a page with a table of contents can be added.
5. Particular attention should be paid to "page rotation", "pagination double" and "stamp removal", as these criteria have been shown to improve visual appearance, as well as OCR. However, we suggest deciding on this on a case-by-case basis.

For the not proprietary file formats PDF, PDF/UA, DOCX, RTF and EPUB, the following software were mostly used to generate the formats: Abbyy Finereader, Microsoft Office Word, Adobe InDesign, Adobe Acrobat Pro and WordToEpub. The results vary among software and among the amount of manual work put into creation of the format so we can not give any specific recommendation.

We should consider the fact that users of mobile devices can also be users of assistive technologies such as speech synthesis, as more and more sighted people enjoy listening to audio publications. In this case, solutions for print-disabled users should be applied in order to ensure access to the widest possible group of users.

5.2 Solutions for print-disabled users

Solutions for the blind and partially sighted, as well as other people who have problems accessing conventional print or electronic publications, are more complex and require more work, time and specialised knowledge. It is important to have access not only to publications, but also to assistive technology and, through this, to achieve a fluid flow of the text, despite complex elements and demanding page structure. In this regard, the most important criteria are: text order, OCR clean-up, primary language and language segments. Due to linear reading, it is necessary to allow navigation to different locations in the publication (via chapters or original pages of the publication or other landmarks). There is also a need to facilitate the understanding of visual elements. Special elements (tables, table headers, captions, footnotes, hyperlinks, complex mathematical notations, etc.) should be adapted to function technically with the help of assistive technologies. Last but not least, it is also important to enable people with residual vision to visually adapt the appearance of the publication to their personal needs (enlargement of the text, background of the text, change of font, etc.)

We recommend:

1. Undertaking an OCR clean-up and fixing the text order.
2. Paying attention to text contrast – scanned PDFs usually have low contrast between text and background due to the colour of the paper.
3. Adding document language and marking segments that are in a different language.
4. Adding navigation segments for chapters, subchapters, original page number, captions, footnotes, hyperlinks, etc.
5. Adding descriptions for visual elements that contribute additional value to the surrounding text, e.g., alt-text for images, graphs, etc.
6. Fixing the structure of tables: table headers, table rows and table cells.
7. Devoting special attention to mathematical expressions. If possible, MathML and/or Latex should be used.
8. Specifically for the blind: using formats that do not contain visual elements and that support assistive technology, e.g., TXT or variations of Microsoft Word documents. In this case, the visual elements are not needed, but alt-text is even more crucial.
9. Specifically for the visually impaired: using formats that are adaptable to screens and that also enable other modern functions for working with the material, e.g., adding

bookmarks, changing the visual appearance, etc. (variations of Microsoft Word documents, EPUB, HTML, MOBI, AZW). Consider open and not proprietary file formats.

10. Using tagged PDF or PDF/UA when using the PDF format.
11. Testing at least one assistive technology or using a test group of blind and partially sighted people and implementing their observations in future workflows.

6 Summary

The aim of the Report on Trial Implementations for Mobile Devices and Print-Disabled Users is to help libraries and other cultural organisations to make digitised content available to a broader community. The Report is based on EODOPEN partners' digitisation experiences at their organisations and complements the EODOPEN Project Deliverable 11: *Guidelines and Recommendations for the Provision of Alternative and Special Formats*, which addresses delivery formats and criteria for increasing the quality of digitisation results for users of mobile devices as well as blind and partially sighted users.

In order to find out which scanning and recognition workflows are optimal for achieving the best results in OCR, a trial implementation among EODOPEN partners was undertaken. One of the goals was to determine which file formats could be generated, as different file formats can give users different user experiences.

The test sample consisted of 16 scans in the TIFF format (see Annex 1), comprising both textual and non-textual elements, such as plain text, chapters and sub-chapters, columns, tables, footnotes, flowcharts, images and text accompanying images (captions). In order to obtain comparable results, it was decided to choose text samples in English and distribute them to all of the project partners. In addition to the scan samples, each partner received a test report questionnaire in which they described the different stages in their digitisation workflows.

For the evaluation of the results, 24 criteria were prepared. These criteria were based on WCAG to ensure the optimal accessibility of the documents and other best practice guidelines. The criteria are: alt-text picture, alt-text picture (chemical formula), caption, footnotes, heading 1, heading 2, heading 3, initial, different language segments, mathematical formulas (simple), mathematical formulas (advanced), OCR errors (text in Picture 4 on Scan 7), page rotation, pagination, pagination-double, picture, picture (chemical formula), primary language setting, special character, stamp removal, table, table header, table rows and text order.

A total of 23 test results from 13 partner institutions were received and analysed. These included results of automatically generated outputs (17), as well as outputs that contained additional manual corrections (6). The software packages used for testing the samples were: ABBYY FineReader, ABBYY FineReader 11, ABBY Recognition server 4, ABBY Recognition server 14, ScanGate by Treventus Mechatronics, ABBYY FineReader PDF 15 Standard, Abbyy Finereader 15 desktop version, Adobe Acrobat Pro, IRIS OCR, LIMB processing, Microsoft Office Word, Scan Tailor Advanced v1.01.16, Tesseract 5.0.0-beta-20210815-22-g386dd, Photoshop 23.2.2., Project PERO OCR and WordToEpub.

The findings showed that manual corrections could significantly improve OCR quality. However, such corrections are time consuming and the focus was therefore on automatic processing. The test results showed that the best outputs were achieved using PDF/UA as the delivery format or tagged PDF. These file formats dealt with all of the criteria better than the other file formats. However, some criteria were only partially met: either they were almost met or they were technically adequate but the content was not related. The alt-text criterion exemplified the most problems, as it is an element that currently requires human input. Formats that do not contain visual elements and support assistive technology are most suitable for the blind, such TXT or variations of Microsoft Word documents. For the partially sighted, the use of formats that are adaptable to screens and enable other modern functions for working with material are most suitable, such as variations of Microsoft Word documents, ePUB, HTML, MOBI or AZW.

Another factor that should be taken in consideration is that delivery formats that meet the needs of the blind and partially sighted also enable a better experience for users of mobile devices. The most useful formats for mobile devices are next delivery file formats: ePUB, MOBI, AZW, HTML and variations of Microsoft Word documents.

Since none of the EODOPEN partners produce audiobooks, so this aspect was not part of the analysis.

At the end of the report, some recommendations and solutions concerning delivery formats for mobile devices and for print-disabled users are presented.

7 Reference

Accessible document solutions. (s.a.). *An Introduction to PDF Tags: The key ingredients in an accessible tagged PDF*. Available at: https://accessible-docs.com/tagging-accessible-pdf/

Learn about sending documents to your Kindle library. (s.a.) Amazon. Available at: https://www.amazon.com/gp/help/customer/display.html?ref_=hp_left_v4_sib&nodeId=G5WYD9SAF7PGXRNA

Guidelines and recommendations for the provision of alternative and special formats based on the survey on special needs of users and technical requirements. (2022). EODOPEN Project Deliverable D11. Available at: https://eodopen.eu/outputs.

8 Vocabulary

ALTERNATIVE TEXT (ALT-TEXT) – Alternative text provides a textual description for non-text content (pictures, graphics, diagrams …).

ASSISTIVE TECHNOLOGIES – "… any item, piece of equipment, software program, or product system that is used to increase, maintain, or improve the functional capabilities of persons with disabilities." (Source: ATIA, https://www.atia.org/home/at-resources/what-is-at/)

DELIVERY FILE FORMAT – the final file formats accessed by the users.

DIGITAL CONVERSION – digitisation

DIGITISATION WORKFLOW – all the processes implemented during the digitisation process from image capturing, image processing, OCR production …, to the conversion of scanning file format to archival and access file formats.

EBOOK – the term eBook usually refers to born-digital publications. However, we use the term of eBook especially referring to digital publications produced as a result of digital conversion, including formats for special needs (audiobooks), which is also the aim of EODOPEN project

IMAGE CAPTURING – scanning.

IMAGE PROCESSING – "Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it." (Source: Digital Image Processing, University of Tartu, https://sisu.ut.ee/imageprocessing/book/1).

MOBILE DEVICES – were mobile or smartphones, laptops, and tablet computers.

PARTIALLY SIGHTED – "People who are partially sighted are not completely blind but are able to see very little." (Source Cambridge Dictionaire, https://dictionary.cambridge.org/dictionary/english/partially-sighted). Use for visually impaired.

PRINT DISABLED – "The term "print disabled" was coined by George Kerscher, Ph.D. around 1989 to describe persons who could not access print. He used it to refer to: A person who cannot effectively read print because of a visual, physical, perceptual, developmental, cognitive, or learning disability." (Source: https://myblindspot.org/mbs-accessibility-defined/).

PROPRIETARY FILE FORMATS – formats that rely on specific software for using and the content of the file can't be read without that software, ex. MOBI, AZW

RESPONSIVE FILE FORMAT – is a format that enables the text to adjust to any screen size.

SCREEN READER – "Screen readers perform a text to speech role, but also allow audio-only access to the menus and other features of the delivery platform" (McNaught and Alexander, 2014)

TAGGED PDF – PDF which contains tags for each page element and enables easier access to document's content with assistive technologies.

TEXT TO SPEECH – "Text to speech is a mature technology that allows text on screen to be voiced by software. (McNaught and Alexander, 2014)

VISUALLY IMPAIRED – see partially sighted.

9 Used acronyms

EBU – European Blind Union.

EODOPEN – eBooks-On-Demand-network Opening Publications for European Netizens – European project cofinanced under Creative Europe program from 2019-2023.

EODOPEN PARTNERS ACRONYMS

BNP -National Library of Portugal

CVTI SR - Slovak Centre of Scientific and Technical Information

MZK - Moravian Library

NCU - Nicolaus Copernicus University in Torun

NLE - National Library of Estonia

NLS - National Library of Sweden

NUK - National and University Library

OSZK - National Széchényi Library

UG, University of Greifswald

UIBK - University of Innsbruck

UREG - University of Regensburg

UT - University of Tartu

VKOL - Research Library Olomouc

OCR – Optical Character Recognition

WCAG – Web Content Accessibility Guidelines

Acronyms for file formats:

AZW – Amazon Word

docx - Microsoft Word Open XML Format

ePUB - electronic publication

HTML - Hyper Text Markup Language

MOBI - MOBI file format (Mobipocket eBook format)

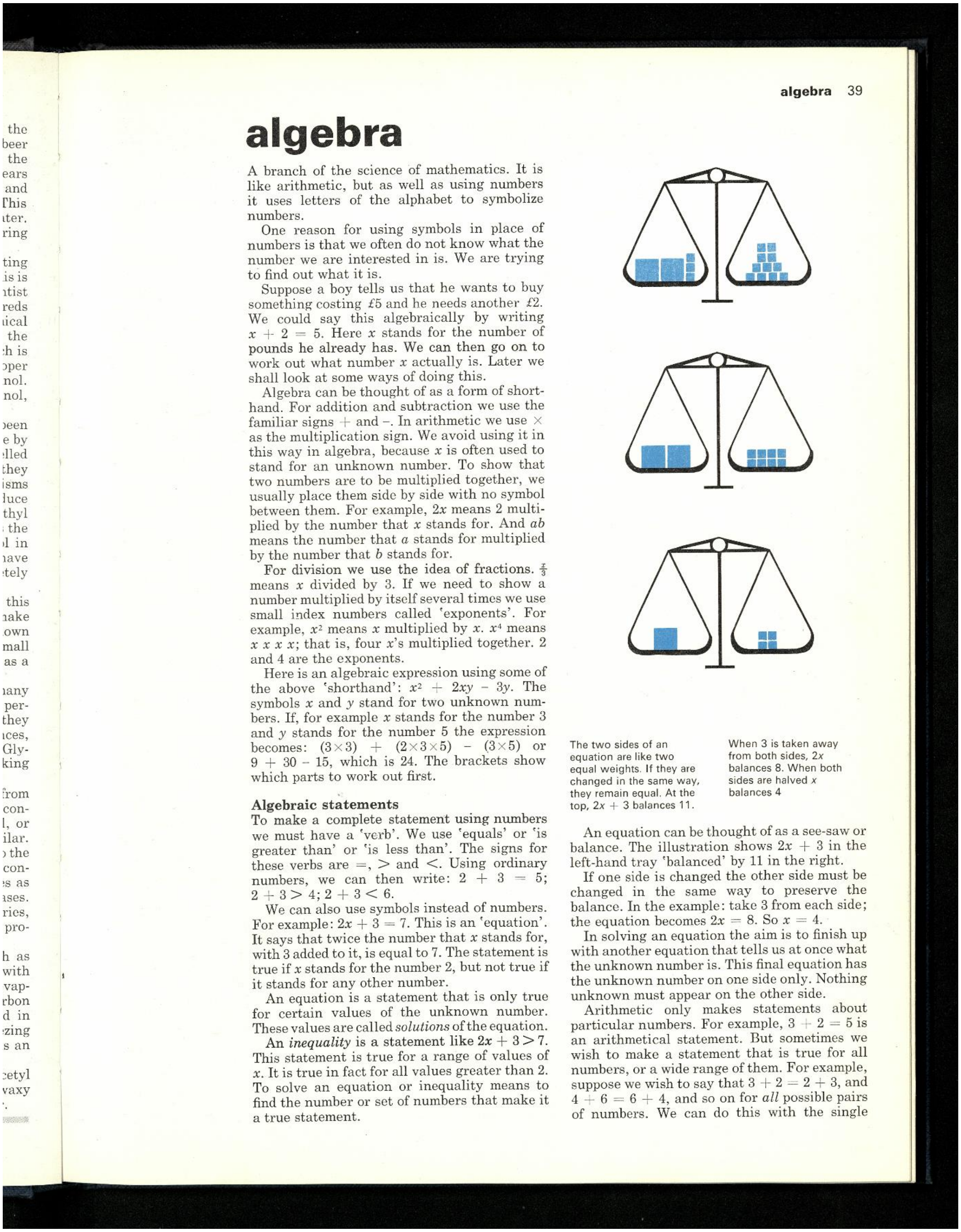
PDF – Portable Document Format

RTF – Rich Text Format

10 Annexes

Annex 1. Testing samples

Scan 1



algebra

A branch of the science of mathematics. It is like arithmetic, but as well as using numbers it uses letters of the alphabet to symbolize numbers.

One reason for using symbols in place of numbers is that we often do not know what the number we are interested in is. We are trying to find out what it is.

Suppose a boy tells us that he wants to buy something costing £5 and he needs another £2. We could say this algebraically by writing $x + 2 = 5$. Here x stands for the number of pounds he already has. We can then go on to work out what number x actually is. Later we shall look at some ways of doing this.

Algebra can be thought of as a form of shorthand. For addition and subtraction we use the familiar signs + and –. In arithmetic we use × as the multiplication sign. We avoid using it in this way in algebra, because x is often used to stand for an unknown number. To show that two numbers are to be multiplied together, we usually place them side by side with no symbol between them. For example, $2x$ means 2 multiplied by the number that x stands for. And ab means the number that a stands for multiplied by the number that b stands for.

For division we use the idea of fractions. $\frac{x}{3}$ means x divided by 3. If we need to show a number multiplied by itself several times we use small index numbers called 'exponents'. For example, x^2 means x multiplied by x. x^4 means $x\ x\ x\ x$; that is, four x's multiplied together. 2 and 4 are the exponents.

Here is an algebraic expression using some of the above 'shorthand': $x^2 + 2xy - 3y$. The symbols x and y stand for two unknown numbers. If, for example x stands for the number 3 and y stands for the number 5 the expression becomes: $(3\times3) + (2\times3\times5) - (3\times5)$ or $9 + 30 - 15$, which is 24. The brackets show which parts to work out first.

