Final public report

Accessible Backlist Ebooks Laboratory (ABE Lab)

TABLE OF CONTENTS

Abstract

The European Accessibility Act (EAA)[1] is a European Union (EU) Directive that *aims to improve the functioning of the internal market for accessible products and services by removing barriers created by divergent rules in Member States*[2]. The EAA will take effect in June 2025, requiring many products and services, including ebooks categorised as services, to comply with accessibility requirements.

One particularly important topic is how a publisher can evaluate the impact of converting the inaccessible ebooks of its backlist into versions compliant with the accessibility requirements of the EAA. Earlier studies have shown very different results. In particular, the issues related to different categories of ebooks have not been addressed. At the same time, some companies are offering tools for converting existing EPUB 2 and PDF files into accessible EPUB 3 versions, but it is not always so clear which issues these tools practically solve.

Given the limited timeframe and the fact that many parties have limited resources to investigate this topic, we have set up the Accessible Backlist Ebooks Laboratory (ABE Lab), a project funded by the EU Creative Europe program. The goal of this project is to centralise some of the necessary research and share the results with anyone interested.

To get insight into the possibilities and costs of remediation, the partners of the ABE Lab project, European Digital Reading Lab (EDRLab), Fondazione LIA and KB – National Library of the Netherlands, conducted research activities to provide information on:

- The size and composition of the ebook backlists at European and national levels,
- Recurrent elements that cause accessibility issues,
- how to categorise clusters of ebooks which may have similar remediation issues,
- tools available on the market that can help with the remediation of ebooks of different formats (EPUB, PDF) and different categories (fiction, non-fiction, STEM, children's book…),
- costs estimates of remediation[3].

This final project report is composed of 5 parts:

1. a short introduction presenting the context and the main issues addressed by the project, detailing the partners, objectives and expected impacts
2. a summary of the activities conducted and of the methodologies
3. a part dedicated to the main findings, detailing the most prominent challenges to consider
4. an outcomes section resuming the main deliverables of the project and
5. a final summary of the state of the art and of possibilities to consider for the next steps.

[1] Directive 2019/882 available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0882#d1e32-115-1
[2] https://ec.europa.eu/social/main.jsp?catId=1202
[3] Since the price of work is strongly different between European countries, we are not regarding costs in money but in minutes. Those estimate times are a point of reference but must be considered in relation to publisher's national realities and per each collection in depth analysis.

Context

According to a World Blind Union (WBU) estimation, persons with reading disabilities - for instance, blind, visually impaired or dyslexic people - have access to less than 10 per cent of the yearly book production[4], depending on the country they live in. The term "book famine" was coined some years ago to describe this. The books they have access to are, in most cases, specific adaptations of printed books created in small quantities by specialised libraries, associations for the blind, or similar organisations, or digital books whose level of accessibility can vary greatly.

The European Accessibility Act (Directive (EU) 2019/882, acronym EAA) will come into force in June 2025. The Directive mandates that various services, including ebooks, adhere to accessibility requirements, in line with the principles of *design for all* and *born accessible*. This means taking into account all the needs and possibilities from the conception of the services to be able to create quality services that can be used also by people with disabilities. It is part of the European strategy for the rights of persons with disabilities 2021-2030[5], which addresses several major themes: accessibility, equality, education, social protection, employment and more. With this, Europe is committed to an inclusive policy that makes people with disabilities a core objective.

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

However, there is still an issue with the majority of ebooks published before 2025: a backlist of 3.5 million titles in Europe. Since one of the objectives of the Directive is to end the book famine endured by visually impaired people, it is important to transform as many titles as possible into accessible EPUB files. To do so, publishers need to have a clear view of the gaps to be filled to make their ebooks compliant and of the related costs per type of ebook.

What is accessibility?

Accessibility means that tools, technologies, websites, ebooks and digital documents are designed and developed so that everyone, regardless of their disabilities - physical, cognitive, motor, situational or age-related - can use them. Accessibility is fundamental to providing equal access to information, study, entertainment and equal opportunities to people with disabilities.

In addition to being required by the regulatory frameworks in many countries worldwide, accessibility is a social value and a basic human right recognised by the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD)[6].

Accessibility is also part of a global quality process. It is good for business, improving the experience and satisfaction for all users. This means that attention has been paid to providing correct hierarchy and semantic tagging of the information, that rich and understandable navigation options are available, and that information about the available features is properly provided. By enforcing standards rules, accessibility also

[4] https://worldblindunion.org/programs/accessibility/accessibility-go-a-guide-to-action/
[5] https://eur-lex.europa.eu/EN/legal-content/summary/strategy-for-the-rights-of-persons-with-disabilities.html
[6] https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities

helps with translations, derivations, and preservation by guaranteeing a more strict homogeneity of the collections.

What are remediation tools?

In recent years, accessibility in the digital world has become increasingly relevant, especially in light of the new European and national regulatory frameworks.

This has led to improvements in ebook production tools, which have progressively integrated specific functionalities to support the production of accessible EPUB and PDF files, allowing publishers to publish and market ebooks that are already accessible, following the *born accessible* paradigm. Not only the tools but the formats themselves have evolved in terms of the accessibility features supported, and accessible guidelines for the EPUB and PDF formats have been published, providing publishers with clear reference specifications to produce accessible ebooks compliant with international and EAA accessibility requirements. As a consequence, production workflows have also started to take accessibility into account.

