Report on Backlist Data and gap analysis

Accessible Backlist Ebooks Laboratory (ABE Lab)

TABLE OF CONTENTS

Table des matières

Overview

Accessible Backlist Ebooks Laboratory (ABE Lab) is a Creative Europe funded project operated by three partners: the European Digital Reading Laboratory (EDRLab), the Fondazione Libri Italiani Accessibili (Fondazione LIA) and the National Library of the Netherlands (KB). It aims to provide publishers with reliable information about options and costs for remediation to make ebooks accessible as requested by the European Accessibility Act (EAA).

Context

To date, persons with reading disabilities - for instance blind, visually impaired or dyslexic people - only have access to 5 to 10% of the yearly book production, depending on the country they live in: the term "book famine" has been coined some years ago to describe this fact. The books they have access to are specific adaptations of print, created in small quantities by associations for the blind or similar organisations, or digital books

The European Accessibility Act (Directive (EU) 2019/882, acronym EAA)[1] will come into force in June 2025. This Directive promotes universal accessibility, which consists in taking into account all the needs and possibilities from the conception of the services, to be able to create quality services in this digital context. It is part of the Strategy for the rights of persons with disabilities 2021-2030, which addresses several major themes: accessibility, equality, education, social protection, employment and more. With it, Europe is committed to an inclusive policy that makes people with disabilities a core concern.

The publishing industry, worldwide, agrees that EPUB is the only digital book format which allows, with a few efforts, the creation of fully accessible ebooks. Most titles are already published in EPUB, and the first born accessible titles are now on sale.

There will still be an issue with most ebooks published before 2025. Since the objective of the Directive is to end the "book famine" endured by visually impaired people, it is necessary to transform as many titles as possible into accessible EPUB files. To do so, publishers need to have a clear view on gaps to be filled and related costs per type of ebook.

Remediating these ebooks can generate high costs, due to the lack of appropriate tools, the level of complexity of the different categories of books or the difficulties in converting them related to the fact that they were created when the standards and the formats did not include accessibility features. In that case there will be a high risk that in many situations publishers may remove those titles from the market, resulting in a great loss of diversity, affecting especially niche and rare titles.

To prevent hundreds of thousands of ebooks from disappearing from the market because of the accessibility requirements, but also to make sure that as many as possible ebooks get an accessible life, stakeholders need an independent analysis and low-cost remediation methodologies and tools.

[1] See online at https://ec.europa.eu/social/main.jsp?catId=1202

Objectives

The main objective of the ABE Lab project is to provide guidelines to European publishers for boosting the remediation of their ebooks currently on sale and convert them into accessible versions. All information about the project can be found on the ABE Lab website.

This document has two objectives The first is to provide stakeholders with insight into an overview of the number of ebooks actually on the European market and their repartition by categories, formats and years of production. This is addressed in the first main section *"Composition of the EU ebook backlist"*. The second objective is to provide a view of the gaps to be filled for the titles in the backlist to reach conformity to the EAA. It is covered by the second main section *"Gap analysis"*.

Composition of the EU ebook Backlist

Without precise data about the composition of ebook backlists in Europe, it is not possible to evaluate the best workflows and the cost of remediation of these ebooks. Without precise data about new releases, it is difficult to evaluate the potential evolution of the number of accessible ebooks before and after the European Accessibility Act comes into force, and therefore the success of the EAA in the ebook publishing industry.

The data presented here were necessary to effectively set up the next steps of the project and to define new research possibilities on the topic; they also provided a basis for future documents and reports that will be published in the context of the ABE Lab project.

This report presents the results of the analysis of the size and the composition of the ebooks backlists and their segmentation and provides trends and insights of use for the scope of our research. It first present our approach including the difficulties we faced and then provides overview and segmentation we achieved. The *Insights* section recapitulates the main findings and the *Outcomes* section details the use of this work for our project.

Approach

For this study we define ebook backlists in Europe as the collection of all ebooks available on the EU Market. Since there is no central source at the European level that provides this information, many publication registration offices, distributors, aggregators, resellers and national libraries were contacted. Given that all these organisations handle only a set of partial data and that these data sets have overlaps between them, we need to be careful to make conclusions. Ebooks made available on the EU market via e-commerce platforms, distributors and retailers operating on the international market also make up a large proportion of the backlist, especially considering ebooks from the UK, the USA and Canada.

We faced two main difficulties in the data collection and analysis activities:

- catalogues overlaps that we resolved going back to registration offices to help with filtering and make sure the true data appears;
- book categorization schemas[2] not harmonised across EU countries. We based our work on Thema, a recognized international standard used in many countries, but national classifications like CLIL in France or NUR in The Netherlands still exist and mapping them to Thema has been challenging even using provided mappings.

More details about these difficulties and how we managed them are given in the following subsections.

Counting titles

To get information about the number of ebooks that have been published in EU member states, we started by contacting all EU ISBN agencies. Summing together the numbers of national productions should give a good first estimate for the total. However there are some caveats:

[2] For data collection and analysis purposes, actors of the book value chain need a way to describe and categorise publications. That's what we call book categorization schemas.

- some titles are published in more than one digital format (like EPUB, PDF, KF8), which results in counting manifestations[3], not unique works;
- some ebook titles get a new ISBN with a new version, so again we count not unique works, but different manifestations of the same ebook;
- some ISBN agencies do not assign ISBNs on a per title basis, instead they give publishers a large range of numbers to use. As a result, these ISBN agencies do not have a record of how many titles are published;
- not all ISBN agencies are informed when ebooks are no longer available on the market.

To elaborate a more complete estimate of what is on the EU market, it is not sufficient to know what is published locally, but we also have to take into account what is imported from abroad. Especially with certain categories of non-fiction ebooks (like computer books), we see that the local production is rather limited and imported ebooks in English are dominant. A part of these titles are distributed via local distributors and sold by local retailers, but next to that we also see online platforms that operate worldwide, such as Amazon, Kobo, Google and Apple. To take this scenario into account when calculating the size of the backlists, we contacted distributors[4], aggregators[5] and resellers[6]. There are caveats here as well:

- some distributors and aggregators handle only a part of the market, for example only trade books or scientific books, or only books in certain languages, meaning that relying on data provided by a single distributor or aggregator is not sufficient to obtain a complete overview, and we must therefore extend the collection to multiple parties. In addition, when we receive aggregated data and not detailed information to the individual ISBN level, it is impossible to know if different collections overlap and how relevant this overlap is, so we need to be careful with adding these together;
- some distributors and aggregators supply both titles from non-EU publishers and titles published in an EU country. Without detailed data and information it is impossible to distinguish between titles that are published in the country itself and titles produced by foreigner publishers, which also might have an overlap with the titles we count in other EU countries;
- retailers sometimes collect titles from multiple distributors and aggregators, so it is also not possible to rely purely on the numbers resellers provide; platforms that operate worldwide hardly provide detailed information about their collections. And since they operate in a different way, their collections are not even directly comparable to those of the traditional ‘publisher-distributor-retailer’ chains. For example, Amazon does not only provide ebooks with an ISBN, that is, produced by publishers, but also a lot of self-published titles, often without an ISBN.

[3] With manifestation we refer to a physical or digital embodiment of a work as defined in bibliographic record standards (https://www.loc.gov/marc/marbi/2009/2009-01-3.html). As example in the digital world an EPUB and a PDF of the same title are counted as two different manifestations.

[4] We define a distributor as an entity that collects and distributes files to selling platforms. The distributor also establishes or collects metadata and sends them to aggregators.

[5] We define aggregators as an entity who collects established metadata and distributes them to selling platforms.

[6] We define resellers as the entity in direct contact with the client.

National libraries are another possible source of information about ebooks published in a country, and sometimes they also play a role in providing ISBNs to publishers. In fact, in some cases, when we requested data from the official ISBN agency, it was the national library of the country that provided us with the answer. Sometimes we even had direct contact to get information on the availability of ebooks. Especially in countries where publishers have the legal obligation to deposit a copy of their publications at the national library, it may have an overview of what is published in the country and therefore represents a valuable source of information.

Classifying titles

Categories

Book categories are a way to isolate and regroup books with common characteristics. Those characteristics may explain production choices and the technologies used.

To group titles for the purposes of our project, we chose to refer to the Thema subject category scheme[7] established and maintained by EDItEUR. Thema aims to be the scheme for a global book trade and is currently the most commonly used classification methodology. Even if the use of Thema classification has progressed a lot in the last few years, it must be noted that is not the only classification used, and since it is still relatively new, not all ebooks have been assigned Thema codes yet.

To resolve this inconsistency, EDItEUR provides a series of documents mapping codes from different book schemas to Thema codes.[8] Because standard classifications have different logics, those mappings may not be one-to-one and sometimes need interpretation. One example is the CLIL[9] to Thema mapping, where the Young Adults category can be mapped to two main Thema categories: Fiction (Thema code F) and Children, Teenage and Educational (Thema code Y).

As a consequence, we had to spend time on understanding differences in classification methodologies and which choices to apply to mappings.

Formats

Ebooks can come in different formats. For the purposes of our project, we focus on the two mainstream formats widely adopted worldwide: PDF and EPUB.

PDF (Portable Document Format) is a document format developed by Adobe for document interchange, and often used for digital documents with a complex layout or when it is necessary to reproduce the structure and graphics of a paper document or book. Initially developed for print reliability and digital conservation, the format is based on PostScript, a computer language for describing the positioning of characters and graphic elements absolutely on the page (similar to having x and y coordinates to position each element on the page). As it went to be used for digital consultation, the format evolved to respond as best as possible to this use principally with the

[7] *Thema – the subject category scheme for a global book trade version 1.5*, EDItEUR, 2022. Available at https://ns.editeur.org/thema/en

[8] *Thema mappings*, EDItEUR, 2023. Available at https://www.editeur.org/151/Thema/#Mappings

[9] *Commission de Liaison Interprofessionnelle du Livre*, the French standard for book classification. Available art https://clil.centprod.com/listeActive.html

addition of a semantic descriptive layer composed of XML[10] language tags. The variety of tagging that can be added is currently limited to 28 elements[11]. The PDF/UA ISO[12] standard provides definitive terms and requirements for accessibility in PDF documents and applications.

EPUB (Electronic PUBlication) is an open file format for electronic publications based on Web Standards (HTML, CSS JavaScript). The first version, named OEBPS 1.0 (Open EBook Publication Structure) was approved in 1999 by the Open eBook Forum, which later became the International Digital Publishing Forum[13] (IDPF). EPUB 2.0 was released in 2010, followed in June 2014 by version 3, in which for the first time in the specifications were included accessibility features deriving from the Daisy specialized format.[14] Just after the release of EPUB 3.1 in January 2017, the IDPF merged into the World Wide Web Consortium (W3C), making the EPUB an official W3C recommendation (standard). The last version, EPUB 3.3, was published on May 25, 2023[15]. EPUB is a native semantic format allowing the use of numerous taggings from different standardised languages such as HTML, ARIA, MATHML, SVG and others.

Years

Another way to classify titles is by year of production. We realised that book categories and formats are not enough and that we also had to make a per-year segmentation. As already mentioned, over the years formats have also evolved in terms of the accessibility features supported. The level of accessibility also depends on the accessibility guidelines available at the time the ebook was produced. The first version of EPUB Accessibility 1.0, the guidelines for creating accessible EPUBs, for example, was published in 2017. We can therefore assume that EPUBs created before this date will have a very low or zero level of accessibility, as publishers lacked clear reference specifications at the time of their production.

The increasing focus on the accessibility of digital content, including ebooks, has also led to improvements in ebook production tools, which in the last years have progressively introduced support for accessibility features, to allow publishers to produce ebooks that are more and more accessible and compliant with international guidelines. In parallel, production workflows have also evolved to take accessibility into account.