Algebraic statements

To make a complete statement using numbers we must have a 'verb'. We use 'equals' or 'is greater than' or 'is less than'. The signs for these verbs are $=$, $>$ and $<$. Using ordinary numbers, we can then write: $2 + 3 = 5$; $2 + 3 > 4$; $2 + 3 < 6$.

We can also use symbols instead of numbers. For example: $2x + 3 = 7$. This is an 'equation'. It says that twice the number that x stands for, with 3 added to it, is equal to 7. The statement is true if x stands for the number 2, but not true if it stands for any other number.

An equation is a statement that is only true for certain values of the unknown number. These values are called *solutions* of the equation.

An *inequality* is a statement like $2x + 3 > 7$. This statement is true for a range of values of x. It is true in fact for all values greater than 2. To solve an equation or inequality means to find the number or set of numbers that make it a true statement.

The two sides of an equation are like two equal weights. If they are changed in the same way, they remain equal. At the top, $2x + 3$ balances 11.

When 3 is taken away from both sides, $2x$ balances 8. When both sides are halved x balances 4

An equation can be thought of as a see-saw or balance. The illustration shows $2x + 3$ in the left-hand tray 'balanced' by 11 in the right.

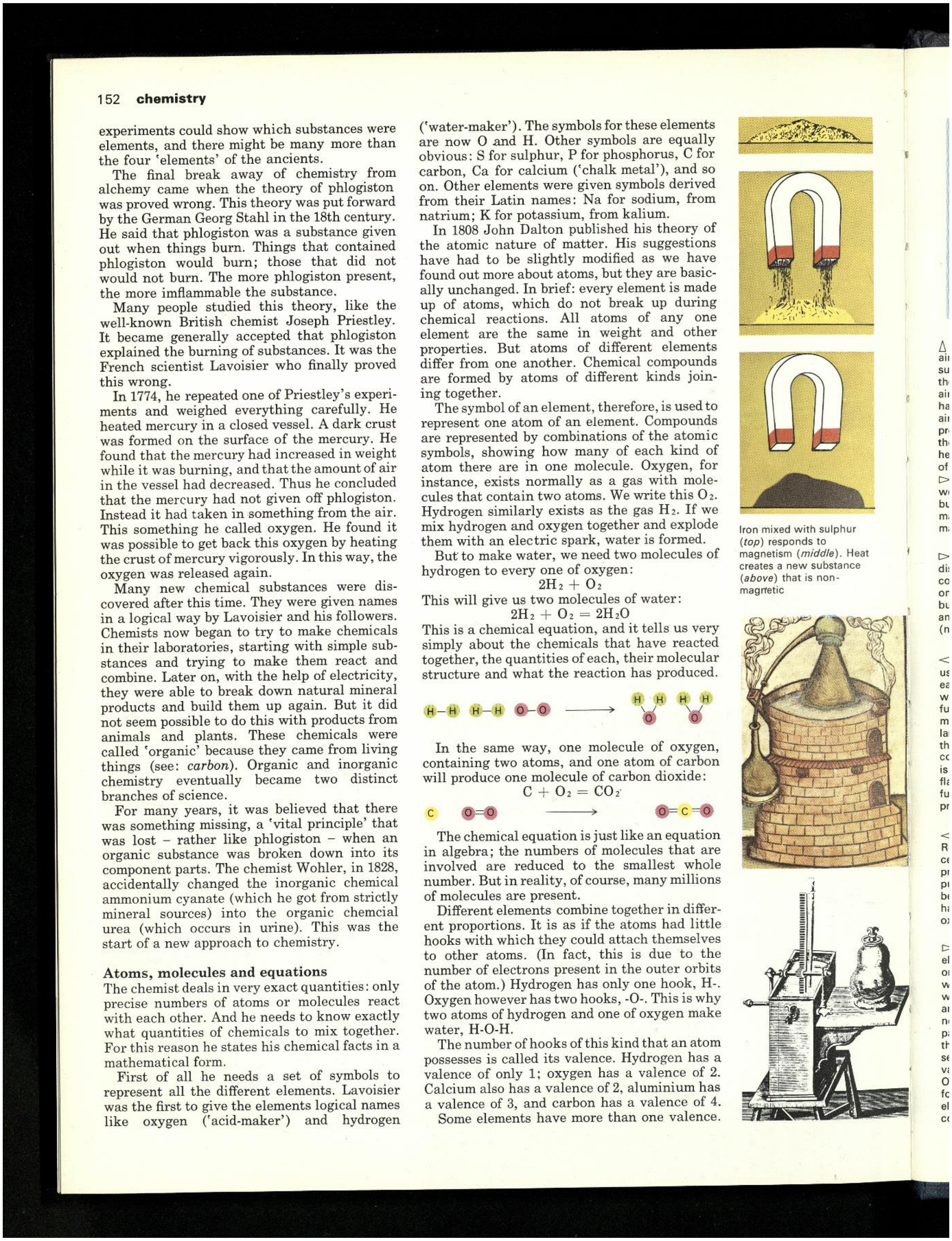
If one side is changed the other side must be changed in the same way to preserve the balance. In the example: take 3 from each side; the equation becomes $2x = 8$. So $x = 4$.

In solving an equation the aim is to finish up with another equation that tells us at once what the unknown number is. This final equation has the unknown number on one side only. Nothing unknown must appear on the other side.

Arithmetic only makes statements about particular numbers. For example, $3 + 2 = 5$ is an arithmetical statement. But sometimes we wish to make a statement that is true for all numbers, or a wide range of them. For example, suppose we wish to say that $3 + 2 = 2 + 3$, and $4 + 6 = 6 + 4$, and so on for *all* possible pairs of numbers. We can do this with the single

Co-funded by the Creative Europe Programme of the European Union

Scan 2



experiments could show which substances were elements, and there might be many more than the four 'elements' of the ancients.

The final break away of chemistry from alchemy came when the theory of phlogiston was proved wrong. This theory was put forward by the German Georg Stahl in the 18th century. He said that phlogiston was a substance given out when things burn. Things that contained phlogiston would burn; those that did not would not burn. The more phlogiston present, the more imflammable the substance.

Many people studied this theory, like the well-known British chemist Joseph Priestley. It became generally accepted that phlogiston explained the burning of substances. It was the French scientist Lavoisier who finally proved this wrong.

In 1774, he repeated one of Priestley's experiments and weighed everything carefully. He heated mercury in a closed vessel. A dark crust was formed on the surface of the mercury. He found that the mercury had increased in weight while it was burning, and that the amount of air in the vessel had decreased. Thus he concluded that the mercury had not given off phlogiston. Instead it had taken in something from the air. This something he called oxygen. He found it was possible to get back this oxygen by heating the crust of mercury vigorously. In this way, the oxygen was released again.

Many new chemical substances were discovered after this time. They were given names in a logical way by Lavoisier and his followers. Chemists now began to try to make chemicals in their laboratories, starting with simple substances and trying to make them react and combine. Later on, with the help of electricity, they were able to break down natural mineral products and build them up again. But it did not seem possible to do this with products from animals and plants. These chemicals were called 'organic' because they came from living things (see: *carbon*). Organic and inorganic chemistry eventually became two distinct branches of science.

For many years, it was believed that there was something missing, a 'vital principle' that was lost – rather like phlogiston – when an organic substance was broken down into its component parts. The chemist Wohler, in 1828, accidentally changed the inorganic chemical ammonium cyanate (which he got from strictly mineral sources) into the organic chemcial urea (which occurs in urine). This was the start of a new approach to chemistry.

Atoms, molecules and equations

The chemist deals in very exact quantities: only precise numbers of atoms or molecules react with each other. And he needs to know exactly what quantities of chemicals to mix together. For this reason he states his chemical facts in a mathematical form.

First of all he needs a set of symbols to represent all the different elements. Lavoisier was the first to give the elements logical names like oxygen ('acid-maker') and hydrogen ('water-maker'). The symbols for these elements are now O and H. Other symbols are equally obvious: S for sulphur, P for phosphorus, C for carbon, Ca for calcium ('chalk metal'), and so on. Other elements were given symbols derived from their Latin names: Na for sodium, from natrium; K for potassium, from kalium.

In 1808 John Dalton published his theory of the atomic nature of matter. His suggestions have had to be slightly modified as we have found out more about atoms, but they are basically unchanged. In brief: every element is made up of atoms, which do not break up during chemical reactions. All atoms of any one element are the same in weight and other properties. But atoms of different elements differ from one another. Chemical compounds are formed by atoms of different kinds joining together.

The symbol of an element, therefore, is used to represent one atom of an element. Compounds are represented by combinations of the atomic symbols, showing how many of each kind of atom there are in one molecule. Oxygen, for instance, exists normally as a gas with molecules that contain two atoms. We write this O_2. Hydrogen similarly exists as the gas H_2. If we mix hydrogen and oxygen together and explode them with an electric spark, water is formed.

But to make water, we need two molecules of hydrogen to every one of oxygen:

$$2H_2 + O_2$$

This will give us two molecules of water:

$$2H_2 + O_2 = 2H_2O$$

This is a chemical equation, and it tells us very simply about the chemicals that have reacted together, the quantities of each, their molecular structure and what the reaction has produced.

In the same way, one molecule of oxygen, containing two atoms, and one atom of carbon will produce one molecule of carbon dioxide:

$$C + O_2 = CO_2$$

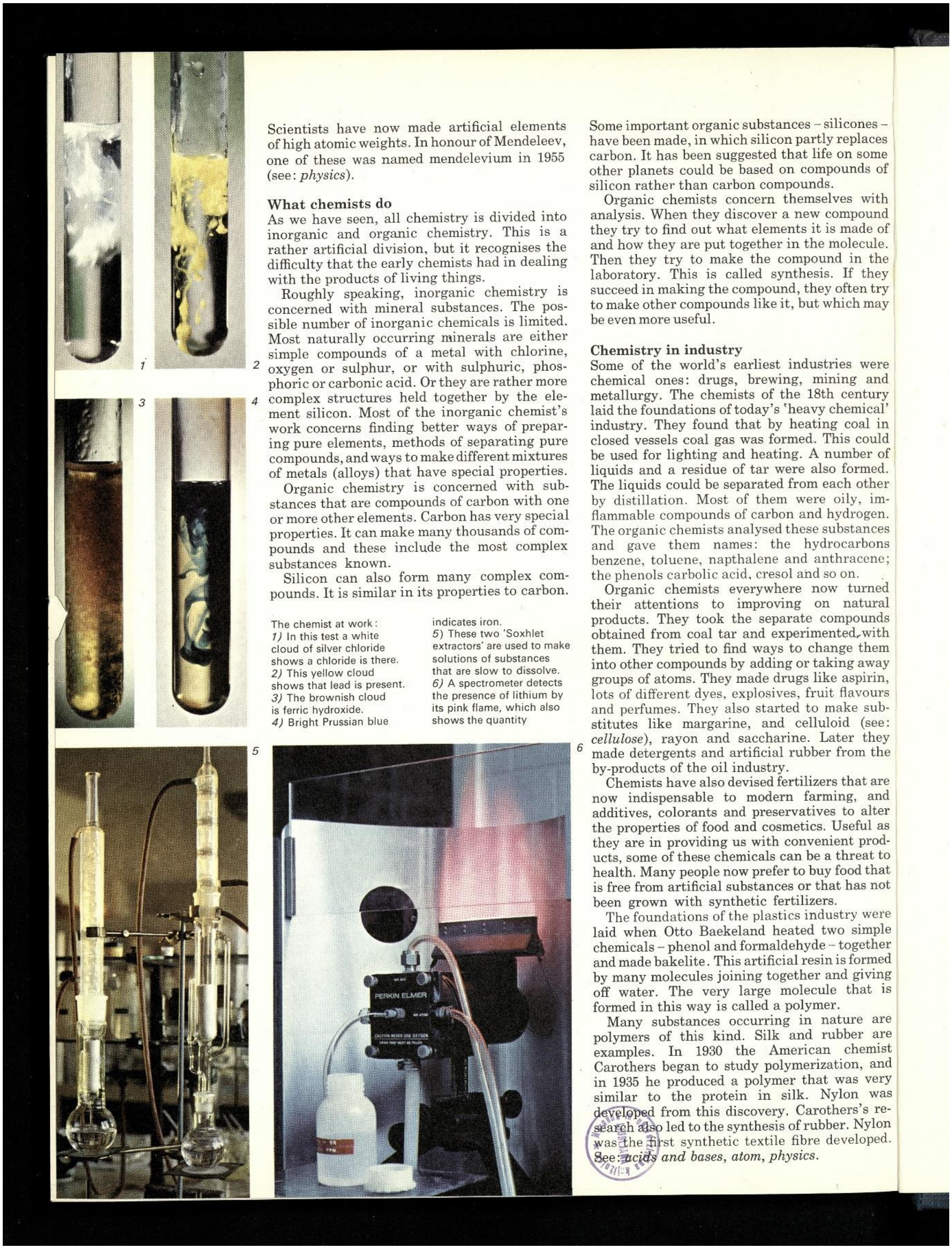
The chemical equation is just like an equation in algebra; the numbers of molecules that are involved are reduced to the smallest whole number. But in reality, of course, many millions of molecules are present.

Different elements combine together in different proportions. It is as if the atoms had little hooks with which they could attach themselves to other atoms. (In fact, this is due to the number of electrons present in the outer orbits of the atom.) Hydrogen has only one hook, H-. Oxygen however has two hooks, -O-. This is why two atoms of hydrogen and one of oxygen make water, H-O-H.

The number of hooks of this kind that an atom possesses is called its valence. Hydrogen has a valence of only 1; oxygen has a valence of 2. Calcium also has a valence of 2, aluminium has a valence of 3, and carbon has a valence of 4. Some elements have more than one valence.

Iron mixed with sulphur (*top*) responds to magnetism (*middle*). Heat creates a new substance (*above*) that is non-magnetic

Scan 3


1
2
3
4
5
6
PERKIN ELMER
CAUTION NEVER USE OXYGEN

Scientists have now made artificial elements of high atomic weights. In honour of Mendeleev, one of these was named mendelevium in 1955 (see: *physics*).

What chemists do

As we have seen, all chemistry is divided into inorganic and organic chemistry. This is a rather artificial division, but it recognises the difficulty that the early chemists had in dealing with the products of living things.

Roughly speaking, inorganic chemistry is concerned with mineral substances. The possible number of inorganic chemicals is limited. Most naturally occurring minerals are either simple compounds of a metal with chlorine, oxygen or sulphur, or with sulphuric, phosphoric or carbonic acid. Or they are rather more complex structures held together by the element silicon. Most of the inorganic chemist's work concerns finding better ways of preparing pure elements, methods of separating pure compounds, and ways to make different mixtures of metals (alloys) that have special properties.

Organic chemistry is concerned with substances that are compounds of carbon with one or more other elements. Carbon has very special properties. It can make many thousands of compounds and these include the most complex substances known.

Silicon can also form many complex compounds. It is similar in its properties to carbon.

The chemist at work:
1) In this test a white cloud of silver chloride shows a chloride is there.
2) This yellow cloud shows that lead is present.
3) The brownish cloud is ferric hydroxide.
4) Bright Prussian blue indicates iron.
5) These two 'Soxhlet extractors' are used to make solutions of substances that are slow to dissolve.
6) A spectrometer detects the presence of lithium by its pink flame, which also shows the quantity

Some important organic substances – silicones – have been made, in which silicon partly replaces carbon. It has been suggested that life on some other planets could be based on compounds of silicon rather than carbon compounds.