A new type of tool and service is increasingly emerging to fix the accessibility of ebooks and digital documents produced in past years. These tools, which may take the form of online cloud platforms, desktop applications, or SDKs, are designed to receive an ebook or digital document that does not comply with international accessibility requirements as input and return an ebook that should be accessible and compliant as output.

They come with different features and access methods: some can be freely downloaded and used, while others require a paid account activation and configuration, which in turn may require direct contact with the developers and marketing or sales managers. The types of licences may also differ by tool provider: they may provide unlimited access to the tool or include a maximum number of files uploaded and remediated; they can provide a monthly, semestrial, annual or other subscription formula, often different for private users and companies. Some are fully automatic, while others allow the manual modification of the document content.

Finally, tools based on Machine Learning and Artificial Intelligence are also beginning to appear and spread; these tools can potentially be trained to respond to the specific needs of a publisher's workflow or the characteristics of a specific ebook series.

As can be seen, the world of these tools is very composite and varied. In this document, we'll refer to these types of technological solutions as *remediation tools*.

The ABE Lab project

Objectives and expected impacts of the project

ABE Lab has two main objectives:

- to collect and analyse in detail the composition of the European ebook backlist;
- to provide guidelines to European publishers to boost the remediation of the ebooks from their backlist.

The main expected impacts in the short term (before 2025) are the following:

- Publishers, Publisher Associations, blind unions, and public services have, for the first time, a clear picture of the status of the European ebook backlist;
- Publishers can use the classification to evaluate the composition of their own ebook backlist;
- Some larger publishers and service providers will build remediation workflows;
- Ebooks that are easy to remediate will be processed through these technical workflows;
- Publishers can calculate the cost of remediation for each segment of their backlist and decide on priorities for remediation.

In the medium term (2025, 2026), the expected impacts are that:

- More publishers and service providers will build remediation workflows;
- The analysis of the first bulk of remediation will enhance the efficiency of the technical workflows. More automated processes and Artificial Intelligence techniques will be integrated into the remediation tools;
- Ebooks that necessitate limited manual processing will be processed through enhanced technical workflows;
- Publishers can use this analysis to evaluate the cost of generation of natively accessible EPUB 3 ebooks.

In the long term (2026, 2030), the expected impacts are that:

- Every European publisher has chosen which backlist ebooks will be made accessible, which titles will be considered disproportionately burdensome, and which titles will be removed from the market;
- Ebooks deemed to be made accessible will be processed through enhanced technical workflows;
- EPUB 3 eventually replaced older ebook formats (PDF, EPUB 2).

The partners

The Accessible Backlist Ebooks Laboratory (ABE Lab) is a European Cooperation project funded by Creative Europe. The project is coordinated by the European Digital Reading Lab (EDRLab[7]); Fondazione LIA[8] and the National Library of the Netherlands (Koninklijke Bibliotheek - KB[9]) are partners.

EDRLab is a non-profit transnational laboratory that develops open standards and open-source tools to accelerate the diffusion of digital technologies in the publishing industry. It is an association of more than 75 members from the publishing industry, including publishers, providers of publishing services, distributors, tool developers, and accessibility specialists. EDRLab is also the developer of the Thorium Reader application, a free and highly accessible reading application for EPUB 3 ebooks, and LCP, a non-proprietary DRM for ebooks.

Fondazione LIA is an Italian non-profit organisation and an internationally recognised expert in the certification of accessibility of EPUB 3 ebooks, having already certified more than 35.000 ebooks available in the *www.libriitalianiaccessibili.it* catalogue. Fondazione LIA carries out research and development activities in the field of digital accessibility, organises awareness-raising events, and offers training courses and consulting activities. It works with 76 publishing imprints to create and distribute new books as accessible publications, and it's part of an international network of organisations dealing with content accessibility.

KB - The National Library of the Netherlands has a long history in library research. It promotes the visibility, usability, and longevity of the Dutch Library Collection, which is defined as the collective holdings of all publicly funded libraries in the Netherlands. It has three main roles: it is the national depot where many Dutch publishers send their ebooks, it operates an ebook lending platform for all Dutch public libraries, and it commissions the conversion of non-accessible books to accessible formats for people with a print impairment. With the EAA on the horizon, the KB strives to make all its public services accessible.

Both EDRLab and LIA are members of the W3C Publishing Business Group, W3C Publishing Working Group, and W3C EPUB Community Group. They are active members of the EPUB Community Group Accessibility Task Force and the EPUB Community Group Fixed Layout Accessibility Task Force and actively shape standards and guidelines.

[7] edrlab.org
[8] fondazionelia.org
[9] kb.nl

Activities

The ABE Lab project aims to provide publishers with reliable information about options and costs for remediation to make backlist ebooks accessible. This project objective was achieved by collecting and analysing data about the European ebook backlist and analysing and testing the processes and workflows by which ebooks of different types and formats (PDF, EPUB reflowable and EPUB Fixed Layout) can be transformed into accessible EPUB in the easiest and less expensive way. This implies knowing which remediation tools are available on the market and what they can do, for which formats and types of books (fiction, complex layouts, STEM) they are most effective, and their shortcomings.

The first part of the project was therefore focused on collecting and analysing data to define the size and composition of the European ebook backlist—defined as the collection of all ebooks available on the EU market—and on identifying the most recurrent accessibility issues in these ebooks. The public deliverable Report on Backlist Data and Gap Analysis details the results.

