

From paper to electronic books

MATIJA KATALENIĆ, MIHAEL LADIĆ, PETAR MILJKOVIĆ

UNIVERSITY NORTH, VARAŽDIN, CROATIA

ABSTRACT

An e-book is an electronic book which we can read on devices such as computers, tablets, mobile phones or special devices intended for such reading. The Oxford Dictionary of English defines an e-book as an electronic version of a printed book, but this cannot be considered as a correct definition since many books exist only in digital form without their printed counterpart. E-book production processes in both the software and graphic parts are really unlimited and this paper presents only one of them in brief steps. The intention is to solely familiarise the user with all that is necessary for production and does not serve for book piracy. If you like a book, we encourage you to buy it and support the author.

Keywords: e-book, scan, EPUB, Sigil, Calibre, Word, OCR, HTML, CSS

Introduction

The very idea of an electronic book reader is emphasised in 1949 by a Spanish teacher Angela Ruiz Robles who had a prototype of an e-reader that could contain more text in one place. However, this did not take root, so the real beginning of book digitalisation is considered to be 1971 when Michael Hart received a copy of the United States Declaration of Independence, typed and placed it on the ARPANET (the precursor of today's Internet) for download. Six individuals downloaded the text and the first "e-book" was born. During the same year, Michael Hart, a student at the University of Illinois, started working on the Project Gutenberg whose goal was to make written works falling into the category of public goods, i.e. works that were not copyrighted available to everyone in an electronic form. In this way, Gutenberg became the first information provider on the Internet and the first digital library in the world, with Michael Hart digitalising the first 100 books. Today, it has over 57.000 books. In 1981, the first digital book (a dictionary) was placed on the market. The real success of electronic books took place in 1998 when the first Electronic-book Fair took place in Gaithersburg, Maryland, in the USA. At this event, a series of norms and regulations contributing to the standardisation of electronic publishing were presented. A year later, HTML and XML formats were

set as the standard file types for electronic books. A great contribution to the development of such books happened in 2001 when Stephen King's famous work *Riding the Bullet* was launched exclusively on the Internet in the form of an e-book. From this moment on, many electronic publishers and online libraries have appeared. [06] [10] The example does neither show all the necessary operations in Word, nor all the necessary HTML and CSS since the same can be done in several different ways.

What are e-books and how can they be read?

An e-book is an electronic book which can be read on different devices. In this paper, we will mention several formats, as well as one of the most famous e-readers: Kindle. Through the advancement of technology and the development of different solutions for reading books, inserting comments, adding external links by the readers themselves and communication between the users (readers) (e.g. Goodreads) is already enabled. Some of the remaining numerous advantages of producing and using e-books are:

- the availability of several books on one device - easier portability,
- enabled access to additional information (links, multimedia, etc.)
- ease of update and the availability of updated book versions,
- better accessibility than printed books to a large number of people with disabilities,
- cheaper production since there is no need for printing, physical distribution, etc
- lower purchase price than printed books, they are often free,
- ecological advantage since they contribute to environmental protection due to non-consumption of paper,
- text can be searched,
- they require little space (approximately 1000 average e-books can be stored on one CD),
- they can be read in dim light, and even in total darkness if the e-reader has a built-in light,
- they do not wear out over time [01]

In our opinion, they almost have no shortcomings. Namely, the battery lasts for weeks which primarily depends on using the light intensity and the Internet, just as with other devices.

EPUB (Electronic Publication) is a free and open format which has been accepted in 2007 and further developed by IDPF (International Digital Publishing Forum), an international organisation gathering different participants and developing standards in the area of digital publishing. EPUB represents a standard with regard to producing e-books, and it is based on different technologies and standards such as XML, CSS, XHTML and Open eBook. The file extension is .epub which can be easily opened and the content seen by only changing the extension into .zip. [01]

Once properly compiled, an e-book will always keep its basic features (paragraphs, chapters, indents) despite our changes to the font (type, size, spacing, margins). A good .epub is easily converted into any other format, for example .azw3 intended for Amazon e-readers.

PDF (Portable Document Format) as a format is good for manuals and magazines which contain a lot of images and graphics, but it is not so desirable in an e-book. It can be read and adapted nicely on a computer. Long-term reading on such screens can lead to eye strain and therefore a headache. With e-readers, there is no such problem. They can also read .pdf, but the font size, spacing and everything

else that would make reading easier cannot be adjusted which is why .pdf for e-readers is not a good choice. Some of the more famous e-readers are certainly Kobo, Nook and Kindle. Most e-readers use e-ink technology. They are special in the fact that they imitate classic paper and ink, showing only white and black colour, as well as grayscale. That is why they are pleasant for reading and do not cause eye strain like LCD screens, with the battery lasting longer on an e-reader because of them. Amazon's 6-inch e-reader of the latest generation is one of the most lightweight e-readers in the world, and with its weight being only 182 grams, it is lighter than many paperbacks. The internal memory of 8 GB enables you to carry thousands of books in your pocket, which you can download directly through Wi-Fi, without connecting your Kindle to a computer. There is also a model with a connection through a 4G signal. The new model comes with an improved screen which will make your reading experience even more pleasant and natural. The built-in backlight allows easy reading even in bright sunlight which is not the case with other devices and tablets. Reading is very pleasant even in complete darkness. The battery does not last for hours, but weeks.