Organic chemists concern themselves with analysis. When they discover a new compound they try to find out what elements it is made of and how they are put together in the molecule. Then they try to make the compound in the laboratory. This is called synthesis. If they succeed in making the compound, they often try to make other compounds like it, but which may be even more useful.

Chemistry in industry

Some of the world's earliest industries were chemical ones: drugs, brewing, mining and metallurgy. The chemists of the 18th century laid the foundations of today's 'heavy chemical' industry. They found that by heating coal in closed vessels coal gas was formed. This could be used for lighting and heating. A number of liquids and a residue of tar were also formed. The liquids could be separated from each other by distillation. Most of them were oily, inflammable compounds of carbon and hydrogen. The organic chemists analysed these substances and gave them names: the hydrocarbons benzene, toluene, napthalene and anthracene; the phenols carbolic acid, cresol and so on.

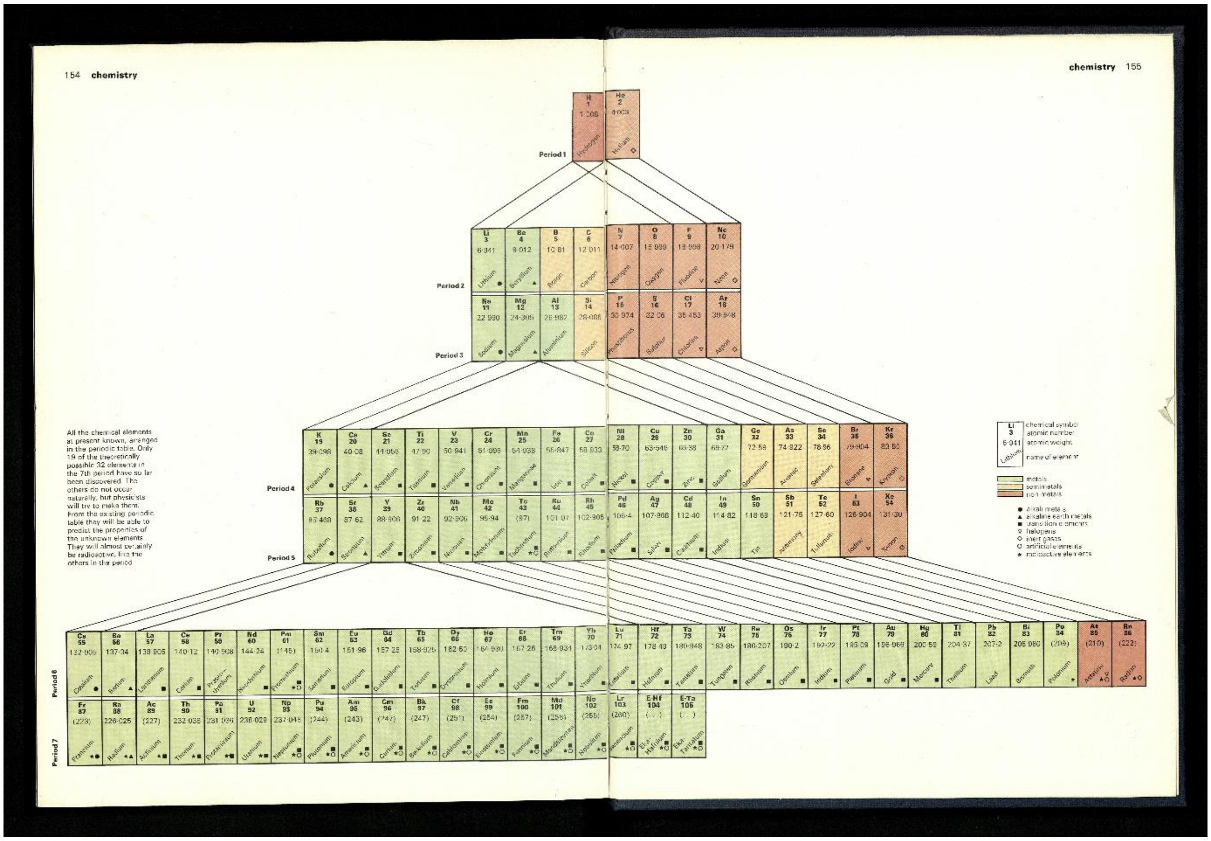
Organic chemists everywhere now turned their attentions to improving on natural products. They took the separate compounds obtained from coal tar and experimented, with them. They tried to find ways to change them into other compounds by adding or taking away groups of atoms. They made drugs like aspirin, lots of different dyes, explosives, fruit flavours and perfumes. They also started to make substitutes like margarine, and celluloid (see: *cellulose*), rayon and saccharine. Later they made detergents and artificial rubber from the by-products of the oil industry.

Chemists have also devised fertilizers that are now indispensable to modern farming, and additives, colorants and preservatives to alter the properties of food and cosmetics. Useful as they are in providing us with convenient products, some of these chemicals can be a threat to health. Many people now prefer to buy food that is free from artificial substances or that has not been grown with synthetic fertilizers.

The foundations of the plastics industry were laid when Otto Baekeland heated two simple chemicals – phenol and formaldehyde – together and made bakelite. This artificial resin is formed by many molecules joining together and giving off water. The very large molecule that is 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.*

Scan 4



All the chemical elements at present known, arranged in the periodic table. Only 19 of the theoretically possible 32 elements in the 7th period have so far been discovered. The others do not occur naturally, but physicists will try to make them. From the existing periodic table they will be able to predict the properties of the unknown elements. They will almost certainly be radioactive, like the others in the period

Period 1	H 1 1·008 Hydrogen	He 2 4·003 Helium ◇

Period 2	Li 3 6·941 Lithium ●	Be 4 9·012 Beryllium ▲	B 5 10·81 Boron	C 6 12·011 Carbon	N 7 14·007 Nitrogen	O 8 15·999 Oxygen	F 9 18·998 Fluorine ▽	Ne 10 20·179 Neon ◇
Period 3	Na 11 22·990 Sodium ●	Mg 12 24·305 Magnesium ▲	Al 13 26·982 Aluminium	Si 14 28·086 Silicon	P 15 30·974 Phosphorus	S 16 32·06 Sulphur	Cl 17 35·453 Chlorine ▽	Ar 18 39·948 Argon ◇

Period 4	K 19 39·098 Potassium ●	Ca 20 40·08 Calcium ▲	Sc 21 44·956 Scandium ■	Ti 22 47·90 Titanium ■	V 23 50·941 Vanadium ■	Cr 24 51·996 Chromium ■	Mn 25 54·938 Manganese ■	Fe 26 55·847 Iron ■	Co 27 58·933 Cobalt ■	Ni 28 58·70 Nickel ■	Cu 29 63·546 Copper ■	Zn 30 65·38 Zinc ■	Ga 31 69·72 Gallium	Ge 32 72·59 Germanium	As 33 74·922 Arsenic	Se 34 78·96 Selenium	Br 35 79·904 Bromine ▽	Kr 36 83·80 Krypton ◇
Period 5	Rb 37 85·468 Rubidium ●	Sr 38 87·62 Strontium ▲	Y 39 88·906 Yttrium ■	Zr 40 91·22 Zirconium ■	Nb 41 92·906 Niobium ■	Mo 42 95·94 Molybdenum ■	Tc 43 (97) Technetium ■ ★○	Ru 44 101·07 Ruthenium ■	Rh 45 102·905 Rhodium ■	Pd 46 106·4 Palladium ■	Ag 47 107·868 Silver ■	Cd 48 112·40 Cadmium ■	In 49 114·82 Indium	Sn 50 118·69 Tin	Sb 51 121·75 Antimony	Te 52 127·60 Tellurium	I 53 126·904 Iodine ▽	Xe 54 131·30 Xenon ◇

Period 6	Cs 55 132·905 Caesium ●	Ba 56 137·34 Barium ▲	La 57 138·906 Lanthanum ■	Ce 58 140·12 Cerium ■	Pr 59 140·908 Praseo-dymium ■	Nd 60 144·24 Neodymium ■	Pm 61 (145) Promethium ■ ★○	Sm 62 150·4 Samarium ■	Eu 63 151·96 Europium ■	Gd 64 157·25 Gadolinium ■	Tb 65 158·925 Terbium ■	Dy 66 162·50 Dysprosium ■	Ho 67 164·930 Holmium ■	Er 68 167·26 Erbium ■	Tm 69 168·934 Thulium ■	Yb 70 173·04 Ytterbium ■	Lu 71 174·97 Lutetium ■	Hf 72 178·49 Hafnium ■	Ta 73 180·948 Tantalum ■	W 74 183·85 Tungsten ■	Re 75 186·207 Rhenium ■	Os 76 190·2 Osmium ■	Ir 77 192·22 Iridium ■	Pt 78 195·09 Platinum ■	Au 79 196·966 Gold ■	Hg 80 200·59 Mercury ■	Tl 81 204·37 Thallium	Pb 82 207·2 Lead	Bi 83 208·980 Bismuth	Po 84 (209) Polonium ★	At 85 (210) Astatine ▽ ★○	Rn 86 (222) Radon ★◇
Period 7	Fr 87 (223) Francium ★●	Ra 88 226·025 Radium ★▲	Ac 89 (227) Actinium ★■	Th 90 232·038 Thorium ★■	Pa 91 231·036 Protactinium ★■	U 92 238·029 Uranium ★■	Np 93 237·048 Neptunium ■ ★○	Pu 94 (244) Plutonium ■ ★○	Am 95 (243) Americium ■ ★○	Cm 96 (247) Curium ■ ★○	Bk 97 (247) Berkelium ■ ★○	Cf 98 (251) Californium ■ ★○	Es 99 (254) Einsteinium ■ ★○	Fm 100 (257) Fermium ■ ★○	Md 101 (258) Mendelevium ■ ★○	No 102 (255) Nobelium ■ ★○	Lr 103 (260) Lawrencium ■ ★○	E-Hf 104 ( ) Eka-Hafnium ■ ★○	E-Ta 105 ( ) Eka-Tantalum ■ ★○													

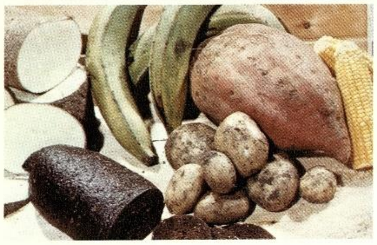
Li	chemical symbol
3	atomic number
6·941	atomic weight
Lithium	name of element

🟩	metals
🟨	semimetals
🟥	non-metals

●	alkali metals
▲	alkaline earth metals
■	transition elements
▽	halogens
◇	inert gases
○	artificial elements
★	radioactive elements

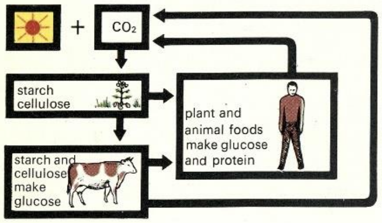
Scan 5

carbohydrate



△ Basic foods, like yams, potatoes, bread, bananas, corn, and sweet potatoes, are rich in carbohydrates

▽ Carbohydrates are made by plants from light, water and CO_2. Man gets energy by eating plants and animals



What do a sugar lump, a loaf of bread and a potato have in common? They all contain carbohydrates. The sugar lump is made entirely of one of the simpler types of carbohydrate, which are all called sugars. The bread is made from flour, which contains a more complicated carbohydrate, starch. The potato also contains starch, food stored by the plant to feed its new shoots in the spring.

There are many types of sugar. Plants make sugars by the process of photosynthesis (see: *photosynthesis*). Plant sap is full of sugar. Both plants and animals change these sugars into one type, glucose, which they use for energy. Plants store glucose by combining many molecules of it to form into a large starch molecule. Starch does not dissolve in water and is too large to enter or leave a cell. When the plant needs the stored energy, it breaks down the large molecule into small glucose molecules and uses them. Glucose is a small molecule that dissolves in water, and passes easily into and out of the cell. Carbohydrates are the main source of energy in our diet. When we eat stored plant starch in flour or potatoes, our bodies too break it down into glucose.

The name carbohydrate refers to the chemical composition of these compounds. The formula shows that for each carbon atom (C) there are two hydrogen atoms (H) and one oxygen atom (O), as in water (H_2O). For instance, a sugar that contains six carbon atoms usually has enough hydrogen and oxygen atoms to make six molecules of water: $C_6H_{12}O_6$.

Much of the sugar we eat is made from the sweet sap of sugar cane. This sap tastes good when it is fresh. It is boiled down to concentrate its sugar content into syrup or crystals. Sugar cane grows only in hot countries. Here it is being harvested in Barbados

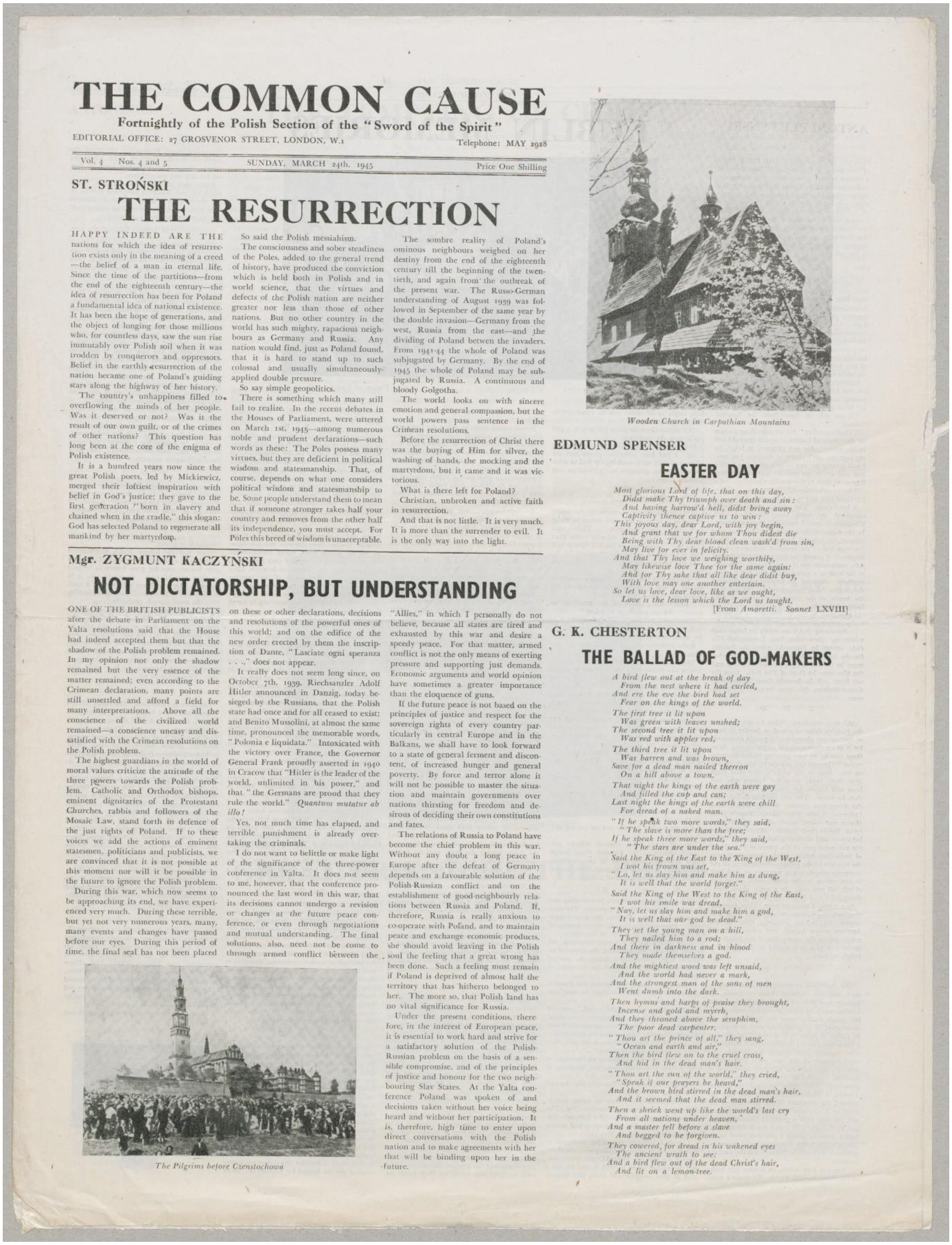
It is more accurate and useful to draw a diagram of the molecule. There are many different sugars with the formula $C_6H_{12}O_6$. Each of these sugars has a different taste and takes part in different chemical reactions, according to the arrangement of its atoms.