In the second part, the partners tested several workflows using existing open-source and commercial tools to remediate inaccessible ebooks of different types and formats (PDF, EPUB 2, EPUB 3 reflowable, and Fixed Layout). The study considered ebooks with different levels of complexity to evaluate the relative effort associated with correcting the accessibility issues.

It is essential to have tools that can correct the accessibility errors identified in the backlist files or offer the human operators options and features that support them in their remediation activity. Consequently, testing existing remediation tools has been fundamental to ABE Lab. The results of these tests form the basis of the whitepaper *Guidelines for remediation tool producers* and the starting point of the guidance we can provide to publishers as an outcome of the project.

Project management (WP1)

The project was coordinated by EDRLab, operated by a core team, and strategically driven by a Steering Committee. The project partners allocated additional resources for each work package. All working documents were stored on a shared drive, and communication was mainly by email. A chat solution was also available for daily discussions.

The project core team met online regularly and performed all activities. The Steering Committee met two times in person and three times online. Two status reports were produced at 6 and 12 months. All decisions were made on consensus.

Analysis of the composition of ebook backlists in Europe (WP2)

This activity was composed of two tasks: collecting data on the European ebook backlist and analysing them, and collecting backlist ebooks from publishers from different EU countries. Details are given in the *Report on Backlist Data and gap analysis*[10].

Backlist Data

Without specific information about the composition of ebook backlists in Europe, it's challenging to determine the best workflows, remediation costs, and potential growth of accessible ebooks before and after the European Accessibility Act (EAA). This lack of data also makes it hard to measure the EAA's success in the ebook industry. To plan the next steps of the project, data was collected and analysed. This formed the basis for subsequent reports publicly published.

As there's no single European source providing data about the number and types of ebooks available, many organisations were contacted. These include publication registration offices, distributors, aggregators, resellers, and national libraries. However, since each organisation only has a portion of the data, and there are overlaps, conclusions must be drawn carefully. A blind spot is that a significant part of the EU market's ebook backlist comes from e-commerce platforms and international distributors and retailers, especially those from the UK, the USA, and Canada.

The report examines the size and composition of EU ebooks backlists and their segmentation and provides useful trends and insights for the research on the topic. It outlines the approach taken, including challenges faced, and gives an overview of the segmentation by subject, format, and year of production. The main findings are summarised in the *Insights* section, and the *Outcomes* section explains how this work was used to shape subsequent project activities.

Gap Analysis

The objective of this activity was to provide an overview of the gaps between the accessibility level of the ebook titles available on the EU market and the EAA target accessibility requirements. The report exposes an analysis methodology to help identify recurrent accessibility issues per category of ebooks that will need remediation to meet the EAA requirements.

This analysis was possible thanks to the collaboration with publishing houses that agreed to provide us with samples of their existing ebooks, matching a wishlist established to reflect the articulation of the backlists as well as possible. These publishers received a document describing in detail the methodology developed and applied to identify the most recurrent accessibility issues and classify the ebooks according to the potential complexity of the remediation activity needed to make them accessible. This document was released on the 30th of October 2023 and was shared with contributing publishers only.

[10] *Report on Backlist Data and gap analysis*, Released on 2023, October the 30th. Diffusion: Public. Available at https://www.abelab.eu/outcomes/deliverables/#report-on-backlist-data-and-gap-analysis

Analysis of remediation processes and workflows (WP3)

This activity was the basis of internal intermediary documents and a *Guidelines for remediation tools producers*[11] available on the ABE Lab website.

It is crucial to know which technological solutions for remediating inaccessible ebooks are available on the market and in the open-source community, what they can do, and with which level of accuracy. This information is key to providing publishers with reliable indications on which tools they can use to remediate their backlist ebooks. The choice may depend on ebook categories (fiction, biography, law, STEM, ...), input formats, desired output formats and specific accessibility issues of the input files.

Therefore, the objectives of this activity were to:

- identify the available tools;
- establish a testing methodology for assigning a score to the tools based on the results produced;
- provide an overview of the tools tested, divided by remediation flows. For each tested tool it includes general information about the product, its evaluation and the score it has obtained.

Discovering and testing the remediation tools available on the market or developed by the open-source community was the focus of the second phase of the project. This phase consisted of the following macro activities:

1. active search for remediation tools. The tools were found through desk research and by exploiting the professional network of the project partners;
2. mapping the functionalities of these tools using a matrix of functionalities;
3. expert people testing the tools and analysing the remediated files to assess their performance, i.e. which accessibility issues they can solve automatically or semi-automatically, and assessing the extent of manual intervention that each tool requires to output a fully accessible ebook.

Our research allowed us to identify 24 remediation tools, of which 13 were selected for the testing phase. Aside from the internal report on this activity, a public white paper was published to provide high-level guidelines that producers and developers of remediation tools can use to design or improve their tools. This white paper highlights the requirements that remediation tools should meet to be effective and to support human operators by making the process as simple, intuitive and fluid as possible, depending also on the competence of the target users. These activities are detailed in internal documents that have been used to build the public deliverables.

This is the first time, as far as we know, that a study has conducted systematic tests on this category of tools to assess their potential and limitations.