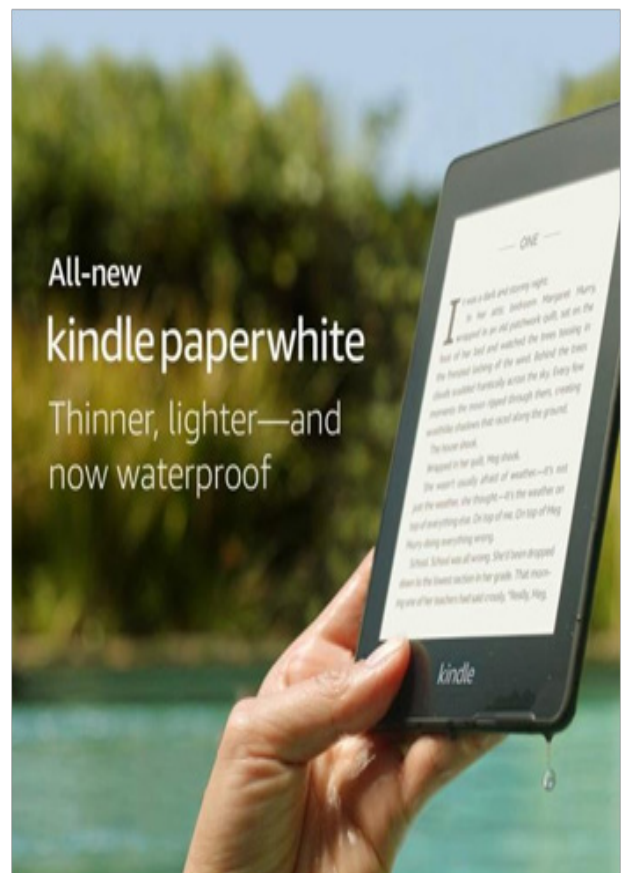


Image 1: The latest Kindle model

E-book production

E-book production is a process that can be conducted in different ways and there is no rule to abide by. From choosing different image scanners, or ordinary photographing of pages with a mobile phone up to professional devices for this purpose. It is desirable to be familiar with the basics of HTML and CSS.

There are many applications that make the job easier such as LibreOffice Writer, Atlantis Word Processor, Foxit, Adobe Digital Editions, Adobe InDesign and many others. In this example, we will use Abbyy, Microsoft Word, Sigil, AlReader2 and Calibre, as well as some ancillary applications like Brackets, WinRAR and Photoshop, and accordingly, we can divide the production process into five basic steps:

1. Book scanning
2. OCR
3. Text correction in Word
4. Producing and processing EPUB
5. Conversion

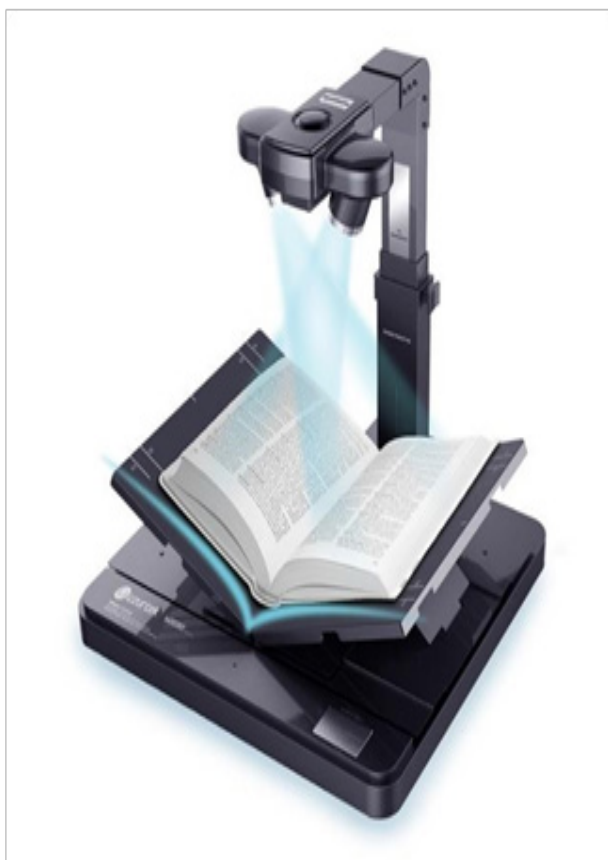


Image 2: Professional book scanner

Book scanning

It is best to scan at a 300 dpi grayscale which will provide the best results for further processing, apart from the cover page which we produce in colour. It is also important that the work is done precisely according to the pages from the book. There are professional devices intended for book scanning, but the cheaper ones we have at home can also be used. Likewise, it is possible to take pictures of the pages with a mobile phone. The duration of the process, as well as OCR quality depend upon the quality of the obtained image or PDF. If the book is older, it is wise to wipe the pages with a dry cloth so that dust or some other particles would not additionally insert some symbols. For our example, we will use a scan of the book Jasmine and Olive written by the Hungarian writer and philosopher Béla Hamvas. The procedure is the same regardless of whether we use PDF, JPEG, or any other format.



Image 3: Scanned book covers



Image 4: Scanned book appearance

OCR - Optical Character Recognition

During the second step, we perform OCR (Optical Character Recognition). For this, we will use the following application: Abbyy. When we open it, we load our obtained scan. We choose black and white so that the loading would be faster and more precise. Coloured text is an exception that can later be corrected if necessary.

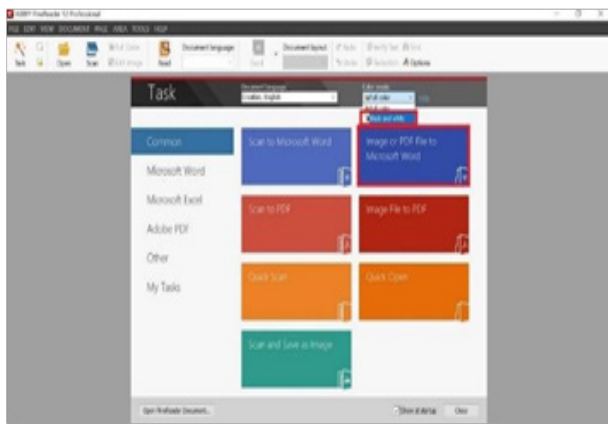


Image 5: Abbyy Main Window

Abbyy will automatically convert everything into editable text which we can immediately modify, but it will be much easier to do in Word. Even though the scan is made in such a way that there are two pages on one, Abbyy will easily recognise and separate them one page at a time.



Image 6: Automatic page separation

In order to skip unnecessary symbols, we have cut out only the part containing text without black lines resulting from the scan, with no images, page numbers, headers and footers, which we will easily add later. Each page of the book is marked separately.



Image 7: Text selection

Subsequently, we chose the language of the selected book, click on the icon “Read”, and the entire scanned book will be converted into text. Upon identifying symbols, the resulting work is saved in Word and PDF which, in the “world” of book making, is called control PDF.

The Abbyy document is rather large, but it is good to keep it until the book is made, in case we delete something, so we do not have to repeat the character recognition again. The obtained work can be immediately saved to .epub, but the OCR is still filled with errors.

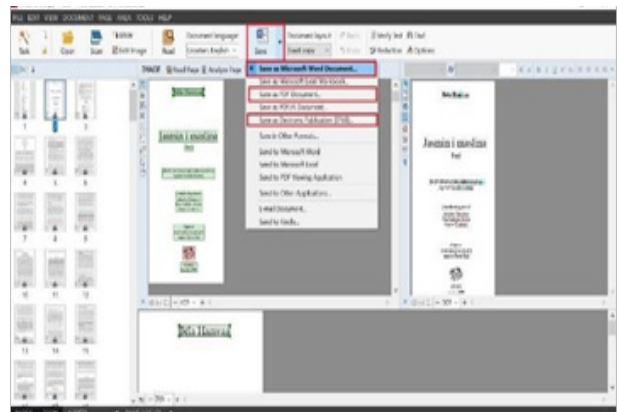


Image 8: Saving OCR

Text correction in Word

Text correction in Word is one of the most important procedures since a quality .epub can immediately be made with ease after the well-formed Word is made. (At least in the eyes of the reader, but, for a professional production, we need Sigil). We will

also demonstrate how to create .epub by using Sigil which is intended precisely for that purpose, making a quality .epub both in appearance and in content format, and editing for further changes according to certain needs can be performed with ease and speed.

During the first step, we set the language used in the book. If we want to perform a Spell Check, alongside the chosen language, the "ABC checkmark" needs to be shown.