Sugars are the simplest carbohydrates. The simplest sugars consist of just one type of sugar molecule. Such a molecule is called a monosaccharide, that is, 'one sugar'. Usually this molecule contains from three to seven carbon atoms. Glucose, which is found in all plants, is the main sugar found in fruit. It has six carbon atoms. Ribose, a sugar that is found in the nucleus of the cell, has five carbon atoms. Two other five-carbon sugars are found in the sticky gums with which plants heal wounds in their bark. These are arabinose and xylose.

Two monosaccharide molecules can be joined together to make a double sugar, a disaccharide. We are most familiar with the disaccharide sucrose. This is the sugar we eat in our food. Sucrose is found in the sap of many plants. Maple syrup is mostly sucrose. The commercial sugar industry makes its brown or white crystals of sucrose from sugar cane and sugar beets. In warm climates these plants are fairly easy to grow in large quantities and yield plenty of sucrose. Honey also contains sucrose, mixed with glucose and another sugar, fructose.

Each molecule of sucrose is made up of a molecule of glucose and one of fructose joined together. These double molecules are too large to be absorbed by our cells. When we eat

Scan 6



THE COMMON CAUSE

Fortnightly of the Polish Section of the "Sword of the Spirit"

EDITORIAL OFFICE: 27 GROSVENOR STREET, LONDON, W.1 Telephone: MAY 2928

Vol. 4 Nos. 4 and 5 SUNDAY, MARCH 24th, 1945 Price One Shilling

ST. STROŃSKI

THE RESURRECTION

HAPPY INDEED ARE THE nations for which the idea of resurrection exists only in the meaning of a creed—the belief of a man in eternal life. Since the time of the partitions—from the end of the eighteenth century—the idea of resurrection has been for Poland a fundamental idea of national existence. It has been the hope of generations, and the object of longing for those millions who, for countless days, saw the sun rise immutably over Polish soil when it was trodden by conquerors and oppressors. Belief in the earthly resurrection of the nation became one of Poland's guiding stars along the highway of her history.

The country's unhappiness filled to overflowing the minds of her people. Was it deserved or not? Was it the result of our own guilt, or of the crimes of other nations? This question has long been at the core of the enigma of Polish existence.

It is a hundred years now since the great Polish poets, led by Mickiewicz, merged their loftiest inspiration with belief in God's justice; they gave to the first generation "born in slavery and chained when in the cradle," this slogan: God has selected Poland to regenerate all mankind by her martyrdom.

So said the Polish messiahism.

The consciousness and sober steadiness of the Poles, added to the general trend of history, have produced the conviction which is held both in Polish and in world science, that the virtues and defects of the Polish nation are neither greater nor less than those of other nations. But no other country in the world has such mighty, rapacious neighbours as Germany and Russia. Any nation would find, just as Poland found, that it is hard to stand up to such colossal and usually simultaneously-applied double pressure.

So say simple geopolitics.

There is something which many still fail to realize. In the recent debates in the Houses of Parliament, were uttered on March 1st, 1945—among numerous noble and prudent declarations—such words as these: The Poles possess many virtues, but they are deficient in political wisdom and statesmanship. That, of course, depends on what one considers political wisdom and statesmanship to be. Some people understand them to mean that if someone stronger takes half your country and removes from the other half its independence, you must accept. For Poles this breed of wisdom is unacceptable.

The sombre reality of Poland's ominous neighbours weighed on her destiny from the end of the eighteenth century till the beginning of the twentieth, and again from the outbreak of the present war. The Russo-German understanding of August 1939 was followed in September of the same year by the double invasion—Germany from the west, Russia from the east—and the dividing of Poland betwen the invaders. From 1941-44 the whole of Poland was subjugated by Germany. By the end of 1945 the whole of Poland may be subjugated by Russia. A continuous and bloody Golgotha.

The world looks on with sincere emotion and general compassion, but the world powers pass sentence in the Crimean resolutions.

Before the resurrection of Christ there was the buying of Him for silver, the washing of hands, the mocking and the martyrdom, but it came and it was victorious.

What is there left for Poland?

Christian, unbroken and active faith in resurrection.

And that is not little. It is very much. It is more than the surrender to evil. It is the only way into the light.

Mgr. ZYGMUNT KACZYŃSKI

NOT DICTATORSHIP, BUT UNDERSTANDING

ONE OF THE BRITISH PUBLICISTS after the debate in Parliament on the Yalta resolutions said that the House had indeed accepted them but that the shadow of the Polish problem remained. In my opinion not only the shadow remained but the very essence of the matter remained; even according to the Crimean declaration, many points are still unsettled and afford a field for many interpretations. Above all the conscience of the civilized world remained—a conscience uneasy and dissatisfied with the Crimean resolutions on the Polish problem.

The highest guardians in the world of moral values criticize the attitude of the three powers towards the Polish problem. Catholic and Orthodox bishops, eminent dignitaries of the Protestant Churches, rabbis and followers of the Mosaic Law, stand forth in defence of the just rights of Poland. If to these voices we add the actions of eminent statesmen, politicians and publicists, we are convinced that it is not possible at this moment nor will it be possible in the future to ignore the Polish problem.

During this war, which now seems to be approaching its end, we have experienced very much. During these terrible, but yet not very numerous years, many, many events and changes have passed before our eyes. During this period of time, the final seal has not been placed on these or other declarations, decisions and resolutions of the powerful ones of this world; and on the edifice of the new order erected by them the inscription of Dante, "Lasciate ogni speranza . . .," does not appear.

It really does not seem long since, on October 7th, 1939, Riechsanzler Adolf Hitler announced in Danzig, today besieged by the Russians, that the Polish state had once and for all ceased to exist; and Benito Mussolini, at almost the same time, pronounced the memorable words, "Polonia e liquidata." Intoxicated with the victory over France, the Governor General Frank proudly asserted in 1940 in Cracow that "Hitler is the leader of the world, unlimited in his power," and that "the Germans are proud that they rule the world." *Quantum mutatur ab illo!*

Yes, not much time has elapsed, and terrible punishment is already overtaking the criminals.

I do not want to belittle or make light of the significance of the three-power conference in Yalta. It does not seem to me, however, that the conference pronounced the last word in this war, that its decisions cannot undergo a revision or changes at the future peace conference, or even through negotiations and mutual understanding. The final solutions, also, need not be come to through armed conflict between the "Allies," in which I personally do not believe, because all states are tired and exhausted by this war and desire a speedy peace. For that matter, armed conflict is not the only means of exerting pressure and supporting just demands. Economic arguments and world opinion have sometimes a greater importance than the eloquence of guns.

If the future peace is not based on the principles of justice and respect for the sovereign rights of every country particularly in central Europe and in the Balkans, we shall have to look forward to a state of general ferment and discontent, of increased hunger and general poverty. By force and terror alone it will not be possible to master the situation and maintain governments over nations thirsting for freedom and desirous of deciding their own constitutions and fates.

The relations of Russia to Poland have become the chief problem in this war. Without any doubt a long peace in Europe after the defeat of Germany depends on a favourable solution of the Polish-Russian conflict and on the establishment of good-neighbourly relations between Russia and Poland. If, therefore, Russia is really anxious to co-operate with Poland, and to maintain peace and exchange economic products, she should avoid leaving in the Polish soul the feeling that a great wrong has been done. Such a feeling must remain if Poland is deprived of almost half the territory that has hitherto belonged to her. The more so, that Polish land has no vital significance for Russia.

Under the present conditions, therefore, in the interest of European peace, it is essential to work hard and strive for a satisfactory solution of the Polish-Russian problem on the basis of a sensible compromise, and of the principles of justice and honour for the two neighbouring Slav States. At the Yalta conference Poland was spoken of and decisions taken without her voice being heard and without her participation. It is, therefore, high time to enter upon direct conversations with the Polish nation and to make agreements with her that will be binding upon her in the future.

The Pilgrims before Czenstochowa

Wooden Church in Carpathian Mountains

EDMUND SPENSER

EASTER DAY

Most glorious Lord of life, that on this day,
Didst make Thy triumph over death and sin:
And having harrow'd hell, didst bring away
Captivity thence captive us to win:
This joyous day, dear Lord, with joy begin,
And grant that we for whom Thou didest die
Being with Thy dear blood clean wash'd from sin,
May live for ever in felicity.
And that Thy love we weighing worthily,
May likewise love Thee for the same again:
And for Thy sake that all like dear didst buy,
With love may one another entertain.
So let us love, dear love, like as we ought,
Love is the lesson which the Lord us taught.

[From *Amoretti. Sonnet* LXVIII]

G. K. CHESTERTON

THE BALLAD OF GOD-MAKERS

A bird flew out at the break of day
From the nest where it had curled,
And ere the eve the bird had set
Fear on the kings of the world.

The first tree it lit upon
Was green with leaves unshed;
The second tree it lit upon
Was red with apples red,

The third tree it lit upon
Was barren and was brown,
Save for a dead man nailed thereon
On a hill above a town.

That night the kings of the earth were gay
And filled the cup and can;
Last night the kings of the earth were chill
For dread of a naked man.

"If he speak two more words," they said,
"The slave is more than the free;
If he speak three more words," they said,
"The stars are under the sea."

Said the King of the East to the King of the West,
I wot his frown was set,
"Lo, let us slay him and make him as dung,
It is well that the world forget."

Said the King of the West to the King of the East,
I wot his smile was dread,
"Nay, let us slay him and make him a god,
It is well that our god be dead."

They set the young man on a hill,
They nailed him to a rod;
And there in darkness and in blood
They made themselves a god.

And the mightiest word was left unsaid,
And the world had never a mark,
And the strongest man of the sons of men
Went dumb into the dark.

Then hymns and harps of praise they brought,
Incense and gold and myrrh,
And they throned above the seraphim,
The poor dead carpenter.

"Thou art the prince of all," they sang,
"Ocean and earth and air,"
Then the bird flew on to the cruel cross,
And hid in the dead man's hair.

"Thou art the sun of the world," they cried,
"Speak if our prayers be heard,"
And the brown bird stirred in the dead man's hair,
And it seemed that the dead man stirred.

Then a shriek went up like the world's last cry
From all nations under heaven,
And a master fell before a slave
And begged to be forgiven.

They cowered, for dread in his wakened eyes
The ancient wrath to see;
And a bird flew out of the dead Christ's hair,
And lit on a lemon-tree.

Scan 7

100TH ANNIVERSARY OF THE LWOW POLYTECHNIC INSTITUTE

by JERZY W. MEIER

Kazimierz Bartel, former Prime Minister and Professor at the Lwow Polytechnic, was executed by the Germans in Lwow for refusing to cooperate with them.

LONG before the war, plans to celebrate the one-hundredth anniversary of the founding of the Lwow Polytechnic Institute were made by the directors of the Institute, now in exile. They undertook the construction of seven large buildings for the Mechanical and Electrotechnical Departments, along with new dormitories for aviation students and testing laboratories for the civil engineering students. All these workshops of the school were used during the development of the Central Industrial Area of Poland.

A special anniversary book was to have been published which would have given a brief review of past achievements and a detailed report of Polish technical work and the contribution of Polish thought to the development of technical study. A series of monographs was to have been published on the work of alumni of the Lwow Polytechnic Institute and of the general Congress of Polish Engineers. This work was disrupted by Germany's aggression and the ensuing occupation. Polish professors—those who still live—are in German concentration camps, young Polish engineers and students are either in the Allied armies, have joined the Home Army, or are prisoners of war.

The beginnings of this Institute reach back to 1825, when Stanislaw Staszic became head of the Polytechnical Council called by the Polish Government to organize a Polytechnic under the direction of Kajetan Garbinski. Staszic opened this school two weeks before his death. The school progressed steadily and its program was enlarged and improved until the Uprising of 1831.

Not until 1844 was another attempt made to establish another such school under the name of "Technical Academy."

Normally, in the life of a nation, a hundred years is no great length of time, but the last century before the rebirth of Poland was a most stormy period during which political and economic events affected the development of the Lwow Technical Academy. Just as the school began to function again, its progress was interrupted by the so-called Spring of the Nations, the revolutionary uprising of 1848 that strove to break the despotism of the Holy Alliance.

The Galician massacre in the Cracow region, the Austrian occupation of hitherto free Cracow, the uprising in Poznan, the temporary freedom of Lwow and the revengeful bombardment of the city by General Hammerstein which destroyed the old University of Lwow followed. All this along with later repressions under Austrian rule developed the character of the Poles and the technical progress of their school far more than the program of education dictated by Vienna. Once again progress was disrupted by the Insurrection of 1863, in Russian-held Poland, when youths from Lwow as well as from every part of Poland crossed the border and flocked to join the colors. This time repression was not so severe, as the artificial dual monarchy was already shaken by its disastrous defeat at the hands of the Germans at Sadowa in 1866. To rescue their decaying empire, the Hapsburgs adopted a constitution and granted limited self-government to their subject peoples.

Monument to the students who fell in the defense of Lwow in 1918-1919, in the garden of the Lwow Polytechnic. The garden had been used as a first aid station during the war.

Education was reorganized by the Provincial Seym in Lwow. In 1872, the Lwow Technical Academy was changed to the Polytechnical School. A rector and senate were placed in charge of it with the right to hold state examinations and grant state diplomas in engineering. This placed the Institute on a par with the Polytechnic School at Vienna. At the same time, classes were transferred to a new building. At last the school was assured of steady growth and progress. This was the era of the pioneer engineer Stanislaw Szczepanowski and of positivist tendencies in Russian-held Poland. Technology grew in popularity, and the school gained an ever-increasing number of students. The failure of the insurrections and the proverbial poverty of bureaucratically governed Galicia further stimulated the desire to create on Polish soil technological schools equal to any found in Western Europe. The Lwow Institute thus became not only an institute of technology but also a school of Polish patriotism. Many of its students were among the first to join the ranks of Pilsudski's Legions in 1914, and fought through the World War at his side.

Independent Poland awarded the Lwow Institute the *"Polonia Restituta"* medal for its achievements in science and the outstanding part played by the school in the fight for independence. The city of Lwow gave the Institute the "Cross of the Defense of Lwow."

Following the first war, after regaining the right to rule itself, the enrollment of the Lwow Polytechnical Institute rose to more than 3,000 students, from the pre-war total of 360. Buildings, class rooms, and dormitories were inadequate. The Institute, however, was fortunate in obtaining the buildings of the Magdalena Convent.

In 1937, construction of new buildings for the Department of Mechanical Engineering began on land donated by the city of Lwow. Seven buildings were to have been erected for the new departments of theoretic and practical engineering, metallurgical technology and the use of scrap materials, industrial and metallurgical laboratories, etc. The first two of these projected buildings were in process of completion and their equipment was ready for installation at the outbreak of the war.