All these aspects - the format and its version, the availability of accessibility guidelines, the support of accessibility by production tools and the adaptation of workflows - are reflected in the way ebook files have been created over the years. For building a representative sample set for our research, we have to take this development into account.

[10] Extensible Markup Language (XML) is a markup language that provides rules to define any data. It is standardised by the W3C and can be found at https://www.w3.org/TR/xml/

[11] A list of Standard PDF Tags is available at https://helpx.adobe.com/acrobat/using/editing-document-structure-content-tags.html

[12] Available at https://www.iso.org/standard/64599.html

[13] https://idpf.org/

[14] DAISY Format. Available at https://daisy.org/activities/standards/daisy/

[15] What is an EPUB file? Available at https://www.edrlab.org/open-standards/epub/

Overview of the EU ebook backlist

We have collected direct basic data from 18 countries and detailed data from 5 countries. The collected data provide precise information about the number of titles and the detailed ones contain repartition by categories and formats. We've integrated those collections with the annual data published by the Federation of European Publishers (FEP) to get an idea of markets.

A considerable amount of ebooks currently on the EU market are provided by e-commerce platforms and resellers operating on the international market. Many of these titles are from countries like the USA, UK and Canada. Since these operators are very unlikely to make their data available, ebooks on the EU market from countries outside the EU are not investigable. However, given the market knowledge we already have, we can assume that the trends identified for the data we have available are also applicable to the data of titles from outside the EU.

Since important differences exist between EU countries when it comes to their ebook backlist and this could appeal to different conclusions, we first present a general overview of the data collected and of the growth in ebook production at the European level, then we present more detailed views per catalogue.

Total

The last available FEP annual statistic report[16] presents data from 2021. Based on declarations from the national book publishing associations, it reported a total of 13.4 million titles available in the active catalogues of European publishers. 3 million of them are ebooks, representing 22% of the titles on the market, for 12% of sales.

Based on the summation of the numbers we collected from individual EU countries, we established that, in early 2023, the backlist of ebooks available in the EU market exceeded 3.5 million.

Since ebooks can be easily traded cross-border, the actual number of titles available to consumers is much higher (see section Titles from outside EU).

Catalogue growth

The evolution of the book market as reported by the Federation of European Publishers[17] is very important to the topic we are discussing. Between 2005 and 2021 we have seen a steady increase in the number of book titles in commerce (+260%). While the increase in the number of new titles has been affected by the COVID-19 pandemic, this does not apply to ebook titles in commerce.

Our data collection reflects strong differences between territories, as we identified in the detailed data provided:

- in France 42.000 unique titles were added in 2013 compared to 115.000 titles in 2022;
- in Italy the production is stable and steady, with 29.000 unique titles added in 2013 versus 24.000 titles in 2022, but the years 2014 to 2021 had around 35.000 titles per year;
- in the Netherlands the number of ebooks added to the market has been relatively stable over the years. In 2012 some 10.000 titles were added, and in 2022 this was around 8.000.

[16] *European Book Publishing Statistics 2021*, FEP 2022. Available at https://fep-fee.eu/European-Book-Publishing-1467

[17] *European Book Market Statistics 2021-2022*. FEP, 2022. Available at https://fep-fee.eu/-Publications-

Number of titles in the backlist per countries

The following list presents the data we've collected and consolidated. It is organised per descendent number of ebooks made available per country. We observed that the detailed data obtained came from 5 of the 6 countries with the biggest backlists.

- **Germany: 1.055.369** currently in distribution (source: MVB GmbH). We notice overlaps with countries with active German speakers (Austria, Switzerland).
- **France: 952.416** currently in distribution, for a total of 1.310.274 ebooks registered (source: Dilicom). We notice exchanges with Canada (Quebec) and overlaps with other countries with active francophones speakers (Belgium, Switzerland).
- **Italy: 376.097** (source: IE - Informazioni Editoriali).
- **Spain: 336.757** (source: Dilve).
- **Poland: 138.415** registered (but expected higher) ISBN service at Bibliotheka Narodowa .
- **The Netherlands: 102.000** registered at ISBN.NL and 70.000 available from CB Logistics.
- **Czechia: 126.229** ebooks registered at Czech National ISBN agency.
- **Sweden: 107.561** ebooks registered at National Library of Sweden, but actual numbers expected significantly higher.
- **Hungary: 84.571** registered at National Széchényi Library.
- **Denmark: 81.324** Danish titles available from Publizon (but only 39.388 ebooks registered at DBC Digital).
- **Portugal: 58.000** ISBNs associated to ebooks according to APEL.
- **Greece: 30.059** ebooks registered at Greek ISBN agency.
- **Lithuania: 18.300** ebooks registered at Lithuania National Library.
- **Slovenia: 17.868** according to Slovenian ISBN agency.
- **Estonia: 11.685** according to Estonian ISBN agency.
- **Bulgaria: 14.129** ebooks registered at ISBN Bulgaria.
- **Latvia: 6.739** ISBNs assigned to ebooks, National Library of Latvia.
- **Ireland: 4.914** Irish titles according to Nielsen Bookdata.
- **Malta: 583** (source: National Book Council).

Missing data

- Austria
- Belgium
- Croatia
- Cyprus
- Finland
- Luxembourg
- Romania
- Slovakia

Titles from outside EU

Buying ebooks from non-European retailers is easy since the deliverable is a file that does not pass through border controls. It is especially the case for ebooks in the English language, which are very popular for certain categories (like computer books or scientific publications) and often outnumber local productions in these categories.

Since the EAA targets ebooks on the European market, we focused on data on ebook sales by European retailers. It is important to notice that international e-book e-commerce platforms operate on the European market too, but getting accurate data from them is not simple as they operate out of traditional distribution channels.

We provide here a quick overview of the collections available through those platforms:

- **Amazon Kindle.** Though Amazon does not publish exact numbers about ebooks that are available for the Kindle devices and the Kindle app, some sources estimate that more than 14 million titles are currently available[18]. However, as this number also takes into account a large number of titles without an ISBN - mainly self-published ebooks -, we can not just add this number to our estimate directly;
- **Apple Books.** Exact numbers are not available, and availability differs from country to country due to licensing deals. Despite this, Apple is known to offer millions of ebooks and audiobooks;
- **Google Books.** According to some sources, Google Books offers more than 40 million ebooks in 50 Languages[19], including 10 million ebooks for free.[20] Here we have the same issue: this number can not be compared directly to the backlist estimate as we defined it;
- **Kobo.** Kobo claims to have over 5 million ebooks and audiobooks available for reading directly on their e-readers and apps[21]. In some countries, they work together with local retailers (like Bol.com in the Netherlands or FNAC in France) to provide subscription services including lots of local content (in the case of Kobo Plus offered by Bol.com, about 'hundreds of thousands').

[18] How Many Ebooks Are There In The Kindle Store On Amazon? Just Publishing Advice, 2023. Available at: https://justpublishingadvice.com/how-many-kindle-ebooks-are-there/

[19] *How the Google Books team moved 90,000 books across a continent*. Ari Mariani, 2023. Available at https://blog.google/products/search/google-books-library-project/

[20] *About Google Books – Free books in Google Books*. Available at https://www.google.com/intl/en/googlebooks/about/free_books.html [Consulted on may 2023]

[21] About Kobo. Available at https://www.kobo.com/us/en/p/aboutkobo

Segmentation

By category

To compare the different EU member state backlists, we needed to split the complete offer of ebooks into several categories, like fiction books, biographies, children's books, books on art, etc. Since we expect that different categories of publications may have different complexity and often specific accessibility issues, we wanted to make sure we have a good representation of ebook categories and genres in this study.

Since many different categorization methods are in use, we needed to standardise as much as possible. For this research, we chose the Thema categorization scheme since this is becoming more and more the international standard many publishers and retailers use. However, given that Thema is relatively new and not all e-books have been assigned Thema codes yet, some mapping from older schemas had to be applied to make proper estimates (see title Categories of the *Approach* section for details).

We observed that the ventilation of ebooks per Thema code in the 5 markets for which we had complete data shows a strong disparity and does not allow for a European level modelisation. For example, in Germany, non-fiction represents about 86%, whereas in the Netherlands it is considerably less: 58%. Fiction ebooks (Thema code F) are the most represented category everywhere but with a range from 13% (Spain) to 42 % (The Netherlands).

The following table and figures show the different shares by Thema code for the 5 countries who provided detailed data.

Table 1: percentage of titles per Thema codes in the backlists of France, Germany, Italy, Spain and The Netherlands.

Thema code	France	Germany	Italy	Spain	Netherlands
A: The Arts	1,74	1,96	4,40	3,71	1,31
C: Language and Linguistics	0,44	2,60	0,95	2,09	0,73
D: Biography, Literature and Literary studies	19,93	5,98	10,55	8,36	5,76
F: Fiction and Related items	20,24	14,81	37,87	13,98	42,82
G: Reference, Information and Interdisciplinary subjects	0,00	4,87	0,25	1,02	0,54
J: Society and Social Sciences	9,70	9,02	8,97	12,63	5,95
K: Economics, Finance, Business and Management	2,44	8,91	3,72	4,35	5,86
L: Law	0,77	4,50	2,57	5,30	5,06
M: Medicine and Nursing	0,34	6,29	1,66	7,96	1,35
N: History and Archaeology	4,51	3,22	3,31	4,32	4,15
P: Mathematics and Science	0,87	7,08	1,10	2,43	0,87
Q: Philosophy and Religion	2,67	4,75	5,70	5,94	5,51
R: Earth Sciences, Geography, Environment, Planning	0,54	1,97	0,54	1,10	0,17
S: Sports and Active outdoor recreation	0,00	0,58	0,70	0,77	1,04
T: Technology, Engineering, Agriculture, Industrial processes	0,48	5,00	0,76	1,98	0,18
U: Computing and Information Technology	0,25	3,65	0,67	0,93	0,90
V: Health, Relationships and Personal development	0,00	3,26	6,65	2,42	2,58
W: Lifestyle, Hobbies and Leisure	6,78	3,30	4,04	2,02	3,71
X: Graphic novels, Comic books, Manga, Cartoons	9,74	0,96	1,47	0,62	0,08
Y: Children's, Teenage and Educational	8,19	3,41	4,13	18,07	9,37
Unknown	10,37	3,86	0,0	0,00	2,08

Figure 1: ebooks per Thema codes, France.