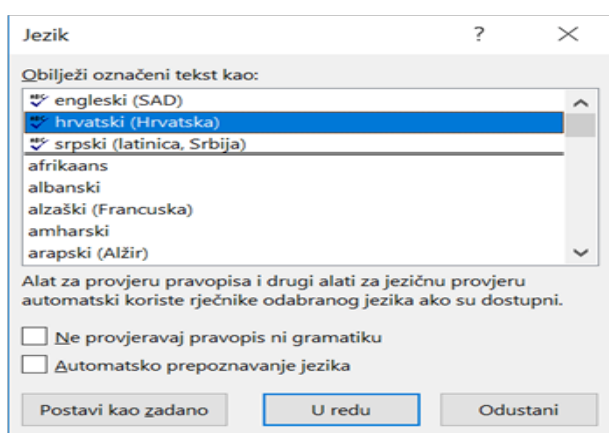


Image 9: Book language selection

The second step is to select the whole text (CTRL + A) and remove all formatting except Italic, with the font being the same size of 12pt. In order to get equal first paragraphs for the entire text which have 1 cm indent for the first row, we adjust the paragraph settings.

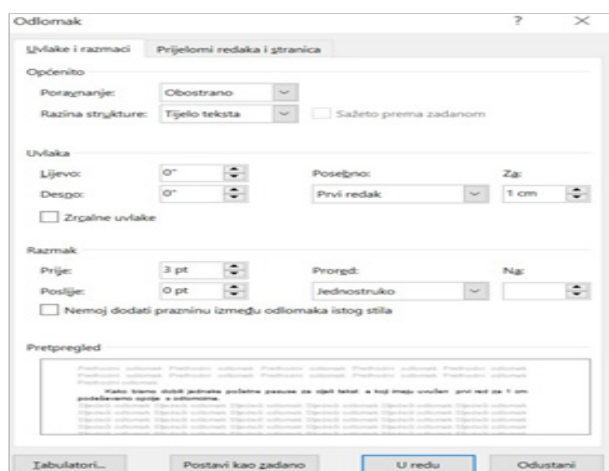


Image 10: Paragraph settings

Now is the time for correcting mistakes by using the option find and replace through which we can truly correct a lot of mistakes such as deleting double spaces, changing two commas into quotation marks, two hyphens into one, two apostrophes into a quotation mark, etc., but, in this example, we will mention only several important ones. An example of finding incorrect paragraphs:

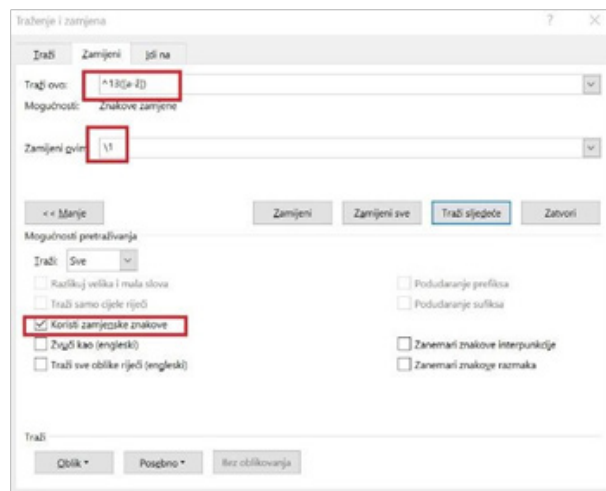


Image 11: Find and replace

Comments are added manually and Calibre will properly recognise and convert them for e-readers.

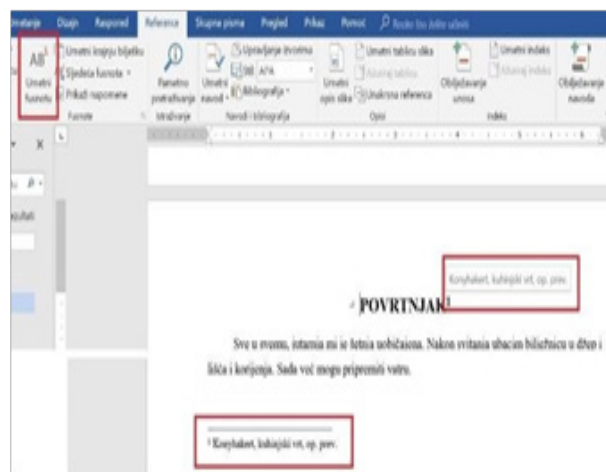


Image 12: Adding comments

Spell Check in the latest versions of Word is really good, but, there is no better verification than reading. This is precisely what we did, we read the whole book and corrected small mistakes that happened during OCR. The most frequent ones were (some could also be corrected through the find and replace method): rn

instead of m, č instead of ć, ii instead of ü and several others. Additionally, in the book itself, there can be grammatical errors made by the language editor, so reading is surely recommended. Alongside reading, we set all titles to "h1" and insert page breaks so that they will always be on a new page regardless of the change in content. After the book is read and corrected, it can be saved as .pdf, which is the final version of the PDF format. For further creation of an EPUB book, we save it as a filtered webpage. If the book is well-made in Word, it can be immediately converted into .epub by using Calibre and we will get a good looking e-book. First, we change the Metadata information and, through the option Convert, we change it into the needed format.