[11] *Guidelines for remediation tools producers*, Released on 2024, April the 25th. Diffusion: Public. Available as HTML, EPUB and PDF at https://github.com/ABELaboratory/publications/deliverables/guidelines-remediation-tools-producers

Analysis of remediation costs (WP4)

Previous activities have helped the project partners better understand the composition of the EU ebook backlist, establish a gap analysis methodology and test remediation tools. The main activity of this WP was to estimate the time and cost of all the activities that publishers and actors involved in ebook production must undertake to remediate backlist titles.

To achieve that, we proceeded in two ways: on one side, we interviewed publishers who have already started remediating their backlist, and on the other, we conducted tests to obtain representative times for the correction of specific accessibility problems and the entire remediation process for selected samples of varying size and complexity.

Interviewing publishers allowed us to determine that understanding the evolution of production processes is crucial. It also emphasised the need for prior analysis and for gathering information about effective practices. Publishers who outsource all of their ebook production mentioned that they struggle to evaluate the technical condition of their backlist titles.. The preferred method for these publishers is to first examine the economic feasibility of making their backlist titles accessible. They estimate which titles would be worth investing a similar amount to creating a new accessible ebook, and then request price quotes from their service providers.

For testing activity, we first selected, based on our analysis, files with an average number of accessibility issues from different categories and formats (PDF, EPUB and EPUB Fixed Layout) out of the samples that collaborating publishers provided in WP2. The choice was a balanced mixture between what is most represented, and the actual files we had at disposal. We selected thema identifiers F, Y, J, M, L, (respectively, for Fiction, Children's books, Society and social sciences, Medicine and Nursing, Law) and refined them with an analysis of the features available in each file.

Based on WP3's findings, a set of tools was selected for remediating each sample based on the starting format and the complexity of the content (images, links, tables, notes, …). The remediation was carried out by an experienced operator who has previously handled conversion and remediation activities, with an in-depth knowledge of publishers' production workflows, a proven track record in page layout programs (InDesign), and a solid understanding of ebooks accessibility requirements. The remediation process was structured into several steps:

1. automatic remediation, to assess what the tools can fix automatically;
2. tool-driven remediation (through accessibility error reports provided by the tool, in-tool guidance, etc.);
3. manual remediation within the tool, using the options and functionalities provided by the tools itself;
4. manual remediation using different tools, to correct those accessibility errors that the tool does not allow to be corrected effectively and, therefore, are still found in the output of step 3. Fixing these errors usually requires direct intervention in the HTML code (we, therefore, used an editor software named Sigil, which allows us to manipulate the EPUB code);
5. semantic enrichment. This was an additional step aimed at getting semantically richer EPUBs beyond the compliance strictly needed to be in line with the accessibility requirements defined by the EAA.

We also conducted two preliminary actions:

- file analysis to assess possible problems and
- file format conversion, when necessary.

We monitored the time required for the execution of each step using a matrix prepared for this purpose. For each step, we reported the time required to correct the different accessibility problems, starting with the most recurring ones from the gap analysis of WP2. This allowed us, for each combination of file and tool, to:

- know the total time needed for each step;
- identify which accessibility issues can actually be fixed in each step (thus, for each tool, which can be corrected in step 1 automatically, and which require manual intervention with the same or a different tool);
- know the time required to solve each accessibility issue in the step in which it was solved (except for step 1, for which only the total time can be measured).

In the case of step 1 (automatic remediation), the total time required for the tool to complete the process was reported aside as time spent by the machine is not priced the same as human time.

This approach allowed us to have an overall, detailed overview of the time associated with the remediation activity for each file and tool. The related internal deliverable *Cost benchmark report* (SEN) was finalised on the 15th of April, 2024.

Communication and Dissemination (WP5)

The project communication and dissemination activities are conducted in 3 main temporalities:

- an initial communication period to inform about the project;
- a continuous information delivery;
- a final communication stage.

The initial communication's objective was to enlist publishers interested in the project. This was done at the first level with national workshops and individual talks in France, Italy and the Netherlands in local languages and organised independently by each project partner. At a second level, we presented the project in English to publishers' associations with the help of the Federation of European Publishers. We also presented the project to Inclusive Publishing members and at the Include! Conference organised by MTM, the Swedish agency for accessible books in Malmö, Sweden. Complementary to those presentations, the three project partners produced news on their website and communication on their social media. A blog article was published by IPA. This phase allowed for widespread knowledge about the project and provided us with more publishers willing to collaborate with the project than expected.

Continuous information was spread throughout the project, with key points for releasing public deliverables, for which webinars were organised to give direct feedback to the publishers involved in the project. The project was also presented at book fairs in Bruxelles and Bologna.

The final communication stage starts with the publication of this final report and will be composed of workshops at national and international levels. Press releases, event presence and material used in public presentations are available on the project website[12].

[12] https://www.abelab.eu/

Findings: a diverse ecosystem

High variability of available data

Collecting data on the backlist ebook titles available over Europe resulted in a challenging activity since no centralised agency exists and the way industry standards are adopted and used is not yet consistent across all European countries (including identification, categorizations, and other bibliographic metadata).

There's a notable blind spot regarding the ebooks available through large online international platforms like Amazon and Kobo. We do not know how much of those contents are exclusive (like self-published titles), and therefore, remediation needs cannot be studied for these ebooks.