The organization and program of the Polytechnic Institute underwent many changes after the rebirth of Poland. In 1919, two other independent educational institutions, the Agricultural Academy in Dublany and the Forestry Academy in Lwow were absorbed and made into the Department of Agriculture and Forestry. To the older departments of architecture, civil and marine engineering, mechanical engineering, and chemistry was also added the department of military and general engineering. The entire program was brought up to date and enlarged.

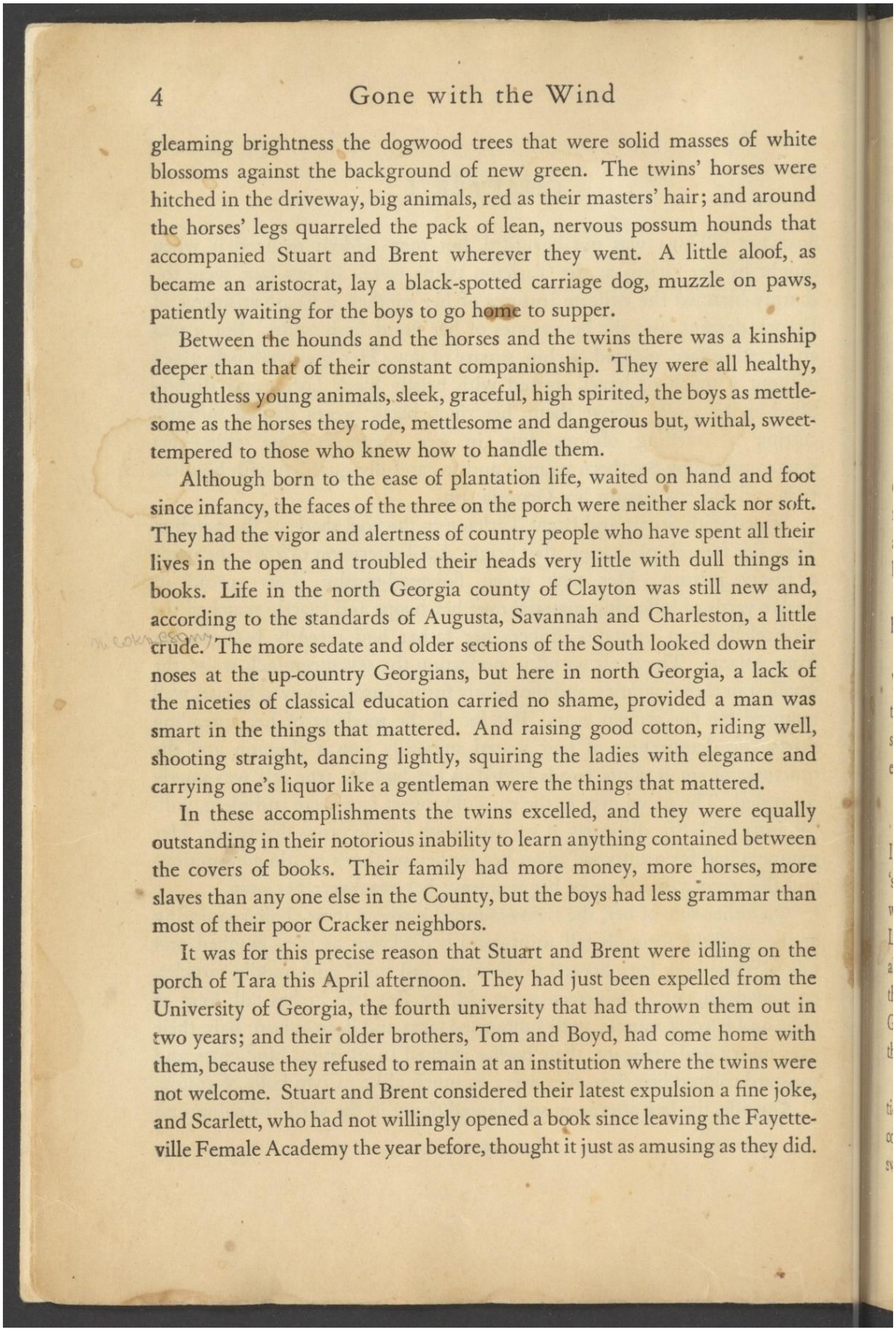
At the beginning of the 20th century, the entire department of mechanical engineering was in the hands of three professors, Franke, Jaksa-Bykowski, and Maryniak. Later they were replaced by Professors Fiedler

(*Turn to p.* 14)

Lwow Polytechnic Institute.

Memorial tablet to "Students of the Lwow Polytechnic who died in the defense of Lwow and their Country in the years 1918-1921."

Scan 8



gleaming brightness the dogwood trees that were solid masses of white blossoms against the background of new green. The twins' horses were hitched in the driveway, big animals, red as their masters' hair; and around the horses' legs quarreled the pack of lean, nervous possum hounds that accompanied Stuart and Brent wherever they went. A little aloof, as became an aristocrat, lay a black-spotted carriage dog, muzzle on paws, patiently waiting for the boys to go home to supper.

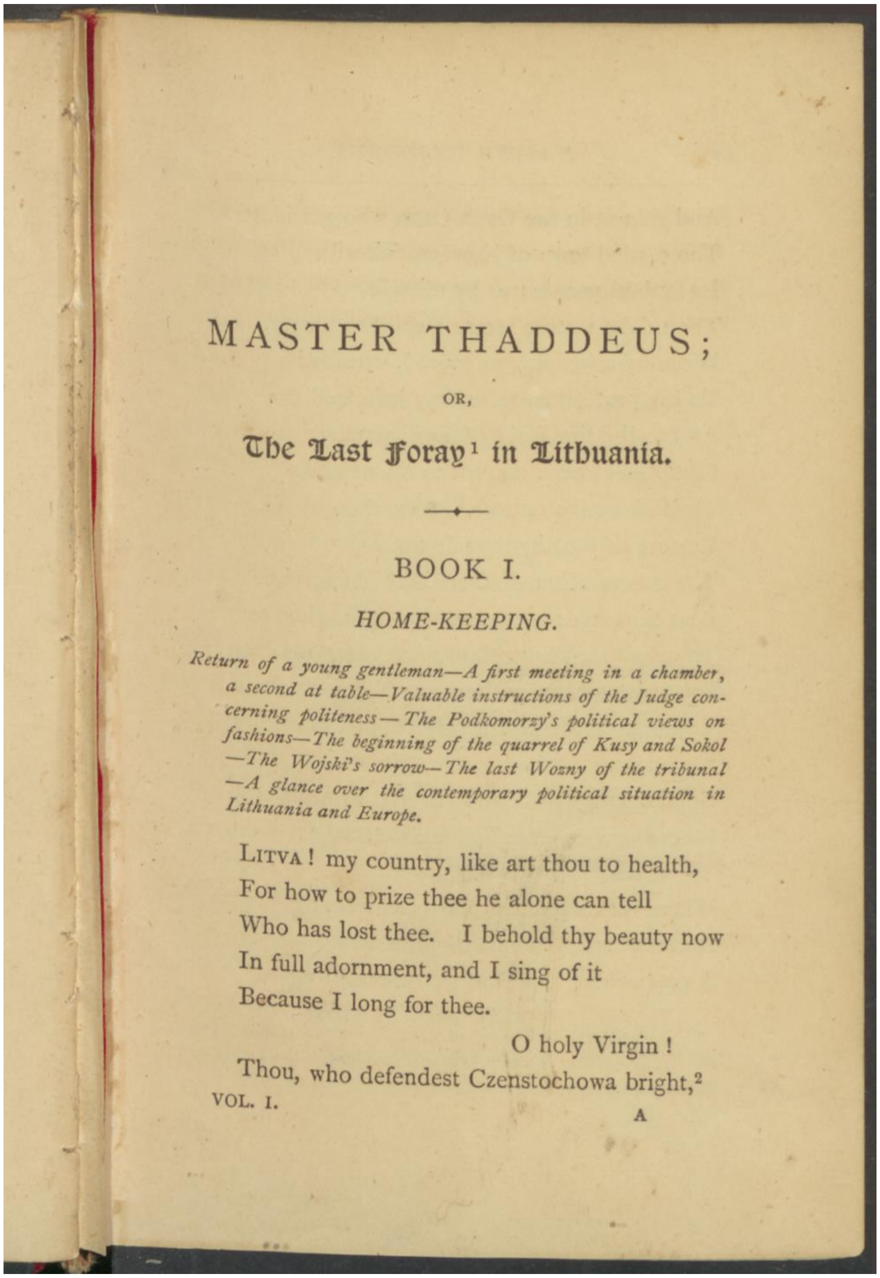
Between the hounds and the horses and the twins there was a kinship deeper than that of their constant companionship. They were all healthy, thoughtless young animals, sleek, graceful, high spirited, the boys as mettlesome as the horses they rode, mettlesome and dangerous but, withal, sweet-tempered to those who knew how to handle them.

Although born to the ease of plantation life, waited on hand and foot since infancy, the faces of the three on the porch were neither slack nor soft. They had the vigor and alertness of country people who have spent all their lives in the open and troubled their heads very little with dull things in books. Life in the north Georgia county of Clayton was still new and, according to the standards of Augusta, Savannah and Charleston, a little crude. The more sedate and older sections of the South looked down their noses at the up-country Georgians, but here in north Georgia, a lack of the niceties of classical education carried no shame, provided a man was smart in the things that mattered. And raising good cotton, riding well, shooting straight, dancing lightly, squiring the ladies with elegance and carrying one's liquor like a gentleman were the things that mattered.

In these accomplishments the twins excelled, and they were equally outstanding in their notorious inability to learn anything contained between the covers of books. Their family had more money, more horses, more slaves than any one else in the County, but the boys had less grammar than most of their poor Cracker neighbors.

It was for this precise reason that Stuart and Brent were idling on the porch of Tara this April afternoon. They had just been expelled from the University of Georgia, the fourth university that had thrown them out in two years; and their older brothers, Tom and Boyd, had come home with them, because they refused to remain at an institution where the twins were not welcome. Stuart and Brent considered their latest expulsion a fine joke, and Scarlett, who had not willingly opened a book since leaving the Fayetteville Female Academy the year before, thought it just as amusing as they did.

Scan 9



MASTER THADDEUS;

OR,

The Last Foray[1] in Lithuania.

BOOK I.

HOME-KEEPING.

Return of a young gentleman—A first meeting in a chamber, a second at table—Valuable instructions of the Judge concerning politeness—The Podkomorzy's political views on fashions—The beginning of the quarrel of Kusy and Sokol—The Wojski's sorrow—The last Wozny of the tribunal—A glance over the contemporary political situation in Lithuania and Europe.

LITVA! my country, like art thou to health,
For how to prize thee he alone can tell
Who has lost thee. I behold thy beauty now
In full adornment, and I sing of it
Because I long for thee.
O holy Virgin!
Thou, who defendest Czenstochowa bright,[2]

Scan 10

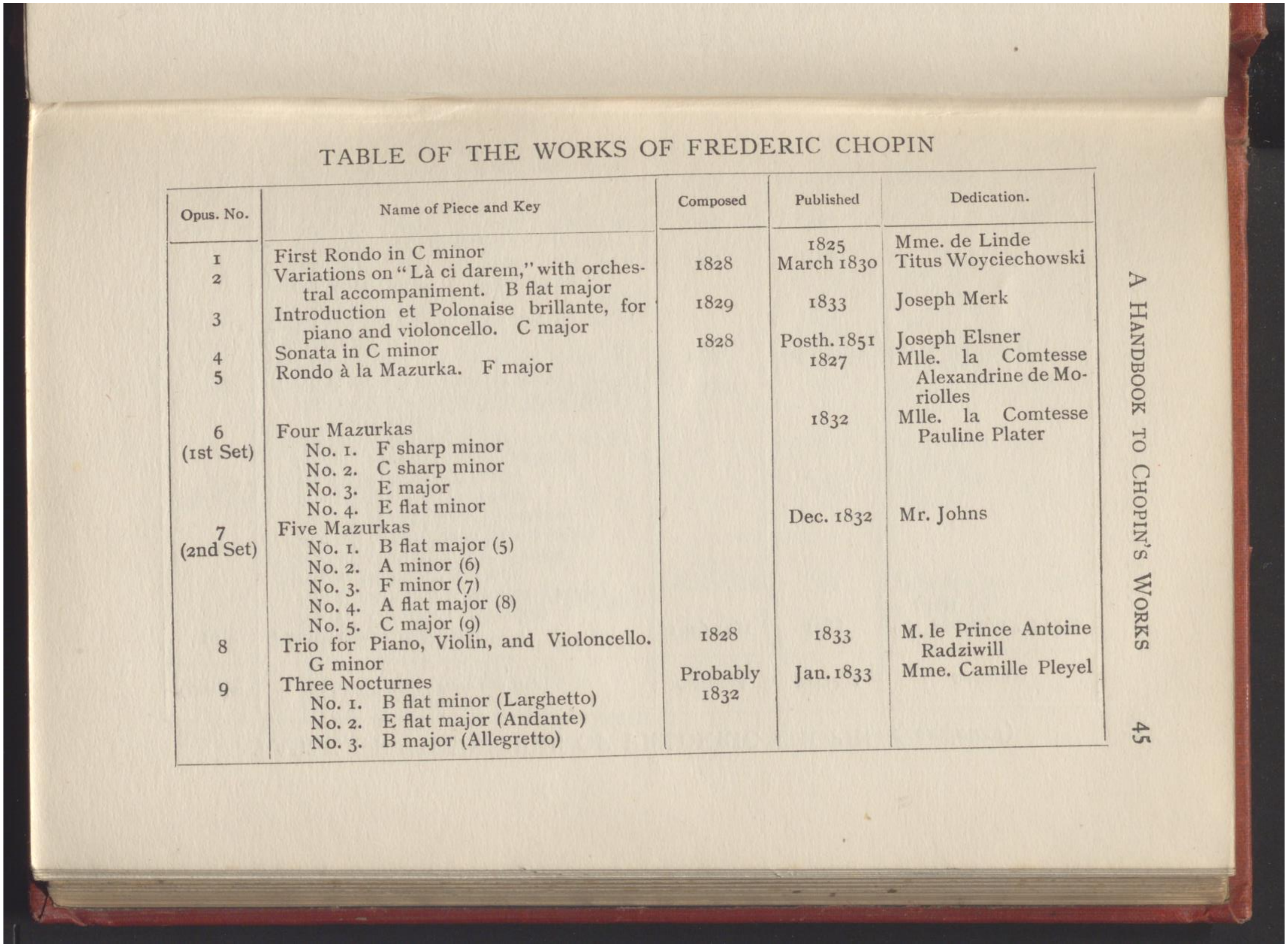
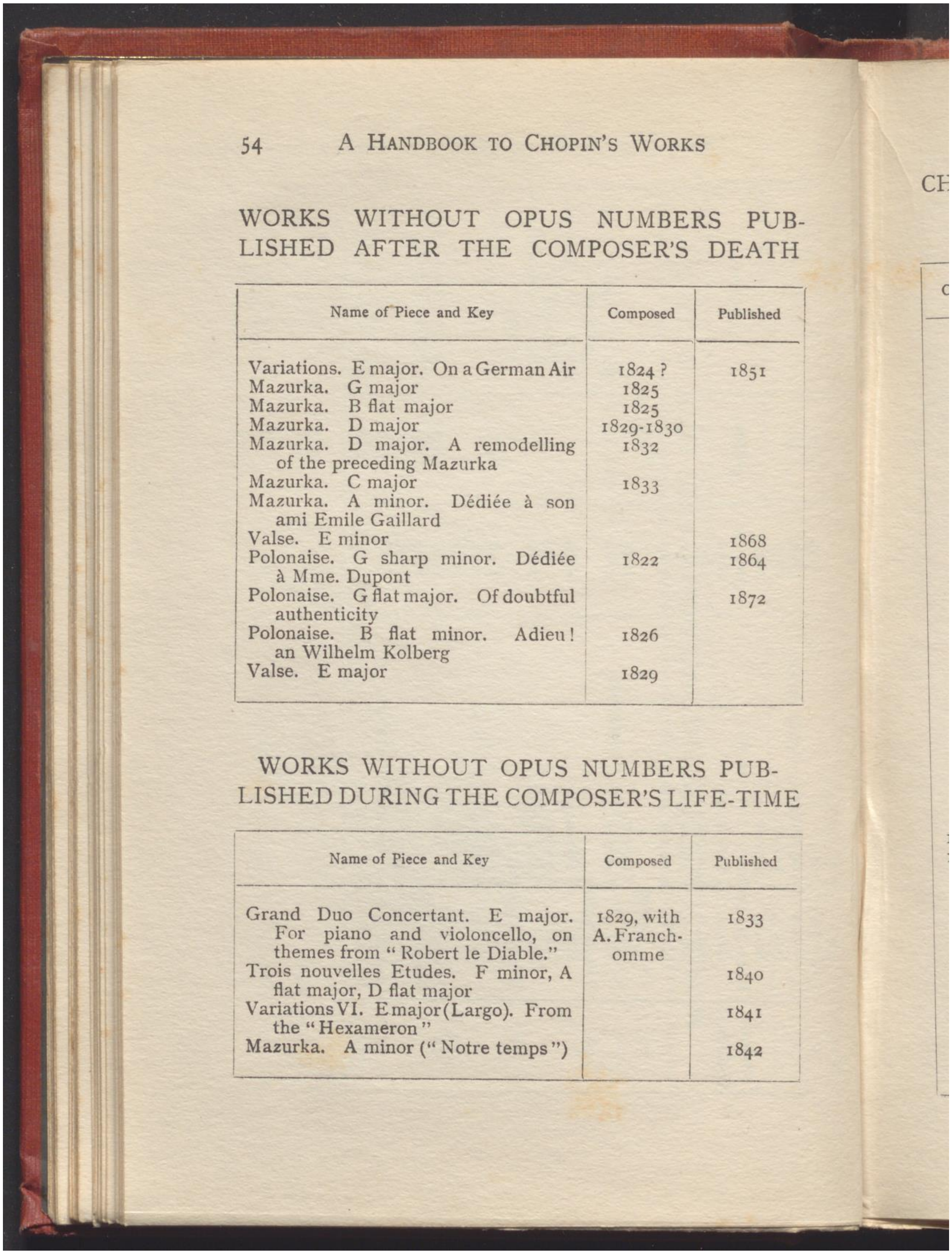