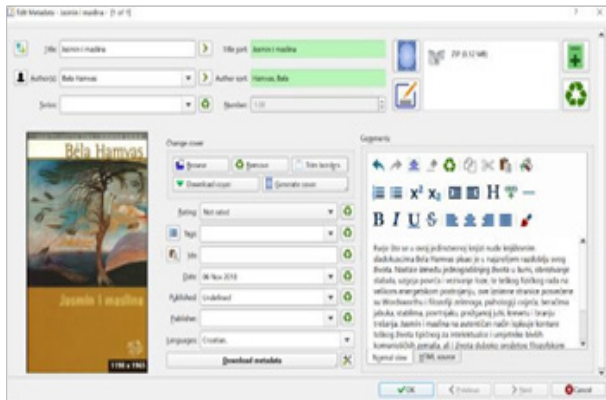


Image 13: Inserting Metadata

Producing and processing .epub

An EPUB document is a compressed file. You can look inside by using a compression programme. In order to do so, it is necessary to rename .epub into .zip, or .rar. WinRAR is used for that purpose.

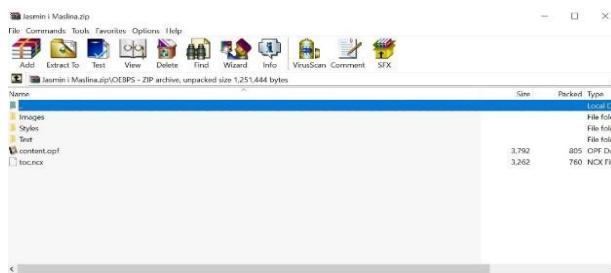


Image 14: WinRAR

In order to get a clean HTML document, .docx is opened in Alreader2 and saved as HTML. Now, apart from the titles (if they are well-made), Italics and Bold, all styles from Word are lost. It does not have all the unnecessary programme codes.



Image 15: AlReader2

Upon saving, we have obtained a HTML document which will be processed in Sigil.

Sigil-0.9.10 (the version used when writing this paper) is a result of numerous hours of volunteer efforts on creating the fastest, most stable and powerful version of the best free software available for editing EPUB. Sigil was started by Strahinja Marković as a school project in 2009, and, from 2011 to date, it was taken over by John Schember.

HTML (HyperText Markup Language) is the starting point for creating our .epub. HTML is the document file format in which webpages are written. HTML documents are text documents with tags to specify formatting. Tags are enclosed between "<" and ">" signs. For example, "<p>" is a tag that marks the beginning of a new paragraph and "</p>" is a tag that marks the end of a paragraph.

CSS (Cascading Style Sheets) is a method to tell the browser how each HTML element (paragraph, header, etc.) should look. One way to do this is to create a separate text file that specifies the format for each element and then include a reference to this file inside our HTML document. That file is called stylesheet.css. When text formatting is made according to the publisher's demands, the same CSS file can be used in other books with minimal change.

The most famous source of assistance with HTML and CSS is the w3schools website. [14]

If the content is properly generated in Word, Sigil will easily recognise it. For all changes, we use the option “TOC” (Table of Contents), as well as a separate .ncx (Navigation Control File) file, a format used for EPUB which e-readers recognise as content. Page numbers are not necessary in the content since they change with every increase or decrease of font or margins.

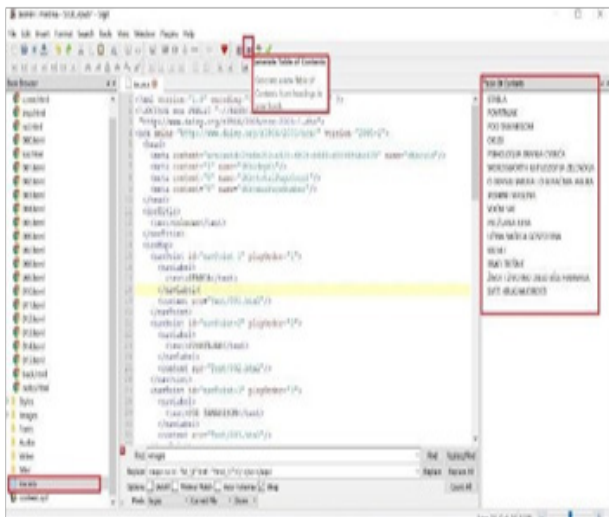


Image 16: Content in Sigil

Everything done in Sigil can be immediately seen, the changes can be seen by clicking the icon book view or code view and, in the same windows, we can easily and rapidly change everything that is necessary.