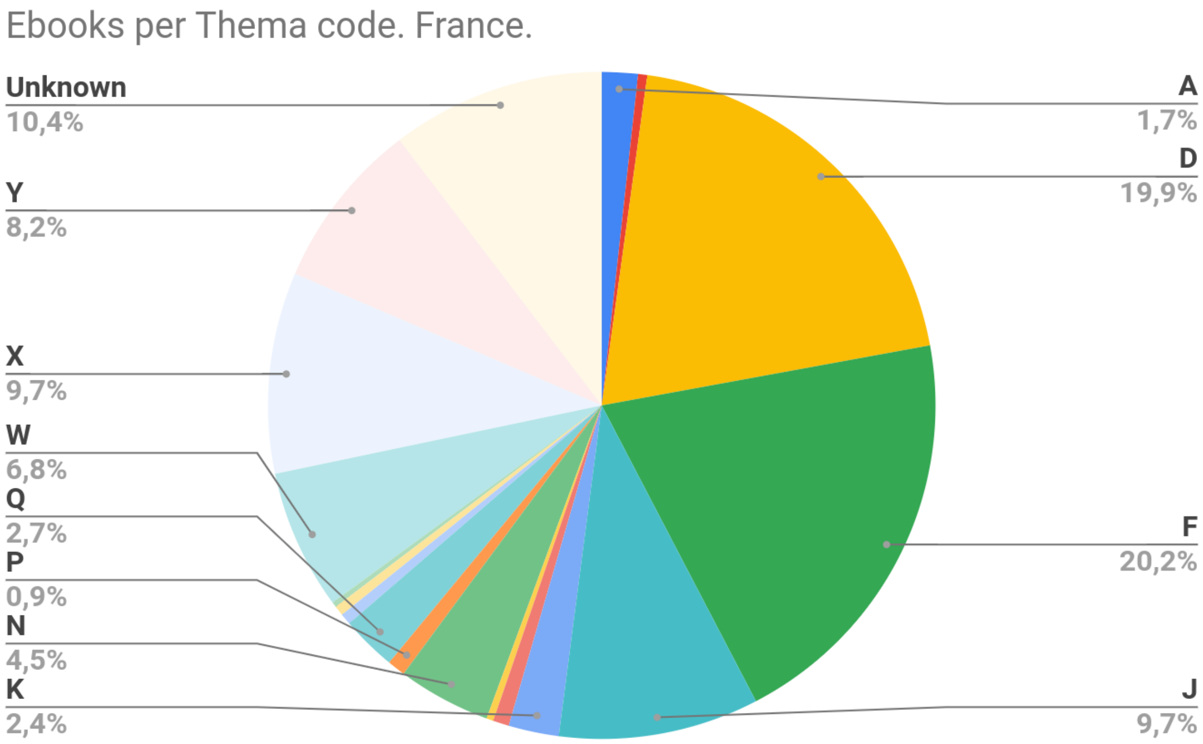


Figure 2: ebooks per Thema codes, Germany.

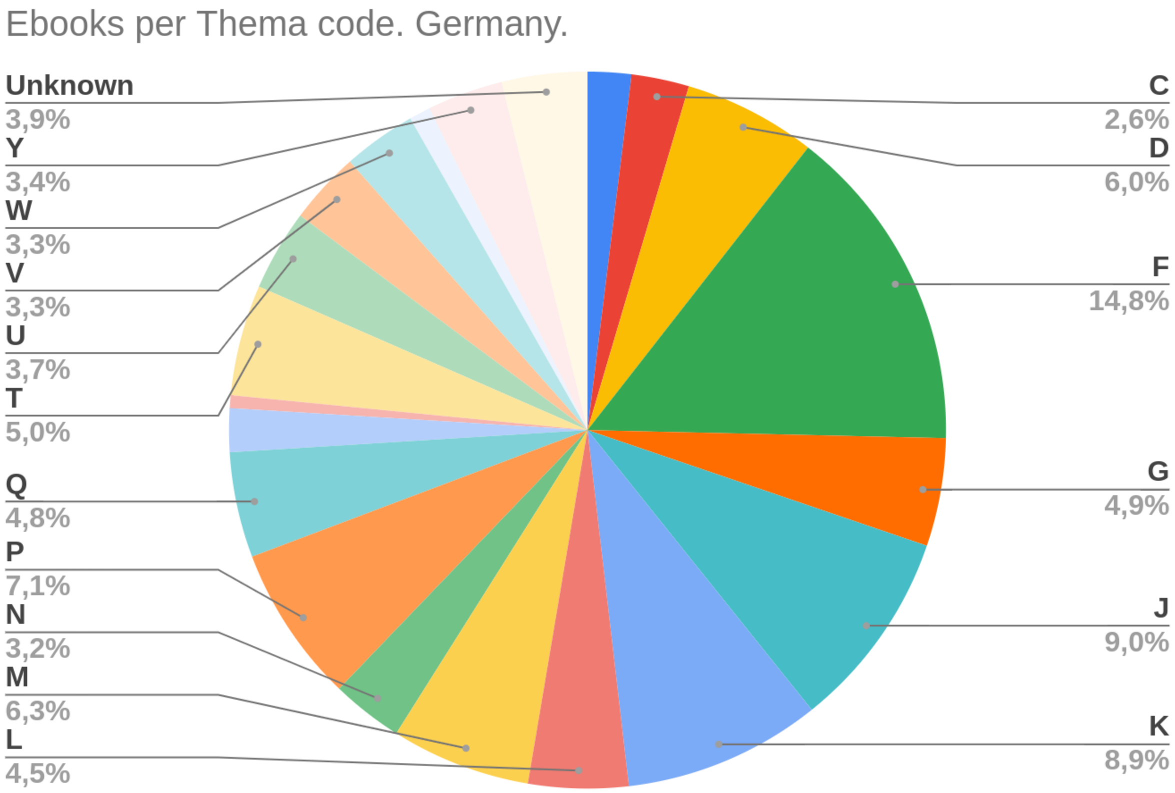


Figure 3: ebooks per Thema codes, Italy.

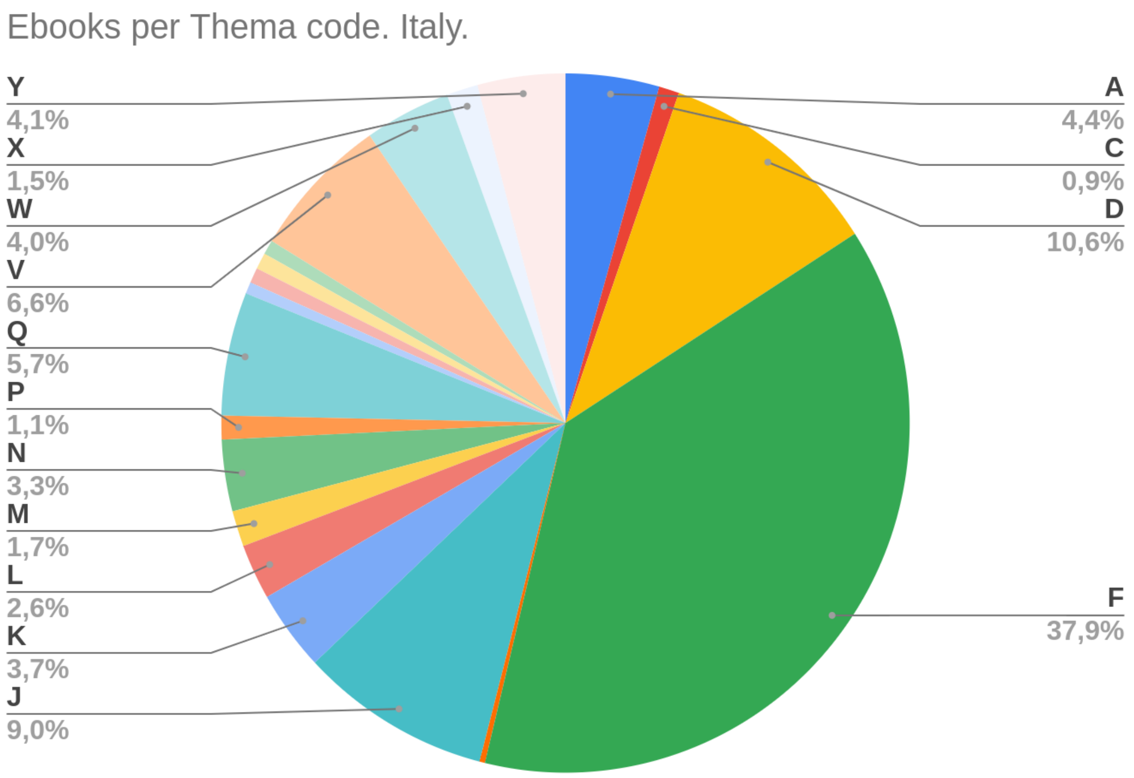


Figure 4: ebooks per Thema codes, Spain.

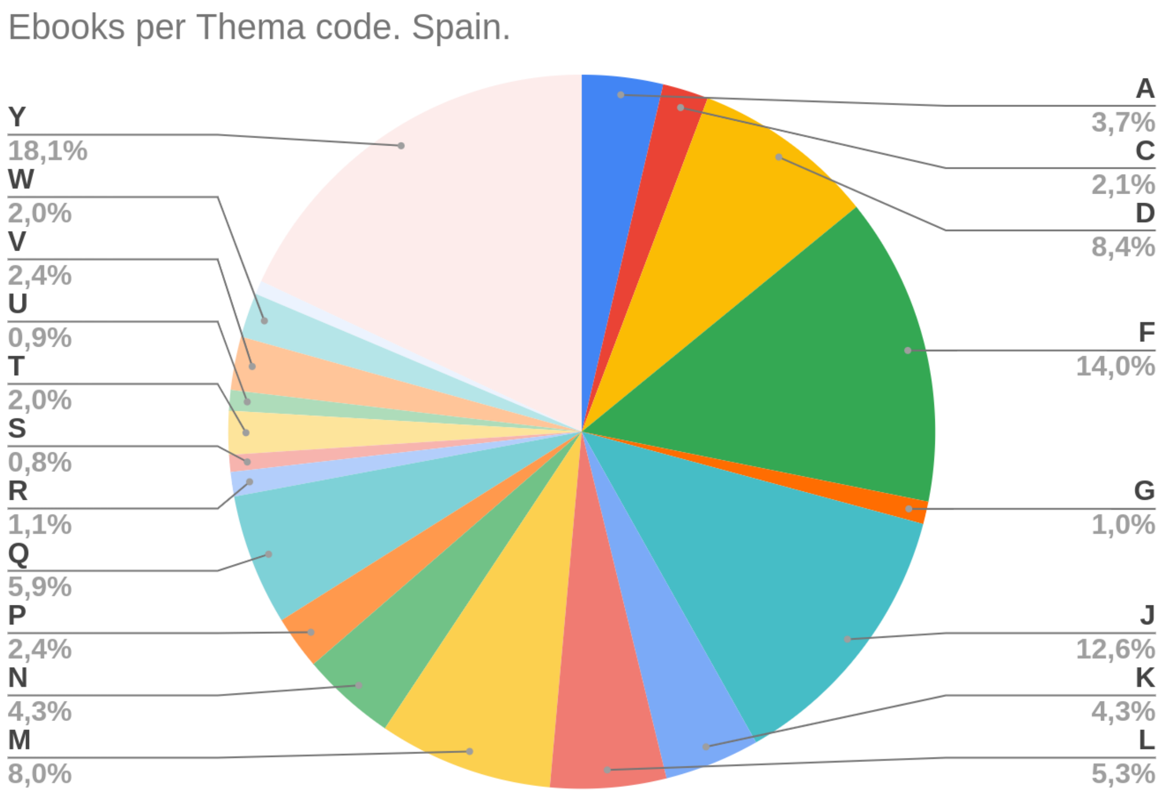
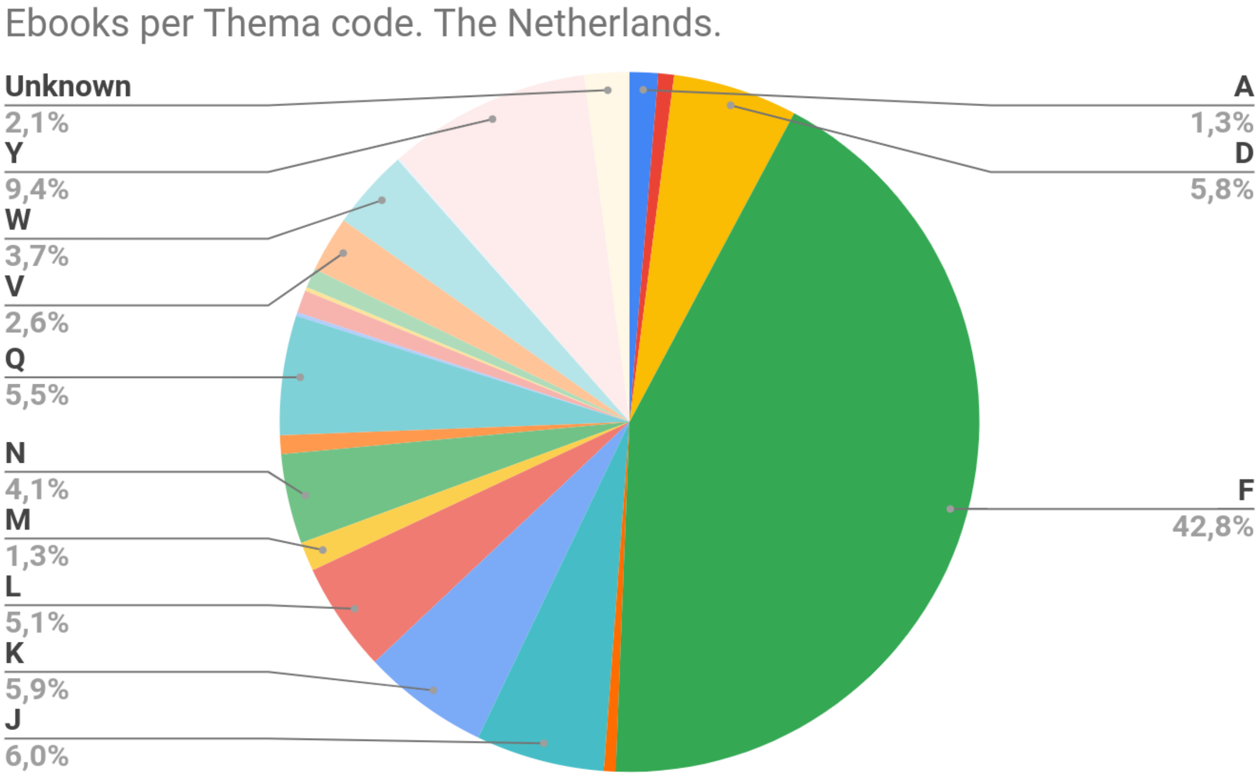