TABLE OF THE WORKS OF FREDERIC CHOPIN

Opus. No.	Name of Piece and Key	Composed	Published	Dedication.
1	First Rondo in C minor		1825	Mme. de Linde
2	Variations on "Là ci darem," with orchestral accompaniment. B flat major	1828	March 1830	Titus Woyciechowski
3	Introduction et Polonaise brillante, for piano and violoncello. C major	1829	1833	Joseph Merk
4	Sonata in C minor	1828	Posth. 1851	Joseph Elsner
5	Rondo à la Mazurka. F major		1827	Mlle. la Comtesse Alexandrine de Moriolles
6 (1st Set)	Four Mazurkas No. 1. F sharp minor No. 2. C sharp minor No. 3. E major No. 4. E flat minor		1832	Mlle. la Comtesse Pauline Plater
7 (2nd Set)	Five Mazurkas No. 1. B flat major (5) No. 2. A minor (6) No. 3. F minor (7) No. 4. A flat major (8) No. 5. C major (9)		Dec. 1832	Mr. Johns
8	Trio for Piano, Violin, and Violoncello. G minor	1828	1833	M. le Prince Antoine Radziwill
9	Three Nocturnes No. 1. B flat minor (Larghetto) No. 2. E flat major (Andante) No. 3. B major (Allegretto)	Probably 1832	Jan. 1833	Mme. Camille Pleyel

Scan 11



WORKS WITHOUT OPUS NUMBERS PUBLISHED AFTER THE COMPOSER'S DEATH

Name of Piece and Key	Composed	Published
Variations. E major. On a German Air	1824 ?	1851
Mazurka. G major	1825	
Mazurka. B flat major	1825	
Mazurka. D major	1829-1830	
Mazurka. D major. A remodelling of the preceding Mazurka	1832	
Mazurka. C major	1833	
Mazurka. A minor. Dédiée à son ami Emile Gaillard		
Valse. E minor		1868
Polonaise. G sharp minor. Dédiée à Mme. Dupont	1822	1864
Polonaise. G flat major. Of doubtful authenticity		1872
Polonaise. B flat minor. Adieu! an Wilhelm Kolberg	1826	
Valse. E major	1829	

WORKS WITHOUT OPUS NUMBERS PUBLISHED DURING THE COMPOSER'S LIFE-TIME

Name of Piece and Key	Composed	Published
Grand Duo Concertant. E major. For piano and violoncello, on themes from "Robert le Diable."	1829, with A. Franchomme	1833
Trois nouvelles Etudes. F minor, A flat major, D flat major		1840
Variations VI. E major (Largo). From the "Hexameron"		1841
Mazurka. A minor ("Notre temps")		1842

Scan 12

and

$$\Delta p = \rho_v g h,$$

being the difference between the vapor pressures at the inner and outer surfaces.

Hence

$$h = \frac{\Delta p}{\rho_v g} = \frac{2T}{R(\rho_w - \rho_v)g},$$

and

$$\Delta p = \frac{2T\rho_v}{R(\rho_w - \rho_v)}.$$

At ordinary temperatures and for droplets whose radii are 10^{-4} cm. (a possible size) the temperature depression, or error, amounts roughly to 0.02° C. According to the equation, the error obviously might have any value, though actually it seems always to be small; that is, this, too, like many other physical equations, has its limitations.

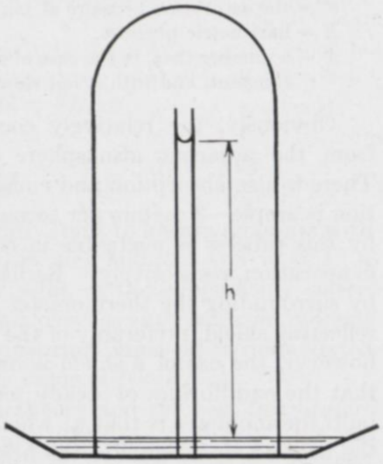


Fig. 6.—Relation of curvature of surface to saturation vapor pressure.

In taking humidity measurements the observer must be careful that his presence does not affect the amount of moisture in the air under examination—he must stand to the lee of his apparatus.

Although the dew-point apparatus and other absolute hygrometers are extremely simple in theory, they generally are too complicated in structure and too difficult to manipulate to be suitable for routine observations.[1] On the other hand, the psychrometer, presently to be explained, which depends on the maximum cooling of water by evaporation when amply ventilated, is less obvious in theory,[2] but both simple in construction and easy to use.

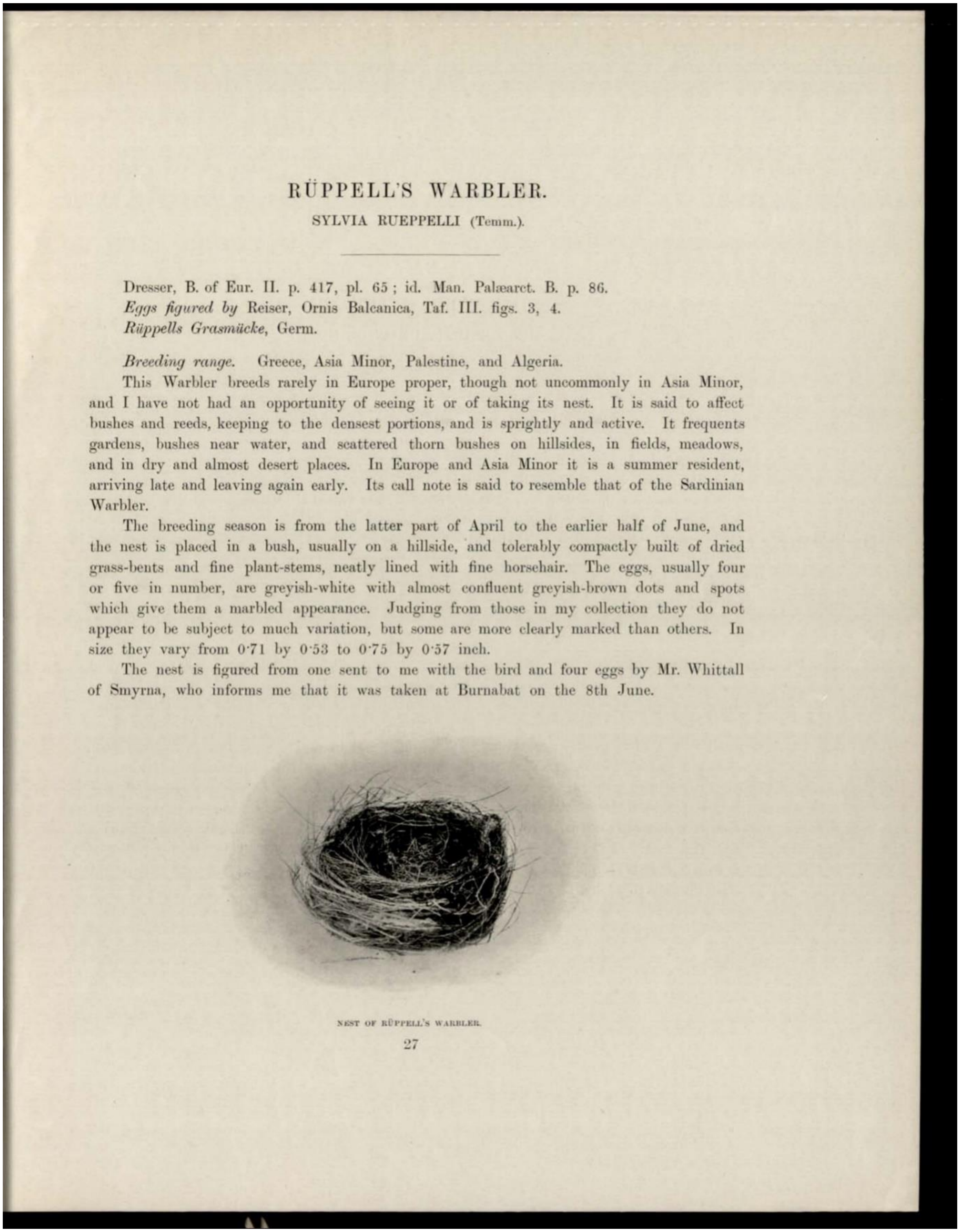
A convenient form of the psychrometric equation is:

$$e = e' - AB(t - t'),$$

[1] For a general discussion of hygrometry see *Phys. Soc. Lon.*, **34,** February, 1922; Glazebrook, "Dictionary of Applied Physics," Vol. 3; and Bongards, Hermann, "Feuchtigkeits Messung," 1926. See also Whipple, F. J. W., *Proc. Phys. Soc.*, **45**; Pt. 2, 307, 1933, and Brooks, D. B., and Allen, H. H., *J. Wash. Acad. Sci.*, **28**; 121, 1933.

[2] Ivory, *Phil. Mag.*, **60**; 81, 1822; August, *Ann. Phys.*, **5**; 69, 1825; Apjohn, *Trans. Roy. Irish Acad.*, 1834; Regnault, *C. R.*, **20**; 1127–1220, 1845; **35**; 930, 1852; Maxwell, Encyc. Brit., 9th Ed., "Diffusion," 1878; Stefan, *Zeit. Oest. Gesell. für Meteorologie*, **16**; 177, 1881; Ferrel, Annual Report, Chief Signal Officer, Appendix 71, "Hygrometry," 1885; Carrier, *Trans. Am. Soc. Mech. Eng.* **33**; 1005, 1912; Grossmann. *Ann. Hydr.* **44**; 577, 1916.

Scan 13



RÜPPELL'S WARBLER.

SYLVIA RUEPPELLI (Temm.).

Dresser, B. of Eur. II. p. 417, pl. 65 ; id. Man. Palæarct. B. p. 86.
Eggs figured by Reiser, Ornis Balcanica, Taf. III. figs. 3, 4.
Rüppells Grasmücke, Germ.

Breeding range. Greece, Asia Minor, Palestine, and Algeria.

This Warbler breeds rarely in Europe proper, though not uncommonly in Asia Minor, and I have not had an opportunity of seeing it or of taking its nest. It is said to affect bushes and reeds, keeping to the densest portions, and is sprightly and active. It frequents gardens, bushes near water, and scattered thorn bushes on hillsides, in fields, meadows, and in dry and almost desert places. In Europe and Asia Minor it is a summer resident, arriving late and leaving again early. Its call note is said to resemble that of the Sardinian Warbler.

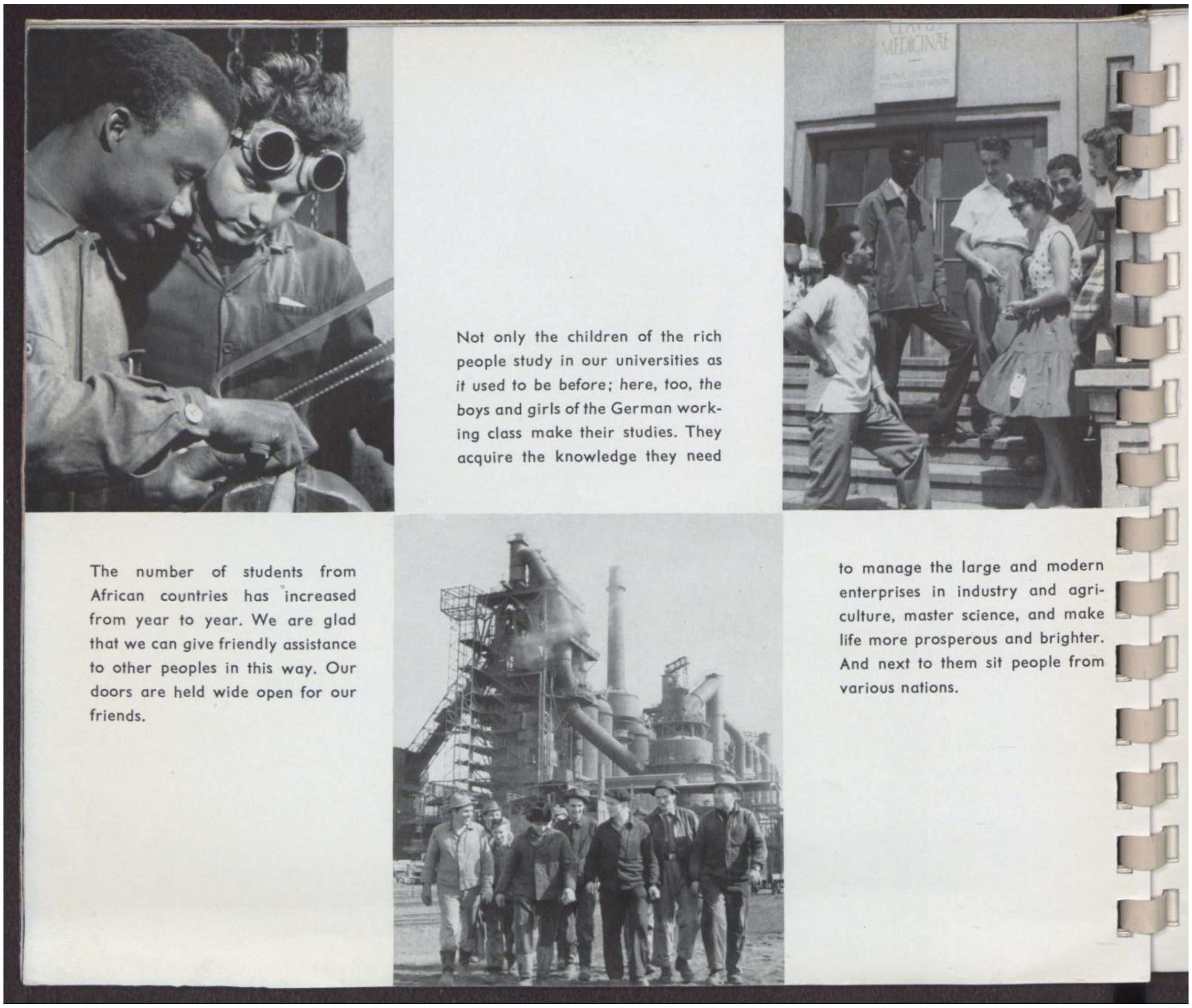
The breeding season is from the latter part of April to the earlier half of June, and the nest is placed in a bush, usually on a hillside, and tolerably compactly built of dried grass-bents and fine plant-stems, neatly lined with fine horsehair. The eggs, usually four or five in number, are greyish-white with almost confluent greyish-brown dots and spots which give them a marbled appearance. Judging from those in my collection they do not appear to be subject to much variation, but some are more clearly marked than others. In size they vary from 0·71 by 0·53 to 0·75 by 0·57 inch.