Additionally, in Sigil, it is possible to correct mistakes using the same method find and replace as in Word.

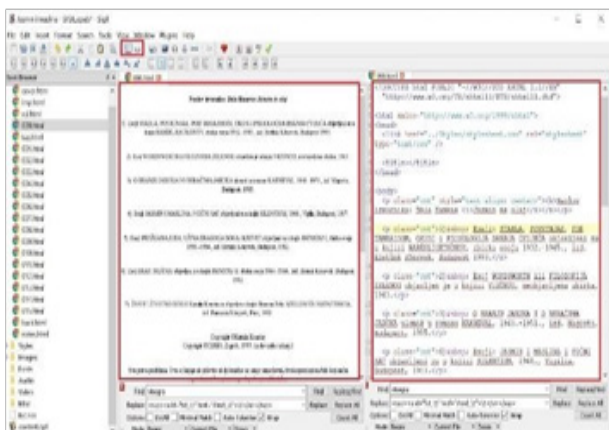


Image 17: Book and Code View

The latest Calibre versions also have their own editor where it is possible to do nearly everything with the code, but Sigil is still way better and simpler when it comes to this.

Indeed, it is possible to do a lot by only selecting the text and choosing the appropriate icon such as adding images, content, titles, alignment, etc. Adding the cover image is simplified, and all the images we have used in the book are located in the folder Images.

The cover image is selected and, with a right-click, we get the option Add Semantics > Cover Image.

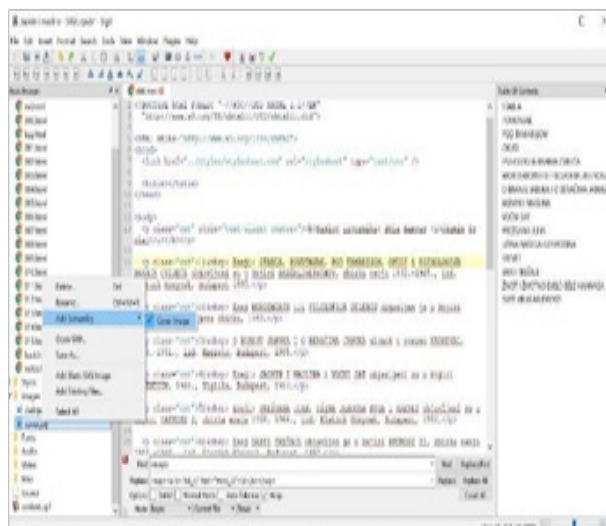


Image 18: Adding a cover image

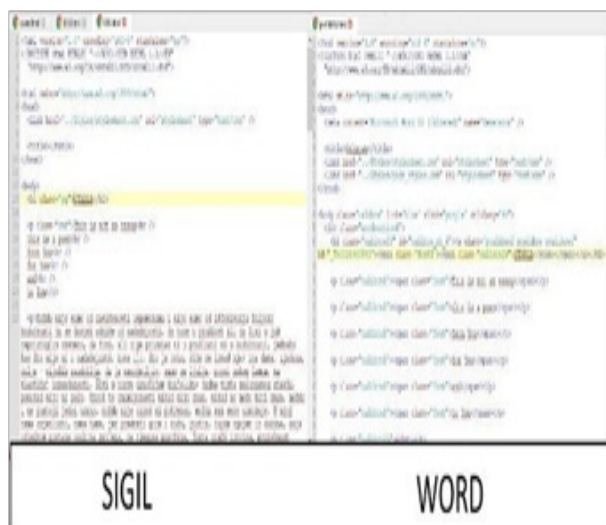


Image 19: Difference between Word and Sigil processing

Conversion

Calibre is used for reading EPUB (and other formats) on the computer. But, it is also the best programme that can convert .epub into any other format supported by our e-readers. In our case, it is most frequently the conversion of .epub into .asw3 Amazon format for the Kindle e-readers.