Figure 5: ebooks per Thema codes, the Netherlands.



By format

We managed to obtain detailed data on digital formats of ebooks on the market only for 5 key markets: France, Germany, Italy, the Netherlands and Spain. We retained only mainstream formats[22]: EPUB and PDF.

The segmentation of the market by format shows very diverse situations, where the German backlist has only 3% of EPUB3 files, but 60% of PDF files, while France and Italy reach nearly 40% of EPUB3 files, versus less than 25% of PDF files.

This disparity does not allow us to make assumptions about the European market as a whole. Consequently, we did not integrate any format-related query in our wishlist for files to be collected for the remediation tests.

The format repartition will have to be studied in more detail at the national and publisher level in order to refine the remediation cost estimate.

Table 2: % of titles per distribution format in 2022.

Market	% of EPUB2	% of EPUB3	% of PDF	% of other formats (ie. HTML, Apps, etc.)
France[23]	12	38	22	18
Germany	34	3	60	3
Italy	36	40	23	1
Netherlands	75	10	15	0
Spain	20	15	40	25

[22] An abstract of ebooks formats is given as Annex Ebooks files formats

[23] Total for France is not heading to 100% because 10% of the titles are in bundle sales including both PDF and EPUB3 formats.

By year

We captured the evolution of distributed files formats by year since 2012 for the 5 key markets. The disparity observed can be compared to that already described in relation to the percentage of titles per distribution formats in 2022. Some countries present a linear evolution reflecting the evolution of formats, while others seem to produce the same ebook formats in 2012 and 2022. This difference will affect remediation activities since countries in which the latest version of the EPUB format, EPUB3, which in recent years has become not only the format of choice for ebook production, but also and especially for the production of accessible ebooks, has not been adopted yet, will face a technological debt in addition to the necessary remediation efforts to make ebooks from the backlist properly accessible and therefore keep them on the market.

Table 3: Evolution of distribution formats from 2012 to 2022

Market	EPUB2 (2012 / 2022)	EPUB3 (2012 / 2022)	PDF (2012 / 2022)
France	-11 (from 23% to 12%)	+13 (from 25% to 38%)	-11 (from 33% to 22%)
Germany	-1 (from 35% to 34%)	+3 (from 0% to 3%)	-2 (from 62% to 60%)
Italy	-12 (from 48% to 36%)	+34 (from 6% to 40%)	-22 (from 45% to 23%)
Netherlands	-8 (from 83% to 75%)	+9 (from 1% to 10%)	-1 (from 16% to 15%)
Spain	Missing data	Missing data	Missing data

Insights

From the data collected we learn that the European backlist is not homogeneous between countries and therefore cannot be averaged and addressed in a similar way. Even a repartition by size of the national backlist is not sufficient to separate different needs.

We also learn that developments in digital publishing go slow. Newer formats (like EPUB3) are not adopted quickly, and newer possibilities (like fixed layout for EPUB) do not cause formats like PDF to be replaced quickly.

A blind spot is caused by the fact that large international platforms offer a lot of content to the European market. We do not know how much those contents are exclusives (like self published titles) and therefore the needs of remediation for those titles cannot be studied.

Outcomes

Based on the backlist overview we decided what type of titles we needed to collect. As previously mentioned, different categories do often need different types of remediation and generalising between these would give an incorrect view. As a result we decided that it was important to get titles of the following types:

- fiction and other ‘mostly text’ publications (including not illustrated children’s books and biographies): we expect those to be mainly reflowable EPUB files with structure issues to address. And within these, we expect older publications posing different challenges compared to the more recent ones;
- illustrated children books and graphics novels: we expect those files to be fixed layout ebooks with strong graphical accessibility challenges and therefore a need for textual alternatives;
- non-fiction publications: we expect to find complex elements in those files like tables and visual resources;
- complex layout publications: we expect those files to be mainly in PDF format with a strong tie between the form and the content with remediation needs including major changes.

As a per format request was not possible, we’ve built a per category wish list[24] divided in 4 time periods :

1. before 2011 where all EPUB files will be EPUB2;
2. 2011-2018 when EPUB3 was available but accessibility principles were not still well understood;
3. 2018-2021 when some publishers started to produce born accessible files ;
4. 2022 and after to represent today’s state of publishing.

[24] Available as annex to this document: ABELab sample collection wishlist.

GAP ANALYSIS

The objective of this section of the report is to share results on the identification of recurrent accessibility issues per categories of ebooks that will need remediation to fit EAA requirements. We'll first define the target, our methodology and the scoring threshold established for this analysis, as well as the identified biases and limits. Then we'll present the results we deemed useful for the next steps of the ABE Lab project, with a list of recurrent accessibility issues detected. Lastly, we'll define the outcomes of this work and the ebook classification we developed which will be used to test remediation tools and workflows.

Target

The European Accessibility Act (EAA)[25] requirements for ebooks are listed in Annex I sections III and IV Linea f). EPUB Accessibility - EU Accessibility Act Mapping[26] is a W3C group note that shows how EPUB files conforming to EPUB accessibility guidelines (EPUB Accessibility 1.1and WCAG 2.1 AA) are responding to the European Accessibility Act requirements. Those two documents help us define our target and basis to establish the accessibility deficiencies.

Pre-paginated ebooks (like PDFs and Fixed Layout EPUBs) do not comply with the EAA requirements for the criterion of *flexibility and choice in the presentation of the content*[27], a key functionality for persons facing cognitive difficulties (like dyslexia) or with sight impairments. As remediation for these types of ebooks would mean a change of format, we choose to introduce middle-way target remediation to allow the study of remediation to today's format state and possibilities offered by remediation tools. The target for these documents will be compliance with the Web Content Accessibility Guidelines (WCAG) 2.1[28]. In addition, PDFs will have to reach PDF/UA conformity, a dedicated standard registered as ISO[29].

Methodology

The backlist data analysis allowed us to define a wish list of categories of files to collect in order to represent the backlist composition in a small but consistent sample. We had a target objective of 200 files from 5 countries. This objective was exceeded, with 351 files collected from 7 EU countries (Denmark, Finland, France, Germany, Italy, the Netherlands, Spain), additionally including some samples also from the United Kingdom. We used the Thema codes as our reference for classification. Some provided samples were classified to different Thema categories by the publisher. As it was not possible to separate the Thema categories, we chose to multiplicate these samples (one per Thema code, i.e. a book with Thema codes D, F was analysed two times, one as D and one as F), resulting in a total of 376 units to analyse.

We added to this sample one accessible EPUB3 target file[30], to be sure that the gap emerging from our analysis was fitting the reality and that files already made accessible would not be considered as files with remediation needs. This target file was produced by LIA as born accessible in 2023.

[25] A detailed list of Norms and Standards is available as annex to this document.

[26] Available at https://www.w3.org/TR/epub-a11y-eaa-mapping/

[27] EAA Annex I, Section IV, f) iii) available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0882#d1e32-100-1

[28] On October 5, 2023 version 2.2 of the WCAG was officially released as a W3C recommendation. This update does not impact or compromise the analysis, research work and testing carried out in the context of the ABE Lab project.

[29] PDF/UA standard, ISO 14289-1:2014: https://www.iso.org/standard/64599.html.

[30] Target file publicly available at https://www.fondazionelia.org/wp-content/uploads/2023/10/European_Stories_2023_EUPL.epub.

We first established a list of key points indicators (KPI) we wanted to evaluate and from them, we could determine the data to extract from the samples. We verified which of these data were available from existing reporting tools (EPUBCheck, ACE, and RGTK for EPUB files; VeraPDF and PDFIX for PDF files, see annex *Tools used in the automated analysis* for a brief description of these tools) and determined the missing ones. Fondazione LIA developed a script to extract the missing data, aggregate all the data, and export a unified report. The details of the tests are available in the *Detailed evaluation of the tests made on ebooks* document, available for project partners and contributing publishers. 15 iterations of the script were made to refine data extraction and the exported reports. We started from a large number of data collected to stretch to a minimum necessary point.

The report was then used to develop calculation methods to define remediation complexity indicators[31]. Iterations were needed for this step as well, as data visualisation produced helped us identify biases, missings and non-relevant information. The results of the evaluation are presented and commented on within this document.

Scoring

Usually, providers of remediation services classify the ebooks per complexity: a book with more images, tables or pages will get a higher score. This method is relevant if the whole set of ebooks to classify is produced from a known production workflow. Looking at the European level, we know that publishers' workflows differ in the quality of files they produce, which consequently may be totally different in terms of accessibility features, accessibility information and, therefore, remediation needs.

That is why in this project we established a new classification related to remediation complexity, considering that an ebook may be very complex but already produced in accordance with accepted accessibility standards, thus resulting in a very low remediation complexity score. To be sure that the scoring was truly reflecting the remediation needs, we referred to our target file known to be fully accessible and with no remediation needs. With some iterations on scoring, we made sure that the target got a score of zero.

Capturing remediation complexities in relation to different file formats was one of the main challenges of the process. PDFs and Fixed-layout EPUBs are known to be the most complex to remediate as the technologies and languages used to build them imply more complexities and a higher level of programmatic abstraction. That's why we decided to represent them apart.

One bias we had to deal with is that PDF format allows for less structure and metadata, resulting in less possibilities for analyses, which resulted in abnormally low scores for files in this format. To address this bias, we had to establish a complementary scoring calculation to apply to these files.

Therefore, each format has specificities related to contents found in the files and accessibility related features missing. To find the correct marker, a threshold of calculated key indicators has been established thru iterations.

[31] Remediation complexity indicators are available for the publishers partners of the project.

Identified limits and bias

As previously commented, files in PDF format do not have the same accessibility possibilities as files in the EPUB format. Therefore, the comparison between the two formats must be done very consciously and should not lead to categorical formulas.

Most of the publishers providing samples are de facto aware of the accessibility subject and therefore the collection we have might be a biased representation of the backlist. A way to verify that would be to do a similar analysis on a large number of files not specifically selected for this type of test. This analysis perspective has been discussed with three members of EDRLab (Beletrina, De Marque and Hachette Livres) and we hope to be able to provide it as a complementary ABE Lab publication in the future.