The nest is figured from one sent to me with the bird and four eggs by Mr. Whittall of Smyrna, who informs me that it was taken at Burnabat on the 8th June.

NEST OF RÜPPELL'S WARBLER.

27

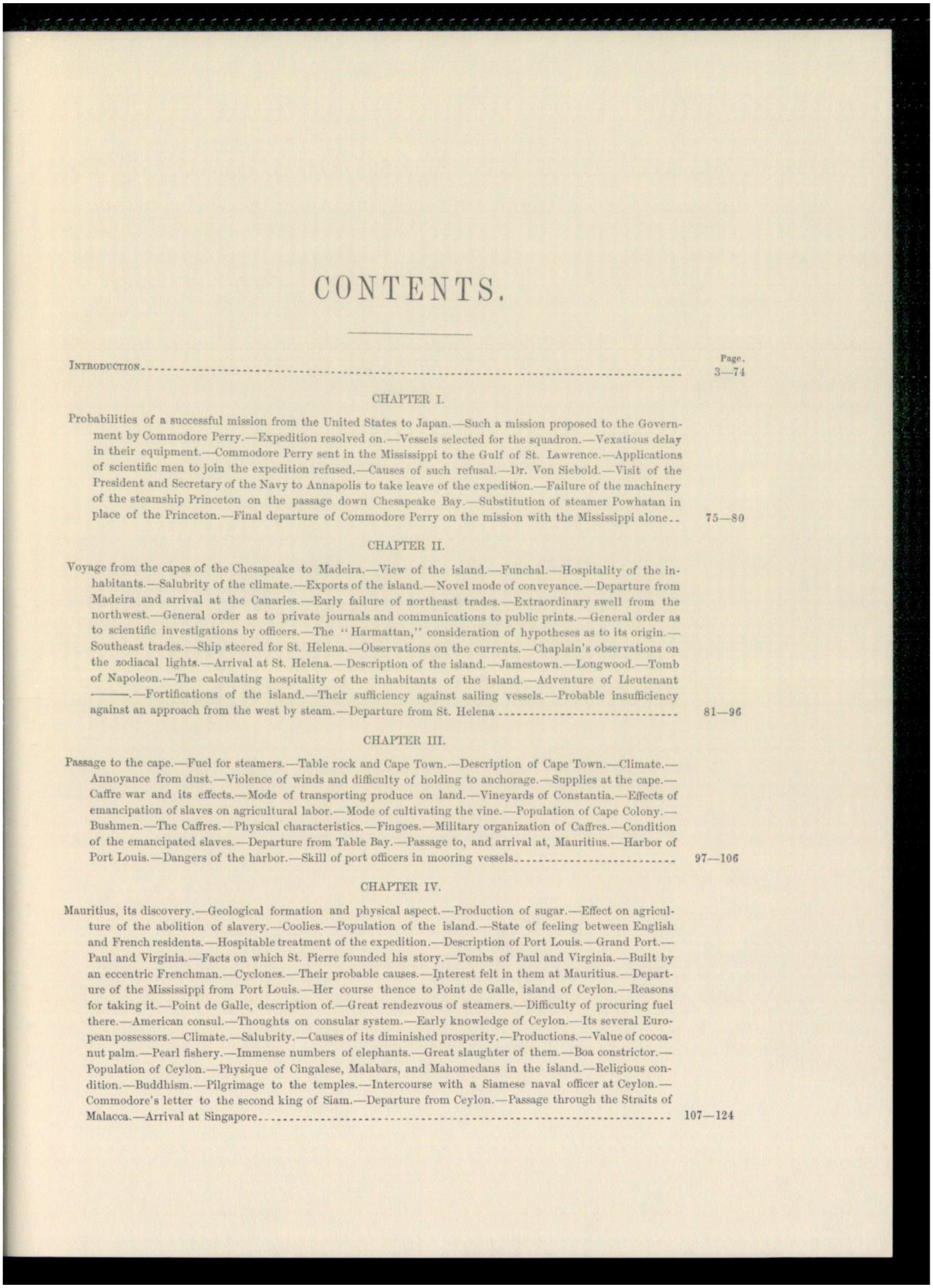
Scan 14



Not only the children of the rich people study in our universities as it used to be before; here, too, the boys and girls of the German working class make their studies. They acquire the knowledge they need to manage the large and modern enterprises in industry and agriculture, master science, and make life more prosperous and brighter. And next to them sit people from various nations.

The number of students from African countries has increased from year to year. We are glad that we can give friendly assistance to other peoples in this way. Our doors are held wide open for our friends.

Scan 15



CONTENTS.

	Page.
Introduction	3—74

CHAPTER I.

Probabilities of a successful mission from the United States to Japan.—Such a mission proposed to the Government by Commodore Perry.—Expedition resolved on.—Vessels selected for the squadron.—Vexatious delay in their equipment.—Commodore Perry sent in the Mississippi to the Gulf of St. Lawrence.—Applications of scientific men to join the expedition refused.—Causes of such refusal.—Dr. Von Siebold.—Visit of the President and Secretary of the Navy to Annapolis to take leave of the expedition.—Failure of the machinery of the steamship Princeton on the passage down Chesapeake Bay.—Substitution of steamer Powhatan in place of the Princeton.—Final departure of Commodore Perry on the mission with the Mississippi alone 75—80

CHAPTER II.

Voyage from the capes of the Chesapeake to Madeira.—View of the island.—Funchal.—Hospitality of the inhabitants.—Salubrity of the climate.—Exports of the island.—Novel mode of conveyance.—Departure from Madeira and arrival at the Canaries.—Early failure of northeast trades.—Extraordinary swell from the northwest.—General order as to private journals and communications to public prints.—General order as to scientific investigations by officers.—The "Harmattan," consideration of hypotheses as to its origin.—Southeast trades.—Ship steered for St. Helena.—Observations on the currents.—Chaplain's observations on the zodiacal lights.—Arrival at St. Helena.—Description of the island.—Jamestown.—Longwood.—Tomb of Napoleon.—The calculating hospitality of the inhabitants of the island.—Adventure of Lieutenant ———.—Fortifications of the island.—Their sufficiency against sailing vessels.—Probable insufficiency against an approach from the west by steam.—Departure from St. Helena 81—96

CHAPTER III.

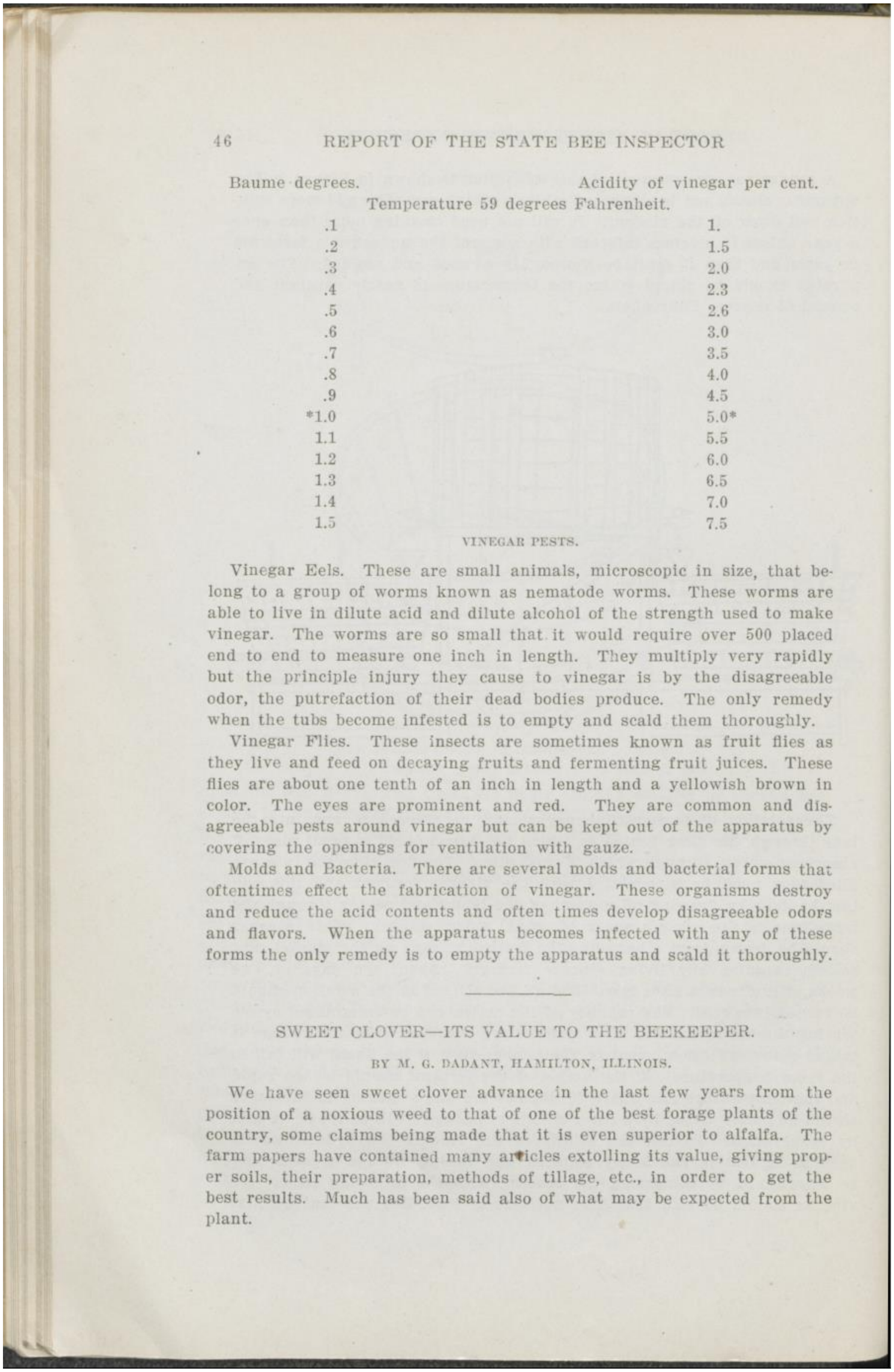
Passage to the cape.—Fuel for steamers.—Table rock and Cape Town.—Description of Cape Town.—Climate.—Annoyance from dust.—Violence of winds and difficulty of holding to anchorage.—Supplies at the cape.—Caffre war and its effects.—Mode of transporting produce on land.—Vineyards of Constantia.—Effects of emancipation of slaves on agricultural labor.—Mode of cultivating the vine.—Population of Cape Colony.—Bushmen.—The Caffres.—Physical characteristics.—Fingoes.—Military organization of Caffres.—Condition of the emancipated slaves.—Departure from Table Bay.—Passage to, and arrival at, Mauritius.—Harbor of Port Louis.—Dangers of the harbor.—Skill of port officers in mooring vessels 97—106

CHAPTER IV.

Mauritius, its discovery.—Geological formation and physical aspect.—Production of sugar.—Effect on agriculture of the abolition of slavery.—Coolies.—Population of the island.—State of feeling between English and French residents.—Hospitable treatment of the expedition.—Description of Port Louis.—Grand Port.—Paul and Virginia.—Facts on which St. Pierre founded his story.—Tombs of Paul and Virginia.—Built by an eccentric Frenchman.—Cyclones.—Their probable causes.—Interest felt in them at Mauritius.—Departure of the Mississippi from Port Louis.—Her course thence to Point de Galle, island of Ceylon.—Reasons for taking it.—Point de Galle, description of.—Great rendezvous of steamers.—Difficulty of procuring fuel there.—American consul.—Thoughts on consular system.—Early knowledge of Ceylon.—Its several European possessors.—Climate.—Salubrity.—Causes of its diminished prosperity.—Productions.—Value of cocoa-nut palm.—Pearl fishery.—Immense numbers of elephants.—Great slaughter of them.—Boa constrictor.—Population of Ceylon.—Physique of Cingalese, Malabars, and Mahomedans in the island.—Religious condition.—Buddhism.—Pilgrimage to the temples.—Intercourse with a Siamese naval officer at Ceylon.—Commodore's letter to the second king of Siam.—Departure from Ceylon.—Passage through the Straits of Malacca.—Arrival at Singapore 107—124

Co-funded by the Creative Europe Programme of the European Union

Scan 16



Temperature 59 degrees Fahrenheit.

Baume degrees.	Acidity of vinegar per cent.
.1	1.
.2	1.5
.3	2.0
.4	2.3
.5	2.6
.6	3.0
.7	3.5
.8	4.0
.9	4.5
*1.0	5.0*
1.1	5.5
1.2	6.0
1.3	6.5
1.4	7.0
1.5	7.5

VINEGAR PESTS.

Vinegar Eels. These are small animals, microscopic in size, that belong to a group of worms known as nematode worms. These worms are able to live in dilute acid and dilute alcohol of the strength used to make vinegar. The worms are so small that it would require over 500 placed end to end to measure one inch in length. They multiply very rapidly but the principle injury they cause to vinegar is by the disagreeable odor, the putrefaction of their dead bodies produce. The only remedy when the tubs become infested is to empty and scald them thoroughly.

Vinegar Flies. These insects are sometimes known as fruit flies as they live and feed on decaying fruits and fermenting fruit juices. These flies are about one tenth of an inch in length and a yellowish brown in color. The eyes are prominent and red. They are common and disagreeable pests around vinegar but can be kept out of the apparatus by covering the openings for ventilation with gauze.

Molds and Bacteria. There are several molds and bacterial forms that oftentimes effect the fabrication of vinegar. These organisms destroy and reduce the acid contents and often times develop disagreeable odors and flavors. When the apparatus becomes infected with any of these forms the only remedy is to empty the apparatus and scald it thoroughly.

SWEET CLOVER—ITS VALUE TO THE BEEKEEPER.

BY M. G. DADANT, HAMILTON, ILLINOIS.

We have seen sweet clover advance in the last few years from the position of a noxious weed to that of one of the best forage plants of the country, some claims being made that it is even superior to alfalfa. The farm papers have contained many articles extolling its value, giving proper soils, their preparation, methods of tillage, etc., in order to get the best results. Much has been said also of what may be expected from the plant.