Regarding the ebooks available for any reseller or library, no centralised or harmonised system exists to collect data, resulting in the need to collect data from different sources (ISBN agencies, national libraries, distributors[13], aggregators[14] and data service providers). Therefore, counting the ebooks available on the EU market revealed some caveats:

- some titles are published in more than one digital format (like EPUB, PDF, KF8), which results in counting manifestations[15], not unique works;
- some ebook titles get a new ISBN with a new version, so again, we do not count unique works, but different manifestations of the same ebook;
- some ISBN agencies do not assign directly ISBNs on a per-title basis. Instead, they give publishers a large range of numbers to use and the publishers themselves assign the ISBNs to their titles. 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 withdrawn from the market;
- some distributors and aggregators handle only a part of the market. For instance, 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 insufficient 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 putting these together;
- some EU distributors and aggregators supply titles from non-EU publishers in addition to titles published in their country. Without detailed data and information, it is impossible to distinguish between titles that are published in the country itself and titles produced by foreign 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 differently, their collections are not even directly comparable to those of the traditional 'publisher-distributor-

[13] 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.
[14] We define aggregators as an entity who collects established metadata and distributes them to selling platforms.
[15] 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). For example in the digital world an EPUB and a PDF of the same title are counted as two different manifestations.

retailer' chains. For example, Amazon does not only provide ebooks with an ISBN, which is produced by publishers but also a lot of self-published titles, often without an ISBN.

The European backlist is not homogeneous

Beyond difficulties in collecting and harmonising available information, the analysis of backlist data showed that big differences exist between the collections of different EU countries, and no common denominator is emerging. Even repartition based on the size of the national backlist is not sufficient to categorise the European market.

To group titles for our project, we chose to refer to the Thema subject category scheme[16] developed and maintained by EDItEUR. Thema aims to be the subject categorization 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.[17] Because different standard classifications have different logics, those mappings may not always be one-to-one and sometimes need interpretation. One example is the CLIL[18] 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 our mappings.

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

In the end, only the extraction of statistics from the files can provide scorings capable of efficiently slicing large collections into sets of ebooks with similar needs. Such an activity can help not only with categorising accessibility needs but also with having a better view of what contents are inside different ebooks (like the number of images, number of pages, etc.).

[16] *Thema – the subject category scheme for a global book trade version 1.5*, EDItEUR, 2022. Available at https://ns.editeur.org/thema/en
[17] *Thema mappings*, EDItEUR, 2023. Available at https://www.editeur.org/151/Thema/#Mappings
[18] *Commission de Liaison Interprofessionnelle du Livre*, the French standard for book classification. Available art https://clil.centprod.com/listeActive.html

Developments go slow

Newer formats (like EPUB 3) are not adopted quickly. The newer possibilities they offer (like reflow for EPUB over PDF, better semantics for EPUB 3 over EPUB 2) do not cause older formats (like EPUB 2 or PDF) to be immediately replaced. This causes difficulties as the remediation work is more complex when dealing with older formats. Like the chicken and egg problem, the adoption of newer formats is slowed by the lack of support by main desktop tools who are not willing to invest in supporting less adopted formats. Consequently, conversion services and proprietary tools have flourished, and too few open-source tools and documentation are available.

Conversion tools as options for remediation

Our research and discussions with the developers of the tools have made it clear that many remediation tools were designed and developed to respond to the internal needs of a specific publishing house or a service provider working for one or more publishing houses, usually in the continuity of the existing tools and contracts. Most of these solutions are marketed as ebook *conversion* tools or used in the context of *conversion services*. Therefore, most of these tools are under continuous development and evolve to meet the needs of different clients' collections and specific requests.

To understand how these technological solutions can concretely fit into the workflows of publishing houses, it is important to underline that publishers can decide to use different tools to correct different accessibility problems, based on what the different solutions are capable of doing, for what formats and with what level of accuracy, which makes the remediation process a composite and complex one.

Since digital accessibility is becoming an increasingly important topic, the issue of remediating inaccessible ebooks and digital documents is becoming increasingly important too. Consequently, the landscape of remediation tools is bound to evolve and expand. For this reason, we will continue to monitor the supply of these tools even after the conclusion of the project and we'll update the dedicated list[19] on the official ABE Lab website should we find new tools, since this represents valuable information for publishing professionals.

Few open-source tools and documentation about automation are available

Our research and discussions with publishers, service providers and tools producers showed that a lot of automation is used at every level, but the rules applied for analysis and automation are established over experience and kept in-house. A reason for that can be found in the fact that it's almost impossible to average the errors in the backlist titles and that there's a will (and a need) to keep a commercial advantage. However, the consequence is that it is not possible to rely on industry-accepted rules to make sure that automation is performed for the best.

Regarding the availability of open-source tools that can be securely incorporated into remediation workflows, we discovered a limited number of available solutions beyond the widely recognised conformance checkers named EPUBCheck and ACE by DAISY.

These tools have become the industry standard and are already integrated into many publishers' production workflows. While incorporating these tools into workflows is a positive step, it's crucial to take note of the

[19] Tools for remediation tested in the ABELab project context: https://www.abelab.eu/activities/tools/

ACE developers' disclaimer: *It is important to keep in mind that only a limited portion of accessibility checks can be automated, and therefore, Ace is not a complete accessibility conformance evaluation tool; instead, it is an aid for a broader, human-driven evaluation process.*

The lack of recognised tools is somewhat worrying, as it suggests that most accessibility issues are addressed using in-house developed solutions. This results in redundant efforts when selecting and evaluating the quality of remediation operations.