At the time of writing this report, some remediation needs can not be spotted automatically, but as technological improvements are occurring very fast, we expect that a better gap analysis could be produced in the coming years. Examples of accessibility problems that cannot be automatically detected are incorrect, non-meaningful or insufficient image descriptions and wrong metadata claims, for which we were not able to establish a valid calculation method during this work.

Results

Per format

The sample contains 84% (316 files) of reflowable EPUB (RFL); 9% (33 files) of pre-paginated EPUB3 Fixed Layout (FXL) and 7% (26 files) of PDFs. This, actually, does not properly represent any of the market segmentations observed in the backlist data analysis.

The low number of pre-paginated files in the sample limits the analysis pertinence. It may be interpreted as an interest of the publishers providing samples to have accurate analysis on the remediation needs of reflowable EPUB files rather than PDF and EPUB3 FXL files, as many ebooks coexist in both reflowable and pre-paginated formats.

The radar diagram and the data table in the next page show the results of the scoring. We resume here the main trendings per format:

* PDF scoring ranges from 29 to 68 with representation in Thema categories A (The Arts), J (Society and Social Sciences), K (Economics, Finance, Business and Management), L (Law,), M (Medicine and Nursing), P (Mathematics and Science) T (Technology, Engineering, Agriculture, Industrial processes) and V (Health, Relationships and Personal development).
* EPUB3 Fixed Layout (FXL) average scoring ranges from 24 to 64 with representation in Thema categories A (The Arts,), C (Language and Linguistics), D (Biography, Literature and Literary studies), P (Mathematics and Science), S (Sports and Active outdoor recreation), T (Technology, Engineering, Agriculture, Industrial processes), W (Lifestyle, Hobbies and Leisure), X (Graphic novels, Comic books, Manga, Cartoons) and Y (Children's, Teenage and Educational) ;
* EPUB3 reflowable average scoring ranges from 4 to 77 with representation in all Thema categories except X (Graphic novels, Comic books, Manga, Cartoons).

This overview shows a concrete difference in ranges, where reflowable formats are almost all below a score of 50 and pre-paginated formats are all over 50. As commented before, the lack of information provided in PDF files might lead to minoring the remediation complexity. We tried to compensate for that in our scoring threshold, but remediation testing will have to establish if the compensation is enough or misleading.

We also detected that pre-paginated are not represented in every Thema code, while reflowables are missing only for category X: *Graphic novels, Comic books, Manga, and Cartoons*. This shows that, except for visual narratives, all types of books can be produced in a reflowable format.

From these results, it seems legit to treat remediation of pre-paginated files apart from the reflowable ones. This result will be represented in our remediation classification through the establishment of a first level of complexity related to file format.

Figure 6: Radar chart showing three curves for PDF (blue continuous line), EPUB Fixed-Layout (orange dashed line) and EPUB reflowable (violet dotted line)

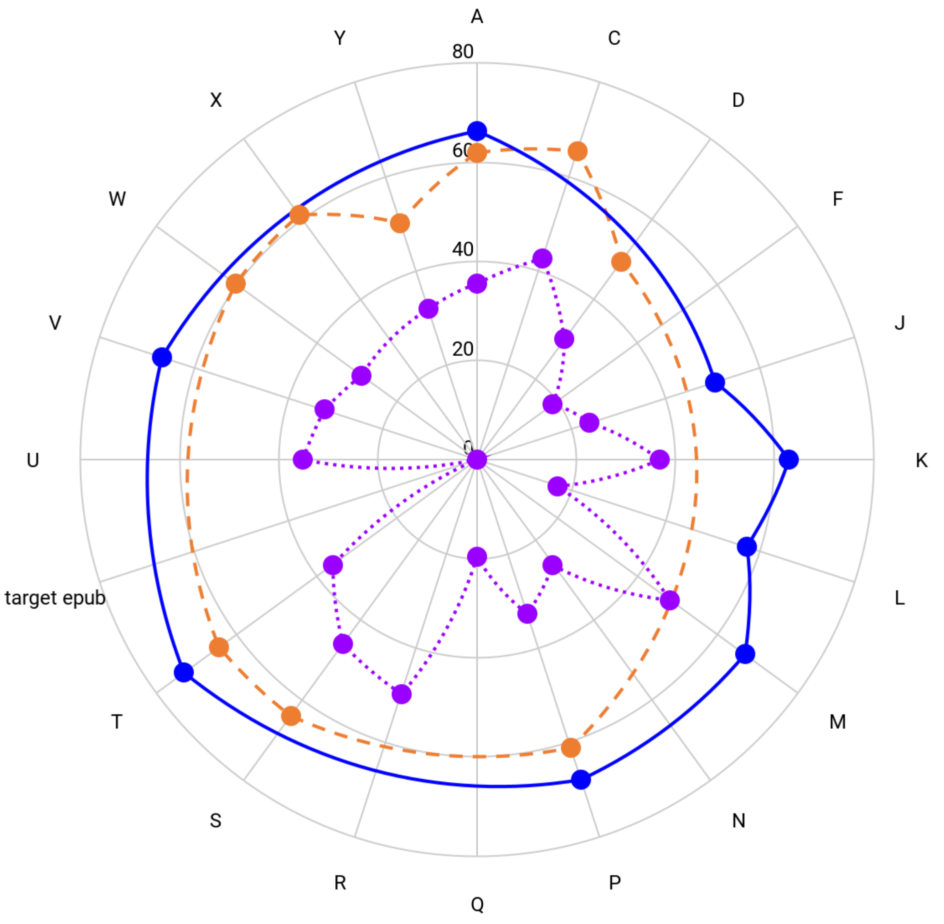


Table 4: Average score per format (rows) and Thema codes (columns). "-" represents absence of files in the sample collection for given format and Thema category

	A	*C*	*D*	*F*	*J*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*
PDF	*66*	-	-	-	*50*	*63*	*57*	*67*	-	*68*	-	-	-	*73*	-	*67*	-	-	-
RFL	*35*	*43*	*30*	*19*	*24*	*37*	*17*	*48*	*26*	*33*	*20*	*50*	*46*	*36*	*35*	*32*	*29*	-	*32*
FXL	*62*	*66*	*49*	-	-	-	-	-	-	*61*	-	-	*64*	*65*	-	-	*60*	*61*	*50*

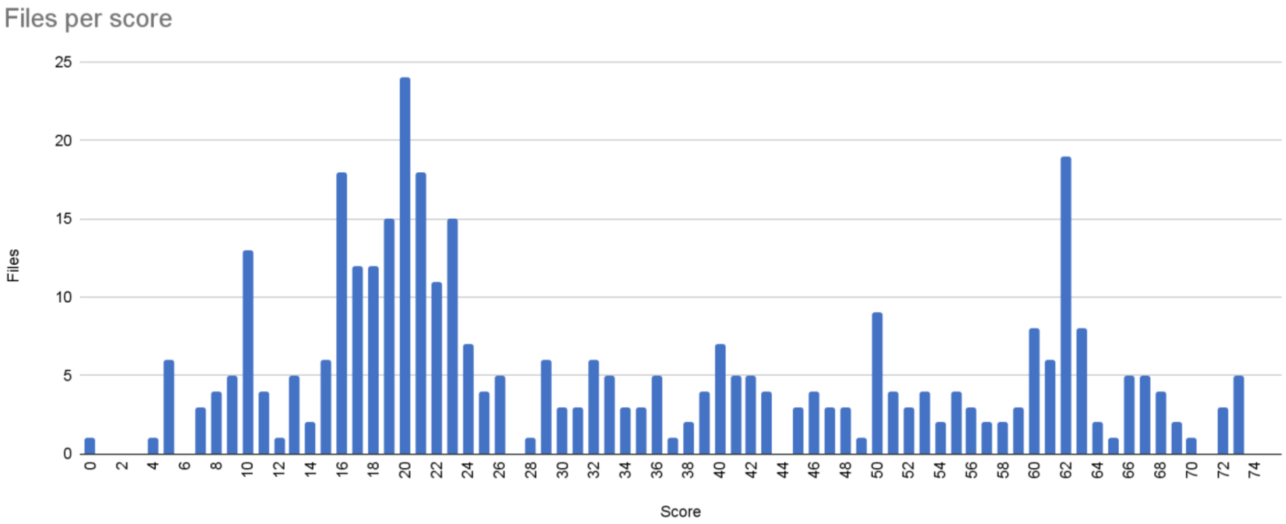
Focus on reflowable EPUB3

As reflowable EPUB3 is the format allowing full compliance to the EAA requirements, we judged it essential to dive deeper in the analysis of the remediation complexity of files in this format. In the collected samples files we found scores from 4 to 73 points. The vast majority have a score between 10 and 30.

The following charts and tables give a full representation. We will summarise here the key information we found:

- most of the files have a medium remediation complexity, but there is also a good number of files with high scores (fig. 4);
- images to fix (meaning textual alternatives to establish) are the heaviest error affecting strongly all categories except for L (Laws) and F (Fiction) (fig. 5);
- most of the categories have a large amplitude of errors per file, meaning that the Thema category alone is not sufficient to establish a segmented average remediation cost (fig. 6).

Figure 7: Bar chart showing the number of files as ordinate by scoring as abscissa.



List data: number of files per score

- Score 0: 1 files
- Score 1: 0 files
- Score 2: 0 files
- Score 3: 0 files
- Score 4: 1 files
- Score 5: 6 files
- Score 6: 0 files
- Score 7: 3 files
- Score 8: 4 files
- Score 9: 5 files
- Score 10: 13 files
- Score 11: 4 files
- Score 12: 1 files
- Score 13: 5 files
- Score 14: 2 files
- Score 15: 6 files
- Score 16: 18 files
- Score 17: 12 files
- Score 18: 12 files
- Score 19: 15 files
- Score 20: 24 files
- Score 21: 18 files
- Score 22: 11 files
- Score 23: 15 files
- Score 24: 7 files
- Score 25: 4 files
- Score 26: 5 files
- Score 27: 0 files
- Score 28: 1 files
- Score 29: 6 files
- Score 30: 3 files
- Score 31: 3 files
- Score 32: 6 files
- Score 33: 5 files
- Score 34: 3 files
- Score 35: 3 files
- Score 36: 5 files
- Score 37: 1 files
- Score 38: 2 files
- Score 39: 4 files
- Score 40: 7 files
- Score 41: 5 files
- Score 42: 5 files
- Score 43: 4 files
- Score 44: 0 files
- Score 45: 3 files
- Score 46: 4 files
- Score 47: 3 files
- Score 48: 3 files
- Score 49: 1 files
- Score 50: 9 files
- Score 51: 4 files
- Score 52: 3 files
- Score 53: 4 files
- Score 54: 2 files
- Score 55: 4 files
- Score 56: 3 files
- Score 57: 2 files
- Score 58: 2 files
- Score 59: 3 files
- Score 60: 8 files
- Score 61: 6 files
- Score 62: 19 files
- Score 63: 8 files
- Score 64: 2 files
- Score 65: 1 files
- Score 66: 5 files
- Score 67: 5 files
- Score 68: 4 files
- Score 69: 2 files
- Score 70: 1 files
- Score 71: 0 files
- Score 72: 3 files
- Score 73: 5 files

Figure 8: Bar chart showing level and repartition of errors per Thema code categories