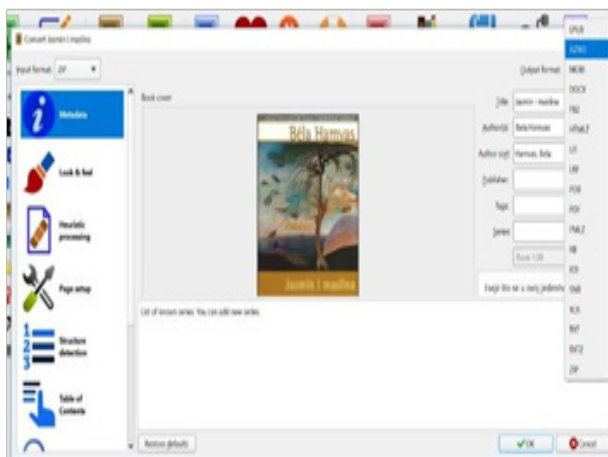


Image 20: Conversion into other formats

Calibre is a very versatile programme which converts our books into a digital library and sorts them by author, language, format, rating, etc. Calibre has weekly updates so that every couple of days we receive some news which make it even better and more popular. When we connect an e-reader with a computer, Calibre automatically recognises it and adds new icons for the e-reader, making it possible to directly send a book to the device through the programme.



Image 21: Sending a book to the e-reader

When we send the book, it is placed on our e-reader.

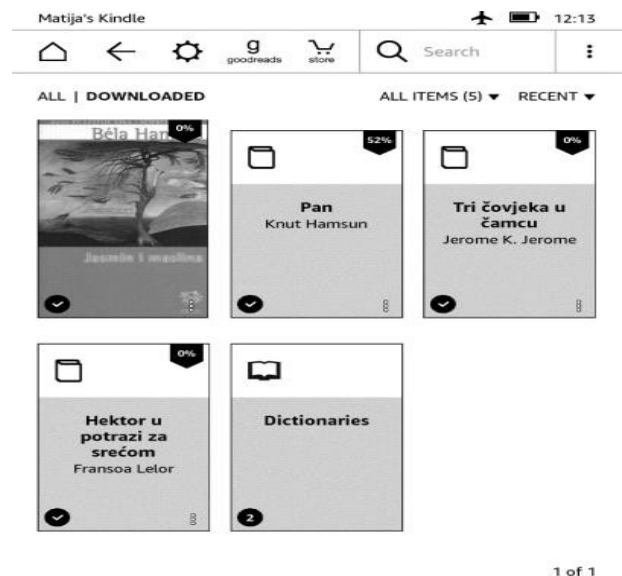


Image 22: Online book

STABLA

This is not an essay
this is a poem
from You
for You
and
to You

Možda nije samo od zasićenosti uspomena i nije samo od iščekivanja željene budućnosti da se čovjek odmiče od sadašnjosti, da tone u prošlost ili da živi u još nepristiglom vremenu, da živi, ali nije prisutan ni u prošlosti ni u budućnosti, jednako kao što nije ni u sadašnjosti, tone ili, što je isto, diže se iznad nje; ima dana, tjedana, dulje – cijelih razdoblja, da je neosjetljiv, samo se ljulja, poput nekog čamca, na vlastitoj ispunjenosti. Živi u posve drukčijem krajoliku; jedna vrsta polusnenog stanja, ponekad nije ni polu. Uzrok te omamljenosti nikad nije znan, nikad ne može biti znan, možda i ne postoji jedan uzrok, možda nije uzrok ni potreban, možda sam sebe uzrokuje. U njoj nema svjetlosti, nema tame, jer predmeti gore i noću, gustim, tupim sjajem iz dubina, boja odjednom postane osobina svijeta, ne njegova površina, forma zrači iznutra, protežnost osjetilnog tijela nije ohlađeni ukras. Život ne teče, nego plavi, pulsira u sebi samome, polaganim, mirmim pulsom, kao da tone u slatku, tamnu, omamljujuću tekućinu.

Opažanje i imaginacija ne mogu se razdvojiti – *man dichtet schon / wenn man erlebt*, međusobno se uljepšavaju i obogaćuju. Tih dubokih plavih boja nigdje nema, ili su samo bile, ili će samo biti – crvene su kao baklje koje gore u noćima bez zvijezda, u nekim dalekim dolinama – crna je poput baršunaste zmijske, koja čovjeka

Time left in chapter: 4 mins

8%

Image 23: E-reader image

Conclusion

Digital books have an ever growing significance where it is only a matter of time when they will mostly replace paper books. The proof for this lies in the fact that print editions of newspapers are being increasingly cancelled and published only in digital formats. We live in a time where technology advances rapidly and the privacy laws are changed, so, consequently, a question of copyrights over print editions DRM (Digital Rights Management) arises alongside the issue of unauthorised distribution where the development of modern

technology (encryption) is of great help. Through certain electronic book readers, it is possible to monitor users, disable printing, or edit the content. A question arises as to whether it is possible to completely protect an e-book from misuse. [03] For the time being, e-readers in Croatia are just a trend of a small number of technology and book lovers, and they do not have a strong connection with libraries. [13] E-books will be used all the more, and they will be increasingly present on the market. The production of electronic books with the mentioned programming tools is simplified and present on all software platforms