The availability of open-source solutions does not imply that remediation tools necessarily have to use them, but knowing what the open-source community has already produced can be a good starting point when designing and developing, especially if these products are actively maintained and updated.

All primary areas of software development extensively utilise existing components provided by the open-source ecosystem. It is vital to the software development landscape, offering scalability, flexibility, and robustness. It helps everyone in the ecosystem save time and money, enhance quality and security, foster innovation and collaboration, and gain a competitive edge in the market.

Is there a market for remediation tools?

Publishing is an umbrella name for activities related to the creation and transmission of culture through books. Many different professional profiles work in the industry. From rights acquisitions to innovative marketing techniques, the shared knowledge about ebooks and digital reading is not so wide and broad.

ebook sales represent around 12%[20] in the best cases, and are still usually a digital replica of the printed object, being as much as possible similar to the paper object.

Publishing houses, except the bigger ones who may have an internal digital department, rely on the expertise of specialised external services for the conversion in digital versions of the printed ones..

Therefore, the following question arises: is there a market for remediation tools dedicated to publishers? Or should we think that service providers are the target of remediation tools vendors?

Post conversion over born accessible

Accessibility experts have long emphasised the importance of considering accessibility from the beginning of the publishing process, as it can make a significant difference at a minimal additional cost. [21]. This is also a key principle of digital-first workflows and single-source publishing.

However, the reality is that the publishing industry is still largely focused on print. Books are typically designed for the printed edition first, and then converted to create an ebook. This means that many ebooks, especially those in the backlist, are missing important digital information about their structure and formatting.

Instead, the structure of the book is often represented visually, such as using italics for quotes or different fonts for headings. Only a few publishing workflows are set up to manage rich semantic structures and create ebooks that are natively accessible.

[20] European Book Publishing Statistics 2022, Federation of European Publishers (FEP). https://fep-fee.eu/European-Book-Publishing-Statistics-2022
[21] See, for example: https://daisy.org/info-help/time-to-use-the-modern-digital-publishing-format/

There are tools that can facilitate the conversion process, but they are often linked to a specific service or offered as part of hybrid licences, such as software as a service or different levels of conversion service.

External expertise

As done for the creation of the digital versions, most publishers view accessibility expertise as a service provided by third-party contractors, rather than an inherent part of their own production process. However, the definition of accessibility expertise is not well-established, and publishers may encounter varying levels of quality and accuracy in the services they receive. This can lead to different experiences, including misleading or deceptive practices that fail to meet the necessary accessibility standards. As a result, publishers must exercise caution and carefully vet potential contractors to ensure they possess the necessary expertise and can deliver high-quality accessible ebooks.

France and Italy are exceptions where publisher associations provide accessibility expertise. In France, this is done through guidelines and discussions established by the *Normes et Standards* working group. In Italy, this is done through the awareness of the topic raised by Fondazione LIA, which is strongly related to the Italian Publishers Association.

Many countries are still in the process of finding an effective approach to accessible publishing. Collaborations between publishers and specialised organization serving people with disabilities which may have specific knowledge on some aspect of accessibility, may be fruitful. The Apace project[22], recently funded in the Creative Europe programme, will promote this collaboration helping to pave the way for a more inclusive publishing future.

[22] https://www.fondazionelia.org/en/project/2024-apace-accelerating-publishing-accessibility-through-collaboration-in-europe/

Outcomes: methods and guidelines

Recurrent accessibility issues

The gap analysis allowed us to establish a list of the most recurrent accessibility issues found in the backlist ebooks, based on the collected samples[23]. This list has been used for other activities and can be considered a baseline for remediation needs that will apply to most collections. The list is detailed in the *Gap analysis* section of the *Report on backlist data and gap analysis*. As a quick reminder, we are just providing here a resume:

- Missing Accessibility Metadata
- Non-reflowable content
- Missing or bad textual alternative for graphical resources
- Missing or bad Language Tags
- Structure errors (Empty-Table-Header, Empty-Heading, Heading-Order)
- Missing semantics (Html-Has-Lang, Aria-Roles, Link-Name, Document-Title, Epub-Type-Has-Matching-Role)
- Possible navigation improvements (Landmark-Unique, Epub-Toc-Order, Epub-Pagelist-Broken)
- Contents and styling errors (Link-In-Text-Block, Color-Contrast)

Additionally, there are also accessibility issues that can't be detected automatically and that requires ad hoc human testing to be identified. This concerns *reflowable restrictions* imposed by style rules applied to the file, as well as s*pecific contents* such as forms, scripts, maths, videos and audio, which are not usually used in ebooks, but as this may happen, it will be necessary to include them in the remediation process should they occur in a file.

Test suite

To test if the identified remediation tools offer functionalities to fix the identified accessibility issues, we've built a set of test files, one for each starting format (EPUB 2, EPUB 3, EPUB Fixed Layout, PDF). We used these test files to carry out functional tests to assess which functionalities were present in a tool and which were missing. In these files, we included all the most recurrent accessibility issues according to our analysis of the publisher's provided ebook backlist.

The test suite is available from the ABE Lab website *Outcomes* section[24] and is composed of five files:

1. an EPUB 2 file, to be used to check the features of the remediation tools that claim to turn inaccessible EPUB 2 into accessible EPUB 3;
2. an EPUB 3 file, to be used to check the features of the remediation tools that claim to turn inaccessible EPUB 3 into accessible EPUB 3;
3. an EPUB 3 with multimedia (audio and video) only;
4. an EPUB 3 Fixed Layout, to be used for the remediation workflows from EPUB 3 Fixed Layout to "accessible" EPUB 3 Fixed Layout and from EPUB 3 Fixed Layout to EPUB 3 reflowable;
5. a PDF file to be used to test the tools that convert a PDF file into a PDF/UA file or in an EPUB 3 file.