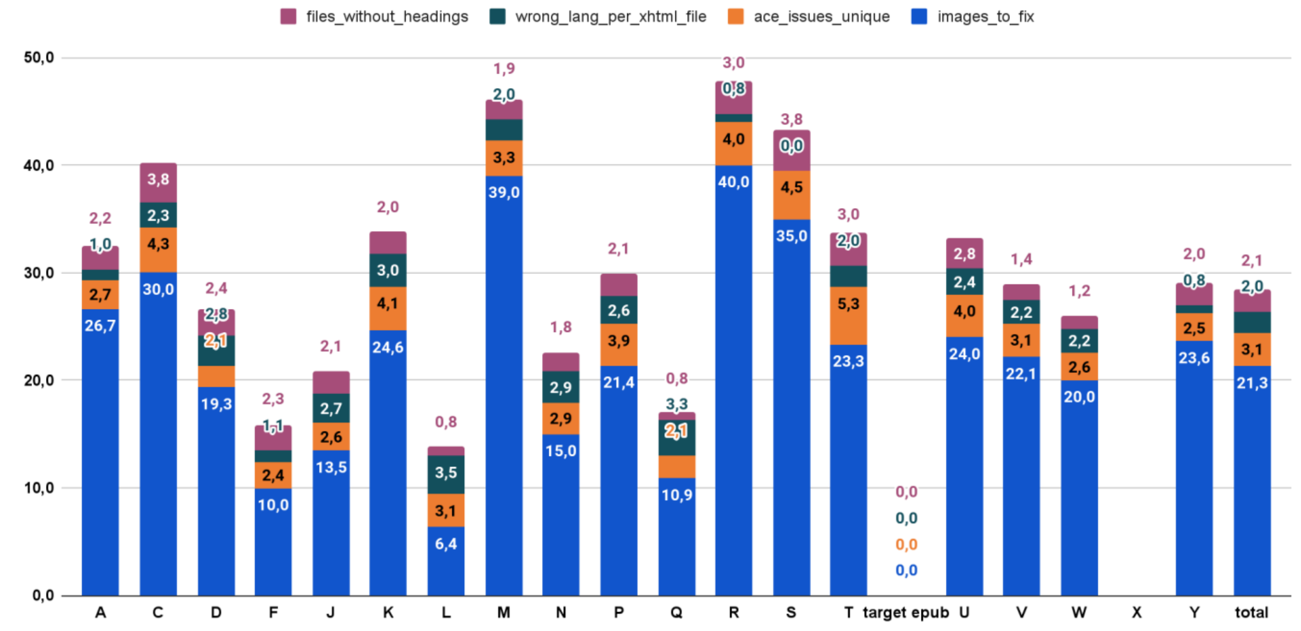


Table 5: level and repartition of errors per Thema code categories

Thema	Publications	images to fix	unique ACE issues	possibly wrong language	files without headings
A	6	26,7	2,7	1,0	2,2
C	4	30,0	4,3	2,3	3,8
D	30	19,3	2,1	2,8	2,4
F	68	10,0	2,4	1,1	2,3
J	23	13,5	2,6	2,7	2,1
K	26	24,6	4,1	3,0	2,0
L	11	6,4	3,1	3,5	0,8
M	20	39,0	3,3	2,0	1,9
N	16	15,0	2,9	2,9	1,8
P	22	21,4	3,9	2,6	2,1
Q	11	10,9	2,1	3,3	0,8
R	5	40,0	4,0	0,8	3,0
S	4	35,0	4,5	0,0	3,8
T	3	23,3	5,3	2,0	3,0
target	1	0,0	0,0	0,0	0,0
U	5	24,0	4,0	2,4	2,8
V	14	22,1	3,1	2,2	1,4
W	13	20,0	2,6	2,2	1,2
Y	33	23,6	2,5	0,8	2,0

Figure 9: Candlestick chart showing number of publications, minimum, average and maximum scores per Thema codes

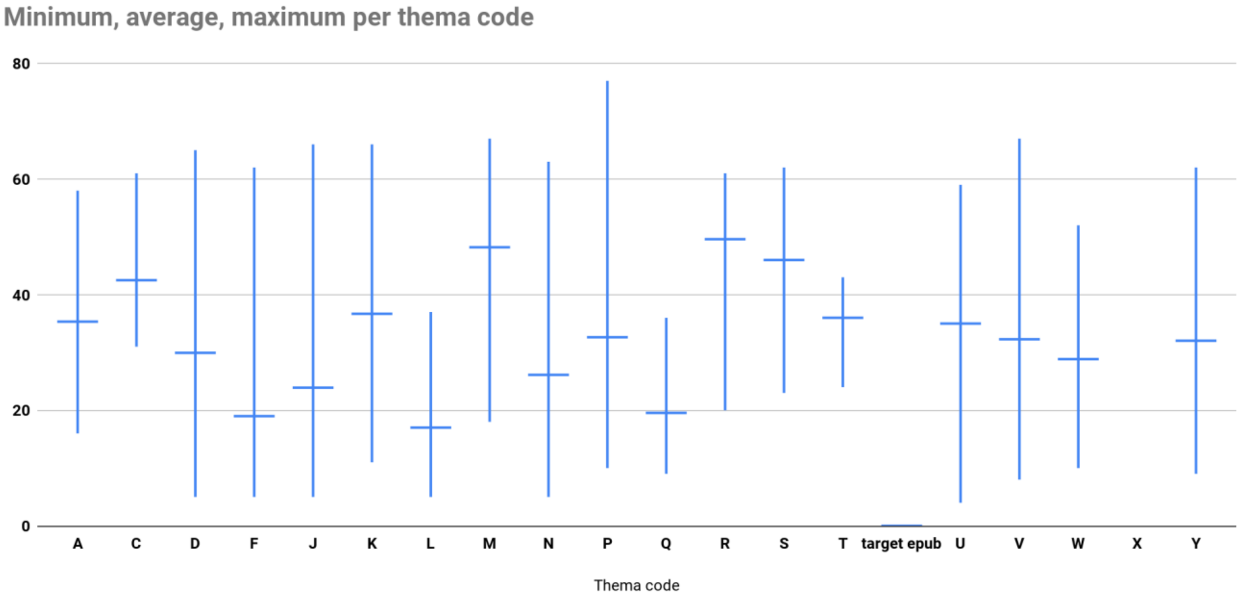


Table 6: number of publications, minimum, average and maximum scores per Thema codes.

Thema code	Publications	Average Score	Standard Deviation	Minimum Score	Maximum Score
A	6	35	17,87	16	58
C	4	43	14,11	31	61
D	30	30	17,10	5	65
F	68	19	7,89	5	62
J	23	24	15,05	5	66
K	26	37	20,54	11	66
L	11	17	8,98	5	37
M	20	48	13,10	18	67
N	16	26	16,76	5	63
P	22	33	17,32	10	77
Q	11	20	9,63	9	36
R	5	50	16,77	20	61
S	4	46	16,47	23	62
T	3	36	10,44	24	43
target	1	0	0,00	0	0
U	5	35	21,25	4	59
V	14	32	20,67	8	67
W	13	29	15,74	10	52
Y	33	32	13,98	9	62

Figure 10: Bubble chart showing average score (X axis) per standard variation (Y axis), one bubble per Thema code, bubble size represents the number of publications in the sample.

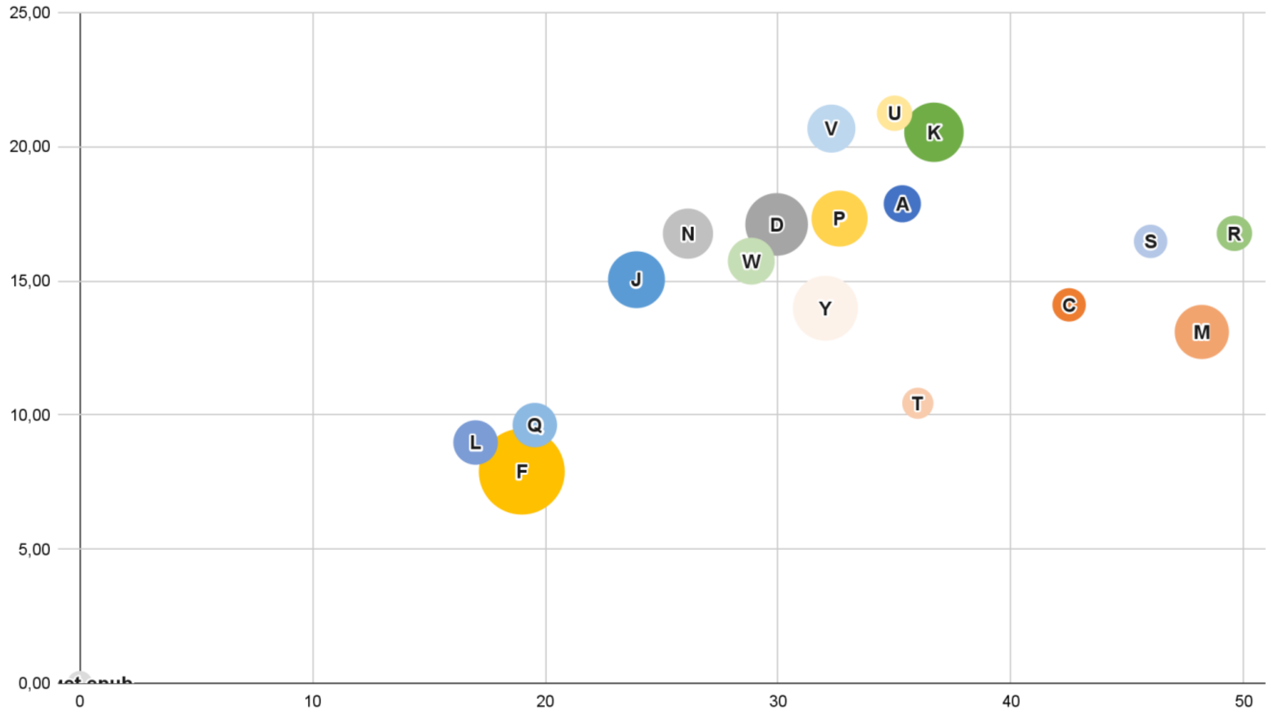


Table 7: Average score, standard variation and number of publications per Thema code.

Thema code	Publications	Average Score	Standard Deviation
A	6	35	17,87
C	4	43	14,11
D	30	30	17,10
F	68	19	7,89
J	23	24	15,05
K	26	37	20,54
L	11	17	8,98
M	20	48	13,10
N	16	26	16,76
P	22	33	17,32
Q	11	20	9,63
R	5	50	16,77
S	4	46	16,47
T	3	36	10,44
target	1	0	0,00
U	5	35	21,25
V	14	32	20,67
W	13	29	15,74
Y	33	32	13,98

Recurrent accessibility issues detected

As a complement to the Thema category level gap analysis, we listed the main known accessibility issues and tried to identify occurrences of these accessibility issues in the collected files. The following table resumes our findings. Results on each accessibility issue are detailed in the following sections.

Table 8: occurrences of main accessibility issues identified in collected files

Accessibility issue	concern	Number of files	in % of the sample
Missing Accessibility Metadata	EPUB files	343	100
Non reflowable content	all formats	59	16
Missing or bad textual alternative for non decorative graphical resources	all formats	312	83
Missing or bad Language Tag	EPUB files	227	66
ACE Issues	EPUB files	319	93

Missing Accessibility Metadata

- **Issue:** no accessibility metadata are present
- **Rule:** EPUBaccessibility 1.1 section '2. Discoverability'
- **Applies to:** EPUB files
- **Problem:** the reader cannot know features or limitations they may experience while reading and the publication can't be discovered through filtering
- **Indicators:** calculated as follows: missing metadata - inferred metadata[32], -3 (conformance metadata are counted as missing per ACE, but are not requested by the EAA). , minimum = 0
- **Collected files affected:** 100%

Non Reflowable content

- **Issue:** the presentation of the content can't be adjusted to fit the reader's needs
- **Rule:** EAA, Annex I, Section IV, f
- **Applies to:** all formats
- **Problem:** fixed displays impeach correct visual adaptation of the content
- **Indicators:** pre-paginated formats
- **Collected files affected:** 16%

[32] Inferred metadata are found per RGTK, meaning that we are able to see an accessibility feature in the file even if the information about it was not provided by the publisher. Therefore no remediation need is necessary except for informing about it, which is already automated per RGTK.