References

1. "Nikola Tesla", Nacionalni portal za učenje na daljinu (2016). Izrada e-knjige u EPUB formatu. [Internet] Available at: <https://tesla.carnet.hr/course/view.php?id=88> (accessed: 4th January 2018)
2. Unless-women (n.d.). Angela Ruiz Robles. [Internet] Available at: <http://www.unless-women.eu/biography-details/items/ruiz-robles.html> (accessed: 4th January 2018)
3. University of Zagreb – Faculty of Graphic Arts (2010) Elektronička knjiga. [Internet] Available at: <http://materijali.grf.unizg.hr/media/Elektronicka%20knjiga.pdf> (accessed: 5th January 2018)
4. E-laboratorij, Tomislav Negulić (2014). Calibre – kreiranje e-knjiga. [Internet] Available at: <http://e-laboratorij.carnet.hr/calibre-kreiranje-e-knjiga/> (accessed: 4th January 2018)
5. Goodereader, Michael Kozlowski, (2010). A brief history of eBooks. [Internet] Available at: <https://goodereader.com/blog/electronic-readers/a-brief-history-of-ebooks> (accessed: 5th January 2018)
6. History Cooperative, (n.d.). A History of E Books. [Internet] Available at: <http://historycooperative.org/a-history-of-e-books/> (accessed: 6th January 2018)
7. The Guardian, (2002). Ebook timeline. [Internet] Available at: <https://www.theguardian.com/books/2002/jan/03/ebooks.technology> (accessed: 6th January 2018)
8. Calibre, Kovid Goyal, (n.d.). Editing e-books. [Internet] Available at: <https://manual.calibre-ebook.com/edit.html> (accessed: 5th January 2018)
9. Zihor, Melita Vrsaljko (2013). Kako doći do besplatnih e-knjiga? [Internet] Available at: <http://www.zihor.hr/kako-doci-do-besplatnih-e-knjiga/> (accessed: 4th January 2018)
10. Škole, portal za škole, Ljerka Čulina (2015). E-knjige (prvi dio). [Internet] Available at: http://www.skole.hr/ucenici/ss?news_id=12571 (accessed: 4th January 2018)
11. Ricount, Danka Jurjević (2014). Uvod u e-knjige. [Internet] Available at: <https://ricountuniri.wordpress.com/2014/05/06/uvod-u-e-knjige/> (accessed: 6th January 2018)
12. Nabava, Anon, (2018). Sve što trebate znati prije kupovine e-čitača knjiga. [Internet] Available at: <https://www.nabava.net/clanci/vodici-za-kupovinu/sve-sto-trebate-znati-prije-kupovine-e-citaca-knjiga-282t6> (accessed: 4th January 2018)
13. Lončarević, Kristina. (2013). E-čitači i e-knjige: budućnost u knjižnicama ili samo trend? Graduate thesis. Faculty of Philosophy in Zagreb. Department of Information and Communication Sciences. [Internet] Available at: <http://darhiv.ffzg.unizg.hr/id/eprint/4346/> (accessed: 4th January 2018)
14. Mobileread, Pablo (n.d.), Quick and Dirty ePub Tutorial - v3. [Internet] Available at: <https://www.mobileread.com/forums/attachment.php?attachmentid=102683&d=1362837219> (accessed: 4th January 2018)
15. Sigil, Dave Heiland (2013), User guide. [Internet] Available at: https://github.com/Sigil-Ebook/Sigil/raw/master/docs/Sigil_User_Guide.epub (accessed: 6th January 2018)

Od elektroničke do papirnatu knjige.

SAŽETAK

E-knjiga je elektronička knjiga koju možemo čitati na uređajima poput računala, tableta, mobilnih telefona ili posebnih uređaja namijenjenih takvom čitanju. Oxfordski rječnik engleskog definira e-knjigu kao elektroničku verziju tiskane knjige, ali to se ne može smatrati ispravnom definicijom jer mnoge knjige postoje samo u digitalnom obliku bez tiskanog primjerka. Procesi izrade e-knjiga u softverskom i grafičkom dijelu su stvarno neograničeni i ovaj rad predstavlja samo jedan od njih u kratkim koracima. Namjera je da se korisnik samo upozna sa svime što je potrebno za proizvodnju i ne služi za piratstvo knjiga. Ako vam se sviđa knjiga, potičemo vas da je kupite i podržite autora.

Ključne riječi: e-knjiga, skeniranje, EPUB, Sigil, kalibar, Word, OCR, HTML, CSS

Received: February 28, 2018 / Accepted: March 09, 2018

Correspondence to: Matija Katalenić, Mihael Ladić, Petar Miljković, University North, Varaždin, Croatia

E-mail: matija.kat@gmail.com