[23] For a complete and detailed list of the most recurrent accessibility issues detected, see *Recurrent accessibility issues detected*, in "Gap analysis", ABE Lab, https://www.abelab.eu/outcomes/gap_analysis/#recurrent-accessibility-issues-detected
[24] https://www.abelab.eu/outcomes/test-suite

For publishers

As there are too few common denominators to classify backlist titles, and because the EAA asks to assess and evaluate disproportionate burden per title, the minimum to do to keep titles in sales is to acquire a good knowledge of the collections.

This part can be done with support of existing automated evaluation tools and distributors' sales reports. These steps will generate a better understanding of the existing and therefore, reinforce the publishing house at different levels, not only those related to accessibility.

How to triage files?

Collecting information and statistics is necessary to start a triage and track decisions per book. At the moment of this report, no known available tool exists to do so. This can be implemented in an integrated information system or, by default, as a spreadsheet. For each title, it is necessary to collect at least the following baseline information:

- File format
- Other formats the title is available in
- Known production information like
    - Year of production
    - Source file (availability and format)
    - Service provider, if known
- Known editorial information like
    - The presence of languages different from the main language
    - Images common to all books (cover, logos)
    - Number of images (except the common ones)

Additionally, it is an easy step to collect EPUBCheck and ACE reports. This will allow publishers to add the following columns to the spreadsheet:

- Number of EPUBCheck errors (critical, serious and minor)
- Number of ACE errors (critical, serious and minor)

More statistics and in-depth analysis can be generated following the methodology used by the ABE Lab gap analysis or requiring services from an accessibility consultancy organisation. We also expect that more detailed analysis tools will be available shortly, as proposed by distributors, service providers, or other actors. In the context of ABE Lab, we have calculated three additional statistics sets:

- statistics on the usage of HTML tags, attributes and values
    - Number of times each tag and attribute appears in the file
    - Attribute values and how many times a certain value appears for a certain attribute (e.g. lang, class, role)
- a check whether the language attribute given by the content creator is congruent with the text or not. We analysed the text of all paragraphs with more than 5 words using a model trained with Machine Learning techniques, which can guess language from plain text[25];
- number of words per XHTML file.

[25] Wrong language assertion is done with the help of the Compact Language Detect 2 (CLD2). We compare the output of CLD2 with language tags specified at the tag level and file level.

How to evaluate cost?

The monetary costs of remediation of various ebooks vary strongly from one country to another. Therefore, it is not possible to provide an estimate in terms of prices and costs. Instead, we share the time frame that, based on our experimentation, requires correcting each type of accessibility error.

Remembering that one-off costs may apply to the organisation is also important. These may include awareness, initial planning, workflow set and tools selection, as well as training people and tools. Since most publishers will first focus on modifying their workflows or specifications for their contracted service providers to make their new ebooks accessible, the associated organisational costs, such as awareness and training, will likely be incurred before the remediation process for backlist titles begins. Certain additional information will be necessary to decide which backlist titles will be remediated, which titles may be kept under disproportionate burden, and which may be removed from the market. This may include legal information, cost estimates for different remediation options and information about the revenues to be expected. Insights on what it takes to remediate existing titles will be the most difficult to get, and for this, some consultancy might be required.

Defining the production cost associated with the backlist ebook remediation activity requires taking into account which interventions can be carried out automatically by the available remediation tools, which adjustments require the intervention of a human operator expert on accessibility, what the post-remediation quality control efforts are, and if there are additional costs related to reintroducing the title in the market. The next subsections give some insights to help with the evaluation of production costs.

Primary considerations

There is currently no remediation technological solution or service on the market that can process a large collection of files in one go, but using the *Readium Go Toolkit*, it is possible to parse EPUB files and add some basic accessibility metadata that can be inferred from the ebook file to large collections (e.g. it can tell if the ebook contains a table of contents, mathematical formulas in MathML or images, but not if the associated alt text is meaningful, and it cannot check whether the language attribute given by the content creator is congruent with the actual language text).

There are tools for reflowable EPUBs that achieve a good level of remediation at an affordable cost. But if the digital version of a publication has been produced as a fixed-layout EPUB or PDF, based on the results of our project, the most achievable option is not to remediate the Fixed Layout or PDF file but to produce a newborn accessible version of the title as a reflowable EPUB.

Graphical resources need dedicated treatment as they require an editorial decision (is it necessary to add an alternative text?) and action (writing an accurate alternative text). ACE report can provide a list of images to help with this process.

Time for basic remediation of EPUB2 to EPUB3

Based on our testing activities, we estimate that with the currently available tools, 1 hour for an average EPUB 2 to be transformed to EPUB 3, and then remediated to meet the accessibility requirements of the EAAis a reasonable time. Further enrichments (like adding page breaks and page numbers) are not included in this evaluation.