Missing or bad textual alternative for non decorative graphical resources

- **Issue:** No textual alternative is provided for informative graphical contents or the alternative is recognized as not meaningful (file name or one word)
- **Rule:** WCAG, Guideline 1.1 Text Alternatives, Success Criterion 1.1.1 Non-text Content, level A
- **Applies to:** all formats
- **Problem:** the non visual readers using TTS or assistive technologies will lose important information necessary to understand the content
- **Indicators:** calculated as follows: content images – content images with alt-text (more than one word and not equal to filename) – contents images decorative
- **Collected files affected:** 83%

Missing or bad Language Tag

- **Issue:** words in different languages from the one of the main content are not identified as such
- **Rule:** WCAG, Guideline 3.1 Readable, Success Criterion 3.1.2 Language of Parts, level AA
- **Applies to:** all formats, but no way was found to identify that in PDF
- **Problem:** non-visual readers using TTS or assistive technologies will experience strange or not understandable reading because of mispronunciation, incorrect braille rendering and bad hyphenations
- **Indicators:** the wrong language assertion is done through a dedicated algorithm. It targets two or more following words in a sentence
- **Collected files affected:** 66%

ACE issues

Issues reported by ACE. The following table shows the number of files and the corresponding percentage of the samples containing errors per severity level. We can note that very few (5% only) files have critical issues, but 92% have serious issues which will need to be evaluated for remediation.

Table 9: number and percentage of collected files affected per ACE issues gravity level.

ACE issue	Number of files	% of the samples
critical	19	5
serious	343	92
moderate	132	35
minor	172	46

A larger table of unique ACE issues has been produced for the use of the project and the building of testing files. The details of those errors are reported in the following tables. One shows the errors for which we proposed a detailed remediation complexity KPI, while the second shows the errors that are not addressed by a specific calculation.

Table 10: percentage of the sample affected by ACE errors for which a detailed remediation complexity KPI has been established.

ACE Issue	% of the samples affected
Epub-Lang:Serious	82.83
Metadata-Accessmode:Serious	46.81
Metadata-Accessmodesufficient:Moderate	71.75
Metadata-Accessibilityfeature:Serious	49.03
Metadata-Accessibilityhazard:Serious	52.91
Metadata-Accessibilitysummary:Moderate	71.47
Image-Alt:Critical	23.82

Table 11: percentage of the sample affected by ACE errors for which no detailed remediation complexity KPI has been established.

ACE Issue	% of the samples affected
Empty-Table-Header:Minor	45.19
Empty-Heading:Minor	69.25
Heading-Order:Moderate	47.33
Html-Has-Lang:Serious	51.07
Link-In-Text-Block:Serious	68.98
Color-Contrast:Serious	27.01
Metadata-Accessibilityhazard-Invalid:Moderate	28.88
Aria-Allowed-Role:Minor	29.41
Aria-Roles:Minor	23.80
Epub-Pagelist-Broken:Serious	5.08
Epub-Type-Has-Matching-Role:Minor	8.02
Landmark-Unique:Moderate	22.99
Epub-Toc-Order:Serious	27.54
Link-Name:Serious	26.20
Document-Title:Serious	15.24
Metadata-Accessibilityfeature-Invalid:Minor	12.57

Potential accessibility issues undetectable through automated analysis

The following are issues that cannot be detected automatically and will require ad hoc human testing.

Reflowable restrictions

- **Issue:** adjusting the presentation leads to letters or sentences overlapping or making the content visually unreadable in any way
- **Rule:** EAA, Annex I, Section IV, f
- **Applies to:** reflowable EPUB
- **Problem:** fixed styles impeach correct visual adaptation of the content

Specific contents to be verified manually if found in files

Some very specialised contents such as forms, scripts, maths, videos and audios are not usually used in ebooks, but as this may happen, it will be necessary to include them in remediation testing. The following table shows that very few occurrences were found in the sample collection.

Table 12: number and percentage of collected files per specific content.

	Number of files	in % of the samples
Forms	1	0.3
Scripts	17	5
Maths contents	7	2
Video contents	0	0
Audio contents	7	2

Classification for remediation

The following classification aims to list the remediation workflows to test. A list of six elements is spread across the different categories, here is a summary of it:

1. PDF to PDF/UA (compliant to WCAG 2.1, level AA)
2. PDF to Reflowable EPUB3 (compliant EPUB Accessibility 1.1, WCAG 2.1, level AA)
3. FXL to «accessible» FXL (compliant EPUB Accessibility 1.1, WCAG 2.1, level AA)
4. FXL to Reflowable EPUB3 (compliant EPUB Accessibility 1.1, WCAG 2.1, level AA)
5. EPUB2 to EPUB3 (compliant EPUB Accessibility 1.1, WCAG 2.1, level AA)
6. Reflowable EPUB to Reflowable EPUB3 (compliant EPUB Accessibility 1.1, WCAG 2.1, level AA)

PDFs

As a representation of the printed page, the PDF format accessibility features are limited in term of flexibility and choice in the presentation of the content (For details about the format and it's known limitations, refer to the Annex Ebooks files formats).

We see two possible remediation options for the PDF files:

1. improve the file to reach PDF/UA standard with WCAG 2.1 AA conformance, allowing the file to support the text zoom functionality provided by most of the reading applications. These files will not totally comply with the EAA's requisites, but will provide state-of-the-art compliance.
2. convert the file to a reflowable EPUB to reach full compliance with EAA requirements.

Fixed Layout EPUBs

Fixed layout EPUBs are subject to the same visual adjustment limitations as PDF: changing font type and spaces between letters, words, lines, or paragraphs is not possible. The possible remediations are:

3. improve the file to reach WCAG 2.1 AA and EPUB accessibility 1.1. As in the case of PDF files, in Fixed Layout EPUB some accessibility features are supported and others are not. If the file is made according to the specifications, however, it must support text zoom functionality provided by most of the reading applications.
4. convert the file to a reflowable EPUB to reach full compliance with EAA requirements.

Reflowable EPUBs

Reflowable EPUBs are known to be fully compliant with EAA requirements[33] if they conform to WCAG 2.1 AA (or superior) and EPUB accessibility 1.1. e found different types of remediation needs:

5. EPUB2 files need to be converted to reflowable EPUB3;
6. Reflowable EPUB3 files need to become compliant with WCAG 2.1 AA and the EPUB accessibility 1.1 .

Outcomes

From this gap analysis, we were able to establish a classification of remediation needs and build test files for each of the classifications.

Direct outcomes of this work are

- a remediation complexity assessment methodology applicable to collections of files;
- a view of the remediation complexity per Thema category;
- a view of main accessibility issues detected.

The heavy presence of images and visual resources appears to be the main criteria of demarcation between categories that will reclaim more efforts to remediate (Medicine, Earth sciences and Sports) and others (Fiction, Philosophy, Religion and Law) that will be easier to remediate.

As per the following steps of the ABE Lab project, it allows us to establish a testing classification and methodology as well as building meaningful files to test for remediation tools.

[33] See EPUB Accessibility - EU Accessibility Act Mapping Group note established by W3C.

ANNEXES

Data sources

1. Dilicom
2. Dilve
3. MVB GmbH
4. Informazioni Editoriali
5. Bibliotheka Narodowa
6. ISBN.NL
7. CB Logistics
8. Czech National ISBN agency
9. National Library of Sweden
10. DBC Digital
11. Publizon
12. Greek ISBN agency
13. Romania National Library
14. APEL
15. National Széchényi Library
16. ISBN Bulgaria
17. Nielsen Bookdata
18. Lithuania National Library
19. Slovenian ISBN agency
20. National Library of Latvia
21. Estonian ISBN agency
22. Malta National Book Council

ABELab sample collection wishlist

Wish list **ebooks** to collect ABE lab	before 2011	2011-2018	after 2018	2022
Fiction (novel, thriller, …)	3	3	3	3
Children's books (simple)	1	1	1	1
Children's books (picture book or FXL)	1	1	1	1
Graphic novel	1	1	1	1
D. Biography, Literature & Literature studies	1	1	1	1
J. Society & Societal Sciences	1	1	1	1
V. Health, Relationships & Personal development	1	1	1	1
N. History & Archaeology	1	1	1	1
Q. Philosophy & Religion	1	1	1	1
W. Lifestyle, Hobbies & Leisure…	1	1	1	1
K. Economics, Finance, Business & Management	1	1	1	1
L. Law	1	1	1	1
A. The Arts..	1	1	1	1
P. Mathematics & Science	1	1	1	1
M. Medicine & Nursing	1	1	1	1
S. Sports & Active outdoor recreation	1	1	1	1
R. Earth Sciences, Geography, Environment, Planning	1	1	1	1
U. Computing & Information Technology	1	1	1	1
C. Language & Linguistics	1	1	1	1
T. Technology, Engineering, Agriculture, Industrial…	1	1	1	1

Publishers partners per country

- Danemark: 1
- Finland: 1
- France: 21
- Germany: 1
- Italy: 21
- The Netherlands: 25
- Spain: 2
- United Kingdom: 3

Composition of the collection of ebook samples

- Per Thema codes:
  - A: 3,21%
  - C: 1,60%
  - D: 8,82%
  - F: 18,18%
  - J: 6,95%
  - K: 7,49%
  - L: 3,74%
  - M: 7,75%
  - N: 4,28%
  - P: 6,68%
  - Q: 2,94%
  - R: 1,34%
  - S: 1,34%
  - T: 1,60%
  - U: 1,34%
  - V: 4,55%
  - W: 4,55%
  - X: 0,53%
  - Y: 12,83%
- Per year:
  - 2023: 15,50%
  - 2022: 26,02%
  - 2021: 7,31%
  - 2020: 5,26%
  - 2019: 4,97%
  - 2018: 9,36%
  - 2017: 3,22%
  - 2016: 3,51%
  - 2015: 1,75%
  - 2014: 2,63%
  - 2013: 3,22%
  - 2012: 1,46%
  - 2011: 8,19%
  - 2010: 5,26%
  - Before 2010: 2,34%

Tools used in the automated analysis

EPUBcheck

EPUBCheck is the W3C official conformance checker for EPUB publications. The DAISY Consortium maintains the project. It is available as open source, under MIT licence.

EPUBcheck is available at: https://www.w3.org/publishing/epubcheck/

ACE

ACE, the Accessibility Checker for EPUB, is a tool developed by the DAISY Consortium to assist with the evaluation of accessibility features of EPUB publications. It is available under MIT licence.

ACE is available at https://daisy.github.io/ace/

Readium Go-toolkit

Developed and maintained by the Readium Foundation, the Go-Toolkit is part of a set of robust, performant, spec-compliant reading system toolkits that support digital publishing formats (e.g. EPUB, Web Publications etc.) and can be deployed in browsers or built into native apps on iOS, Android or the desktop. It is available under BSD-3-Clause licence.

Readium Go-toolkit is available at https://github.com/readium/go-toolkit

Vera PDF

VeraPDF is a purpose-built, open-source, file-format validator covering all PDF/A and PDF/UA parts and conformance levels. It is developed and maintained by the VeraPDF consortium and it is available under dual-licensed GNU General Public License v3 or later (GPLv3+) and Mozilla Public License v2 or later (MPLv2+).

VeraPDF is available at https://verapdf.org/

PDFix SDK

PDFix SDK provides a platform to work with PDF files. Developed and maintained by pdfix-inc, it is provided under a commercial licence. We use the Lite version provided under the Free licence. It allows us to get statistics information on PDFs, including the information on the presence or absence of tags.