Annex 2: 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: _______

OCR: editing reading order of recognized layout elements

- yes
- no

Other notes: _______

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? _______

Annex 3. Testing report questionnaires by partners

P1 - UIBK, University of Innsbruck – RTF format

A12 TEST REPORT

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

Partner organisation: University of Innsbruck
Which software for image processing and OCR did you use for this sample?
Abbyy FineReader 14, Adobe Indesign

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?
Convert from tif to jpeg with irfan view because of error messages in abbyy fine reader 14 for 3 files ("Möglicherweise ist die Datei beschädigt")

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: _______

OCR: editing reading order of recognized layout elements

- [✓ yes]
- no

Other notes: _______

OCR: additional work on recognised text

- [✓ fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)]
- no

Other notes: additional work: add origpage, caption, alt-text

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.

We use Abbyy Finereader only for layout analysis, text recognition and correction. All other processes (markup of headers, adding elements such as origpage, caption, footnotes, etc.) are then carried out in Adobe Indesign. Finally, the table of contents is created in Microsoft Word.

5. EXPORT

Any additional comments?

We export and deliver the file as RTF.

P1 - UIBK, University of Innsbruck – ODM workflow – PDF and RTF format

A12 TEST REPORT

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

Partner organisation: University of Innsbruck
Which software for image processing and OCR did you use for this sample?
Abbyy FineReader recognition server 4 - testing the ODM workflow

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?
Convert 3 files because there were some problems with uploading

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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? We exported the files in PDF, PDF/A, alto, RTF, xml

P2 - UT, University of Tartu

A12 TEST REPORT

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

Partner organisation: University of Tartu Library
Which software for image processing and OCR did you use for this sample?
ABBYY FineReader PDF 15 Standard; ABBYY FineReader Server 14

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: rotated one image

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

OCR: additional work on recognised text

- [✓ fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)]
- [✓ no]

Other notes: Only done for special cases, projects – rarely.

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. no

5. EXPORT

Any additional comments? _______

P3 - NUK, National and University Library – usual workflow

A12 TEST REPORT

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

Partner organisation: National and University Library (Slovenia)
Which software for image processing and OCR did you use for this sample?
We use internally developed workflow software which uses Abbyy FineReader engine for image processing and OCR.

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: We usually use equalizing the dimensions of the scans but in this example, we did not use it because scan sizes were too diverse. Because of it, OCR didn't function correctly.

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English
- other (add): latin

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: Automatic recognition recognizes just the columns (ex. newspapers).

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments? _______

P3 - NUK, National and University Library – PDF edited with Adobe Acrobat Pro

A12 TEST REPORT

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

Partner organisation: National and University Library (Slovenia)
Which software for image processing and OCR did you use for this sample?
We use internally developed workflow software which uses Abbyy FineReader engine for image processing and OCR.

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: We usually use equalizing the dimensions of the scans but in this example, we did not use it because scan sizes were too diverse. Because of it, OCR didn’t function correctly.

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English
- other (add): latin

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: Automatic recognition recognizes just the columns (ex. newspapers).

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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?

For testing purposes the final PDF was edited using Adobe Acrobat pro for manually adding the images, autotagging, editing tags, manually fixed reading order, language segments were added, TOC on scan 15 was nested and footnotes on scan 12 were nested. Tables were turned into images although we know it is not the correct way. We used adobe's accessibility check to fix any other problem (name and language of the document for example).

P3 - NUK, National and University Library – PDF made with Abbyy FineReader 15

A12 TEST REPORT

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

Partner organisation: National and University Library (Slovenia)
Which software for image processing and OCR did you use for this sample?
For testing purposes was used Abbyy FineReader 15 desktop version.

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: _______

OCR: editing reading order of recognized layout elements

- [✓ yes]
- no

Other notes: _______

OCR: additional work on recognised text

- fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)
- [✓ no]

Other notes: _______

3. 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. no

4. EXPORT

Any additional comments?

At export from Abbyy we chose that it is compliant with PDF/A and PDF/UA standard.
Additionally, we used Adobe Acrobat Pro for autotagging but tags were not checked.

P3 - NUK, National and University Library – PDF and ePUB made from Word

A12 TEST REPORT

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

Partner organisation: National and University Library (Slovenia)
Which software for image processing and OCR did you use for this sample?
We use internally developed workflow software which uses Abbyy FineReader engine for image processing and OCR.

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: We usually use equalizing the dimensions of the scans but in this example, we did not use it because scan sizes were too diverse. Because of it, OCR didn't function correctly.

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English
- other (add): latin

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments?

For testing purposes the final TXT file was taken to Microsoft Office Word for further work. We added images, did full OCR clean-up and fixed reading order. We added structure, page numbers, footnotes, hyperlinks, captions to images and tables. We exported as PDF/A-3A. Using Adobe Acrobat pro we did accessibility check, checked reading order and fixed the title of the document.

For ePUB we used the workflow we do on EODOPEN for ePUB production. The above clean Microsoft Word file was converted with the tool WordToEpub and then we used Sigil for fixing mistakes and did accessibility check with EpubCheck and Ace by Daisy.

P4 - MZK, Moravian Library – small edited

A12 TEST REPORT

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

Partner organisation: Moravian Library
Which software for image processing and OCR did you use for this sample?
PROJECT PERO OCR
https://pero.fit.vutbr.cz/
https://pero-ocr.fit.vutbr.cz/
https://github.com/DCGM/pero-ocr
This report is for documents in folder 01_MZK_Small edited (SE)

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?
Crop images (only selected)
Rotate the images (only selected)
Resize the images (we need in this tool max. 8 Mb).

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: no further adjustments

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English
- [✓ other (add):]

Czech Printed Model +
Language Model - English Wikipedia

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments?

Export is in Page and Alto format (+txt with plain text).

P4 - MZK, Moravian Library – edited

A12 TEST REPORT

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

Partner organisation: Moravian Library
Which software for image processing and OCR did you use for this sample?
PROJECT PERO OCR
https://pero.fit.vutbr.cz/
https://pero-ocr.fit.vutbr.cz/
https://github.com/DCGM/pero-ocr
This report is for documents in folder 01_MZK_Edited (E)

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?
Rotate the images (only selected)
Resize the images (we need in this tool max. 8 Mb).

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: no further adjustments

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English
- [✓ other (add):]

Czech Printed Model +
Language Model - English Wikipedia

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: _______

OCR: editing reading order of recognized layout elements

- [✓ yes]
- no

Other notes: _______

OCR: additional work on recognised text

- fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)
- [✓ no]

Other notes: _______

3. 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. no

4. EXPORT

Any additional comments?

Export is in Page and Alto format (+txt with plain text).

P5 - UG, University of Greifswald

A12 TEST REPORT

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

Partner organisation: University of Greifswald
Which software for image processing and OCR did you use for this sample? Abby

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? Imageframes and JPEG-Compression (Usually done automatically by our workflowsystem)

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: no

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: no

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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.

Yes in regular workflows we use Intranda Layout wizard: automatic deskewing, page cropping, semiautomatic separation of pages, hard quality management no bright pictures, good contrast, but no postprocessing after OCR.

5. EXPORT

Any additional comments?

- good scans are half the battle
- Preliminary work would have to be done by marking the image areas

- Reworking the Images and wrongly transscriped Images areas for Accessibility
- A workflow without preliminary work or reworking

UG Notes about A12

ArchEmig	Lines and captions are not identified Stains on the paper (newspaper) are identified as punctuation marks (Epub) Columns are not recognized throughout. The continuous text jumps (txt)
Chemestry	Formulars were not detected Variables (Greek alphabet) are not identified Lines and captions are not identified
EOD-Open	Headers are missing (Epub) Headers and pagination not in the right positions (txt)
Gromdzenie	Table structure are not identified (ePUB and txt) Different fonts are identified (Epub) different grades of transcription Font of recitation could not get transcripted (txt) Problems with fracture and antiqua in cusiv
Magazyn	Variables (Greek alphabet) identified as special characters Formulars were not detected (in every format)
Eegs	Problems with fracture and antiqua in cusiv Some normal characters were not detected and transscripted
Internat-ariculture	Layout and framing is not compatible Transcription is in ePUB right Txt and PDF alright
Narrative	Upper and lower case wrong (PDF und Epub) Lines and captions are not identified Probably too pale scan. Text not always translated correctly (txt und Epub)
Report	Translation of the tabular display distorted Probably too pale scan. Text not always transscriped correctly (txt und Epub) Upper and lower case wrong (PDF und Epub)
UG	To be fair, it is an addendum and it is a relatively recent publication. The page is clean and the typesetting regular. The cusive font has enough spacing between letters.

- Problems with the txt and PDF are pretty much the same
- good scans are half the battle
- Preliminary work would have to be done by marking the image areas
- Reworking the Images and wrongly transscribed Images areas for Accessibility

P6 - NLS, National Library of Sweden

A12 TEST REPORT

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

Partner organisation: National Library of Sweden
Which software for image processing and OCR did you use for this sample?
ABBYY Finereader 11

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: None of the above

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: no

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: no

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments?

Up until now we have used ODM when digitising orders for the EODOPEN project. We are just about to start using a new system (Limb Processing) and the file we uploaded are made via this system.

P7 - NCU, Nicolaus Copernicus University in Torun

A12 TEST REPORT

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

Partner organisation: Nicolaus Copernicus University in Toruń
Which software for image processing and OCR did you use for this sample?
ABBYY FineReader Server 14.0

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? Resolution to 300 dpi

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: Resolution of all scans is set to 300 dpi

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments?

We have some problems with software after the last update.

PDFs are linearized before publication.

P9 - VKOL, Research Library Olomouc

A12 TEST REPORT

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

Partner organisation: Research Library Olomouc
Which software for image processing and OCR did you use for this sample?
ScanTailor Advanced v1.01.16
Tesseract 5.0.0-beta-20210815-22-g386dd

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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? _______

P10 - BNP, National Library of Portugal – PDF

A12 TEST REPORT

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

Partner organisation: National Library of Portugal
Which software for image processing and OCR did you use for this sample?
LIMB Processing and IRIS OCR

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments? _______

P10 - BNP, National Library of Portugal – docx

A12 TEST REPORT

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

Partner organisation: National Library of Portugal
Which software for image processing and OCR did you use for this sample?
LIMB Processing and IRIS OCR

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: _______

OCR: editing reading order of recognized layout elements

- [✓ yes]
- no

Other notes: _______

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? _______

P11 - NLE, National Library of Estonia

A12 TEST REPORT

Please, add detailed information! You can also add screenshots or record the testing process.
Partner organisation: National Library of Estonia
Which software for image processing and OCR did you use for this sample?
For books files: ABBYY FineReader 11 for image processing and ABBYY Recognition Server 4 for OCR.
For newspaper/periodicals: ABBYY FineReader 11 and CCS docWorks 7.1.0.90 for image processing and ABBYY 12 OCR-engine

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: At the moment we have 2 different workflows for 1) books and 2)newspapers/periodicals. We do segmentation (with software CCS Docworks) only on periodicals at the moment (book files as the most sample files were are just deskewed, cropped and OCR-d), but there is plan in the near 1-2 years to switch the books also to the segmentation workflow.

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: _______

OCR: editing reading order of recognized layout elements

- [✓ yes]
- no

Other notes: _______

OCR: additional work on recognised text

- [✓ fixing OCR mistakes (wrongly recognized characters, words, decorative initial or any other mistakes)]
- no

Other notes: We are fixing only OCR mistakes of periodical's Headlines, Captions and Authors, seldom in Textblocks

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. no

5. EXPORT

Any additional comments?

Export files are also different at the moment depending on the type of the item – book files (as most of the testing samples here were) are OCRed PDFs for the user but segmented newspapers/periodicals are jpeg2000 and PDF (1 sample file).

As we have 2 different workflows for the books and periodicals, we also have 2 different portals for them as well. But this is going to change in the near future as we are starting to implement a new archival system soon and all the materials will go under segmentation and hopefully in the same portal as well. There are a lot of changes ahead of us in this field :)

P12 - OSZK, National Széchényi Library

A12 TEST REPORT

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

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

1-6, 8-11, 14-15
ScanTailor Advanced (1.0.16)
Photoshop (v 23.2.2)
ABBYY Recognition Server 4.0

7, 13, 16
Photoshop (v 23.2.2)
ABBYY Recognition Server 4.0

12
ScanTailor Advanced (1.0.16)
ABBYY Recognition Server 4.0

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 (15)]
- [✓ no (others)]

If yes, what did you change? _______

2. IMAGE PROCESSING

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

Deskewing:

- [✓ automatic (1-5, 8-12, 14-15)]
- [✓ manual (6-7, 13, 16)]
- automatic and manual

Cropping:

- automatic
- [✓ manual]
- automatic and manual

Additional steps:

- [✓ lines straightening (dewarping) (12)]

- noise removal (denoising)
- contrast enhancement
- correction of geometric distortion
- binarization (1-6, 8, 12, 15)
- removal of stamps, written notes (1-5, 8)
- equalising the dimensions of the scans (all same size after cropping)

Other notes:
1-5: “Mixed” output was chosen while processing the samples using Scan Tailor: the textual content was selected and binarised while the illustrations were remained in color mode. During the binarisation we’ve thickened the letters to make it easier for the OCR-algorithm to recognise the characters.
6: Converting to grayscale, increasing contrast and adjusting levels using Photoshop.
7: “Smart” sharpening, Adjusting levels.
9: “Smart” sharpening, Neural Filters, Removal of negative visual effects caused by JPEG-compression (middle), Adjusting levels.
10-11: “Smart” sharpening, Neural Filters, Removal of negative visual effects caused by JPEG-compression (middle), Adjusting levels.
12: Automatic dewarping.
13: “Smart” sharpening (with noise removal), Neural Filters, Removal of negative visual effects caused by JPEG-compression (middle), Adjusting levels
14: “Mixed” output was chosen while processing the samples using Scan Tailor: the textual content was selected and binarised while the illustrations were remained in color mode. After Scan Tailor processing we adjusted the levels using Photoshop.
16: “Smart” sharpening, Neural filters, Removal of negative visual effects caused by JPEG-compression (middle), Converting to grayscale, Adjusting levels

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- English (1-5, 7-16)
- ~~other (add):~~ Polish (6)

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 (1-6, 8, 12-16)
- additional manual corrections (7)
- we do not use it (9-11)

Other notes:
7: The automatic recognition had skipped the text on the 5. page, therefore we selected it manually after the automatic processing.

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments? _______

P13 - CVTI SR, Slovak Centre of Scientific and Technical Information

A12 TEST REPORT

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

Partner organisation: Slovak Centre of Scientific and Technical Information
Which software for image processing and OCR did you use for this sample?
ScanGate by Treventus Mechatronics for image post processing
ABBYY Recognition Server 4.0 for OCR

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: some pictures - background normalisation, unsharp masking
We are using equalising the dimensions of the scans but in this example, we did not use it because images sizes were too difference.

3. MULTILEVEL DOCUMENT ANALYSIS AND RECOGNITION OF ELEMENTS

OCR: character recognition:

- [✓ English]
- [✓ other (add): German, German (new spelling), Slovak, Czech]

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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. no

5. EXPORT

Any additional comments?

We are doing two outputs pdf files. One output is only for long time archive. The second pdf file output is for using to digital library (mostly with smaller file size with some conversion).

P14 - UREG, University of Regensburg

A12 TEST REPORT

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

Partner organisation: University Library of Regensburg
Which software for image processing and OCR did you use for this sample?
ABBYY Recognition Server 4.0

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: _______

OCR: editing reading order of recognized layout elements

- yes
- [✓ no]

Other notes: _______

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?

The exported formats are XML, Text and PDF containing the recognized text. The last is served to the users.