PDFix is available at https://pdfix.github.io/

Glossary

Abbreviations

- EAA: European Accessibility Act
- ACE: Accessibility Checker for EPUB
- EU: European Union
- EPUB: Electronic Publication
- FXL: Fixed Layout
- FEP: Federation of European Publishers
- KPI: Key Point Indicators
- RFL: Reflowable
- RGTK: Readium Go Tool Kit
- PDF: Portable Document Format
- WCAG: Web Content Accessibility Guidelines

Thema codification

- A: The Arts
- C: Language and Linguistics
- D: Biography, Literature and Literary studies
- F: Fiction and Related items
- G: Reference, Information and Interdisciplinary subjects
- J: Society and Social Sciences
- K: Economics, Finance, Business and Management
- L: Law
- M: Medicine and Nursing
- N: History and Archaeology
- P: Mathematics and Science
- Q: Philosophy and Religion
- R: Earth Sciences, Geography, Environment, Planning
- S: Sports and Active outdoor recreation
- T: Technology, Engineering, Agriculture, Industrial processes
- U: Computing and Information Technology
- V: Health, Relationships and Personal development
- W: Lifestyle, Hobbies and Leisure
- X: Graphic novels, Comic books, Manga, Cartoons
- Y: Children’s, Teenage and Educational

Ebooks files formats

- EPUB (Electronic PUBlication) is an open file format for electronic publications based on Web Standards (HTML, CSS ,JavaScript). The first version, named OEBPS 1.0 (Open EBook Publication Structure) was approved in 1999 by the Open eBook Forum, which later became the International Digital Publishing Forum[34] (IDPF). EPUB 2.0 was released in 2010, followed in June 2014 by version 3, in which for the first time in the specifications were included accessibility features. Just after the release of EPUB 3.1 in January 2017, the IDPF merged into the World Wide Web Consortium (W3C), making the EPUB an official W3C recommendation (standard). The last version, EPUB 3.3, was published on May 25, 2023[35]. EPUB is a native semantic format allowing to use numerous taggings from different standardised languages such as HTML, ARIA, MATHML, SVG and others.
  Accessibility features of the format are detailed in EPUB Accessibility 1.1 W3C recommendation specifying content conformance and accessibility metadata requirements for the EPUB publications.
- PDF (Portable Document Format), created by Adobe in 1992 and standardised as an open standard, maintained by the International Organization for Standardization (ISO). The PDF ISO declines in special purposes such as PDF/A for archiving, PDF/E for engineering, and PDF/X for printing, PDF/UA for accessibility. Initially purposed for print reliability and digital conservation, the format is based on PostScript, a computer language for describing the positioning of characters and graphic elements absolutely on the page (similar to having x and y coordinates to position each element on the page). As it went to be used for digital consultation, the format evolved to respond as best as possible to this use principally with the addition of a semantic descriptive layer composed of XML[36] language tags. The variety of tagging that can be added is currently limited to 28 elements[37].
  The PDF/UA ISO[38] standard provides definitive terms and requirements for accessibility in PDF documents and applications. It supports accessibility features like the reading order, semantic tagging, and incorporation of non-structured alternative text for graphic resources. Intrinsic accessibility known limitations are:
  - lack of possibilities to adjust the font type and spaces between letters, words, lines, or paragraphs, meaning that the format may provide a barrier-free reading experience for screen readers users but would provide a difficult reading experience for users with low vision or cognitive-related disabilities;
  - lack of possibilities to provide information on accessibility features within the file.

[34] https://idpf.org/

[35] What is an EPUB file? Available at https://www.edrlab.org/open-standards/epub/

[36] Extensible Markup Language (XML) is a markup language that provides rules to define any data. It is standardised by the W3C and can be found at https://www.w3.org/TR/xml/

[37] A list of Standard PDF Tags is available at https://helpx.adobe.com/acrobat/using/editing-document-structure-content-tags.html

[38] Available at https://www.iso.org/standard/64599.html

Bibliography

- *Thema – the subject category scheme for a global book trade version 1.5*, EDItEUR, 2022. Available at https://ns.editeur.org/thema/en
- *About Google Books – Free books in Google Books*. Available at https://www.google.com/intl/en/googlebooks/about/free_books.html [Consulted on May 2023]
- *How the Google Books team moved 90,000 books across a continent*. Ari Mariani, 2023. Available at https://blog.google/products/search/google-books-library-project/
- *How Many Ebooks Are There In The Kindle Store On Amazon?* Just Publishing Advice, 2023. Available at: https://justpublishingadvice.com/how-many-kindle-ebooks-are-there/
- European Book Market Statistics 2021-2022. FEP, 2022. Available at https://fep-fee.eu/-Publications-
- *European Book Publishing Statistics 2021*, FEP 2022. Available at https://fep-fee.eu/European-Book-Publishing-1467
- *Thema mappings*, EDItEUR, 2023. Available at https://www.editeur.org/151/Thema/#Mappings
- *Thema: the subject category scheme for a global book trade. Executive briefing*. EDItEUR, 2022. Available at: https://www.editeur.org/files/Thema/20220422_Thema%20Executive%20briefing.pdf

Index of tables

Index of figures

Addendum - Per year analysis

As part of the elaborations to define the gap analysis between the accessibility requirements of the EAA and the accessibility issues of the ebooks in the backlist, we also checked the breakdown by year to see if we were able to identify meaningful patterns. Due to the context of our research, this part was not considered so relevant and therefore was not published in the initial version of this report. For clarification and completion of the information related to the study, we have decided to add this information as an Addendum by May 2024.

In the context of the ABE Lab research, the per-year analysis resulted in being fewly relevant because all Thema categories of books are present for most of the years, leading to averages scores per year variation from 20 to 37. This does not indicate that a per-year analysis is not relevant and we suggest that it should be done on defined collections. We expect that a refined per-year analysis crossed with Thema categories done on higher numbers of samples would help to identify errors patterns. This was confirmed by publishers who already started the remediation of their backlists.

This analysis showed that our sample is composed of a third of files produced after 2021, which will probably be the case for most of the ebooks that will be considered to be made accessible to comply with the EAA. It showed higher levels of complexity around the years 2017 to 2020, mainly because the samples collected for those years have a larger number of images.

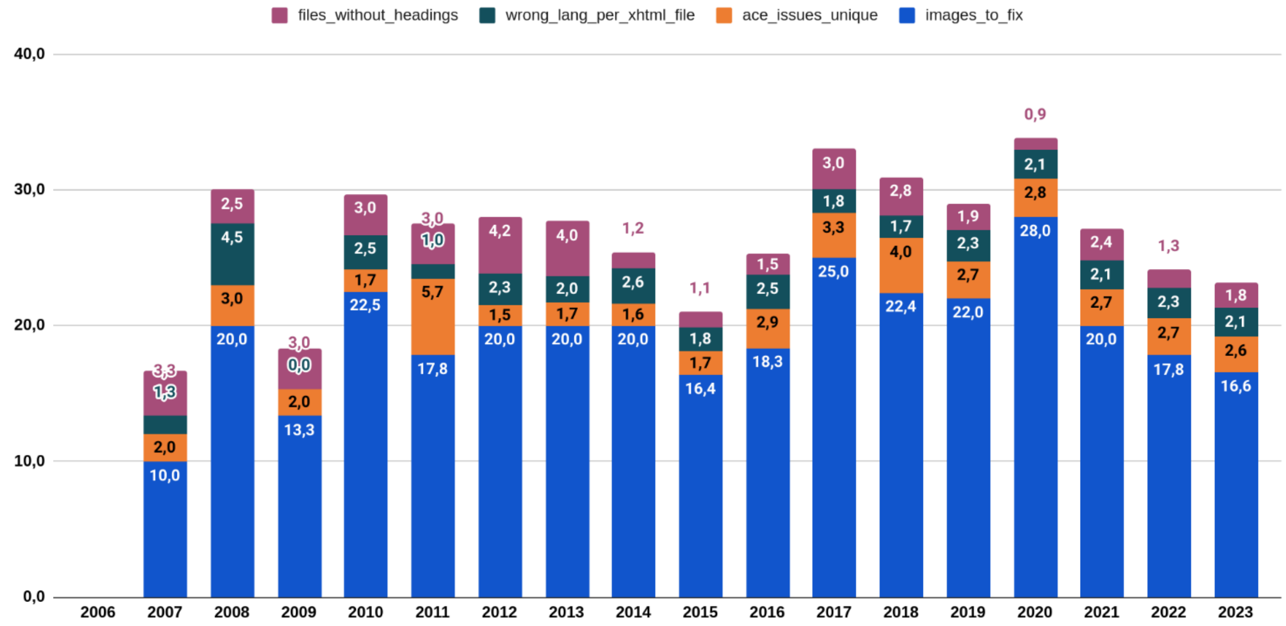
Table A1: reflowable EPUBs, per year, number of publications and scoring.

Year	Publications	Average Score	Standard Deviation	Minimum Score	Maximum Score
2007	3	20	4,58	16	25
2008	2	34	0,71	33	34
2009	3	22	4,04	18	26
2010	12	34	19,69	8	61
2011	23	30	11,54	20	56
2012	6	32	16,24	20	62
2013	3	31	16,20	21	50
2014	5	28	15,87	16	55
2015	11	25	13,74	14	55
2016	12	29	17,54	9	67
2017	8	37	17,50	20	63
2018	33	33	15,48	16	66
2019	15	32	16,15	10	62
2020	10	36	16,07	10	55
2021	24	30	20,05	4	77
2022	104	27	17,90	5	67
2023	38	27	18,08	0	66

Table A2: level and repartition of errors per year.

Year	Publications	Average Score	Standard Deviation	Minimum Score	Maximum Score
2007	3	10,0	2,0	1,3	3,3
2008	2	20,0	3,0	4,5	2,5
2009	3	13,3	2,0	0,0	3,0
2010	12	22,5	1,7	2,5	3,0
2011	23	17,8	5,7	1,0	3,0
2012	6	20,0	1,5	2,3	4,2
2013	3	20,0	1,7	2,0	4,0
2014	5	20,0	1,6	2,6	1,2
2015	11	16,4	1,7	1,8	1,1
2016	12	18,3	2,9	2,5	1,5
2017	8	25,0	3,3	1,8	3,0
2018	33	22,4	4,0	1,7	2,8
2019	15	22,0	2,7	2,3	1,9
2020	10	28,0	2,8	2,1	0,9
2021	24	20,0	2,7	2,1	2,4
2022	104	17,8	2,7	2,3	1,3
2023	38	16,6	2,6	2,1	1,8

Figure A1: Bar chart showing level and repartition of errors per year (visual representation of the data presented in table A2).



Document Control Information

- Document Title: Report on backlist data and gap analysis
- Project Title: ABE Lab
- Project Manager (PM): EDRLab
- Author co-authored by the project partners
- Doc. Version: 1.0
- Sensitivity: Public — fully open
- Date: 30/10/2023

Document Approver(s) and Reviewer(s):

All Approvers are required. Records of each approver must be maintained.
All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Action	Date
EDRLab	PM & co author	*Approved*	30 oct. 2023 / 15 may 2024
Fondazione LIA	Project partner & co author	*Approved*	30 oct. 2023 / 15 may 2024
KB	Project partner & co author	*Approved*	30 oct. 2023 / 15 may 2024

Document history:

The Document Author is authorised to make the following types of changes to the document without requiring that the document be re-approved: *Editorial, formatting, and spelling Clarification*. Changes to this document are summarised in the following table in reverse chronological order (latest version first).

1. Initial publication on 30/10/2023 created by EDRLab, Fondazione LIA & KB.
2. Update on 15/05/2024 including an Addendum after the Annexes.

Statement of originality:

This deliverable 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.

Document Location

The latest version of this controlled document is stored at:
https://github.com/ABELaboratory/publications/deliverables/report-on-backlist-data-and-gap-analysis