The following list gives more details including minimum, maximum and average times spent by the operator who performed the remediation tests:

- for the 'simple' fiction title (320 pages, 14 images), the total time was 43 minutes minimum, 52 minutes on average, and 67 minutes maximum

- for a children's book with mainly text, only a cover image (288 pages, 19 images) the total time was 39 minutes minimum, 43 minutes on average and 48 minutes maximum
- for a title with 'medium complexity' (category Society and Social Sciences) (277 pages, 60 images, 20 tables, 150 links, 192 footnotes, 46 chapters) the total time was 55 minutes minimum, 64 minutes on average and 73 minutes maximum. Note that time refers to time needed to copy-paste a generic image descriptions, not the time required for writing them.

Time for conversion and remediation of PDF to reflowable EPUB3

Based on our testing activities, we estimate that converting a PDF into an EPUB 3, and remediating the output file will take more time than converting an EPUB 2 into an EPUB 3 and making it accessible (based on the results of the 'average' files tested). Typically a day per file would be plausible. Actual time per real book needs to be assessed case by case.

The following list gives more details including minimum, maximum and average times spent by the operator who performed the remediation:

- a sample with 'medium complexity' (Category Law) including 83 pages, 1 image, 55 links (hyperlinks and footnotes) took 54 minutes minimum, 86 minutes on average and 118 minutes maximum. The full book is 492 pages including the cover, 1 image, 57 links, and around 770 footnotes, an EPUB would be around 24 chapters);
- An extract of category Medical (22 pages, 13 images, 1 table and 4 links) took 63 minutes minimum, 72 minutes average and 81 minutes maximum. The full book is 359 pages including the cover, 164 images, 21 tables, and 104 links (not counting TOC and index);
- an illustrated children's book (57 pages, 37 images) took a minimum of 22 minutes, an average of 25 and a maximum of 28;
- another non-fiction extract, including 40 pages, 15 images, 31 links (hyperlinks and footnotes) took a minimum of 34, an average of 44 and a maximum of 55 minutes. The full book is 136 pages, 75 images, 146 links, and 28 footnotes).

When to start remediation?

Of the accessibility correction actions, the addition of accessibility metadata is the one that can be accomplished most quickly. This can be partially done by a batch use of the *Readium Go Toolkit* (an open-source tool that can derive some basic and partial accessibility metadata directly from files), by an in-house build script reusing the inference rules proposed by the Readium Foundation, or by publisher-specific inference rules created from the knowledge of the collections. The result will be the addition of accessibility information to ONIX records. We expect that some distributors will propose prefilled formulas or Excel files depending on the usual interexchange procedures in place. However, in most countries, publishers remain legally accountable for the information provided, so understanding the meaning of the most important accessibility information is necessary. This learning can be achieved by reading the resources proposed in the dedicated appendix.

For tool producers

Based on the recurrent accessibility issues found in the sample collections, and built on the results of our testing activities, we have detailed high-level guidelines including the main functionalities and information that tool producers should provide to help publishers solve efficiently accessibility problems found in their files. Tools producers should refer to the *Guidelines*[26]. We are providing here a quick summary of key points.

Find, *fix*, and *check* are the three key aspects of accessibility remediation. Tools are expected to support expert human operators by identifying the accessibility problem, reporting it to them, and guiding them for applying corrections. For the remediation to be effective, the tools should therefore:

- find the accessibility issues contained in the ebook;
- provide a way to correct the identified accessibility issues, for example, via dedicated options;
- check the conformance of the file to ensure the output will be valid with the file format specification and compliant with the target accessibility guidelines or standards. Accessibility checks carried out before export, or at the operator's request, are also strategically important to ensure that the ebook does not have accessibility problems before it is exported and checked with external tools.

Publishers' needs vary greatly, depending on the size of the collections, in-house knowledge, and many other factors. Therefore, the remediation tools that will be used can take many different forms and levels of complexity. It is important to identify which operator audience the individual tool is aimed at by providing clear, detailed documentation on how the remediation tool works and what it offers to make the work easier and faster. Minimum viable information that must be found is as follows:

- Input and output formats
- Accepted contents complexity
- Accessibility Target (EPUB Accessibility, WCAG, EAA, ADA, PDF/UA)
- Type of licence and plan
- Safety of the data
- Machine Learning policy

Although the objective of the ABE Lab project is not to guide the design, usability, and accessibility of graphical interfaces of remediation tools, it would be great if these were compliant with the requirements defined by the Web Content Accessibility Guidelines (WCAG) 2, the User Agent Accessibility Guidelines (UAAG), and the technical standard EN 301 549. An accessible interface would represent added value and make the remediation tool usable by many operators. The following key points should be regarded with caution:

- In-tool support. There should be attention to providing as much support as possible to guide the human operators in fixing the accessibility errors, to reduce the number of iterations needed to the minimum;
- Improve, don't disrupt. The remediation tool must not cause a loss of information, must not add accessibility errors and must not compromise the technical integrity of the file. When converting a PDF into an accessible PDF, the original layout and design should be maintained: fonts, colours, and graphic elements should not be affected during the remediation phase, and the accessible file respects the graphics of the source file.

[26] *Guidelines for remediation tools producers*, Released on 2024, April the 25th. Diffusion: Public. Available as HTML, EPUB and PDF at https://abelab.eu/outcomes/deliverables/guidelines-remediation-tools-producers

An ebook can contain many different types of accessibility errors. Analysing a sample of backlist files, we found that some accessibility problems recur with particular frequency; these are detailed in the Gap analysis[27] and resumed here:

- Accessibility metadata
- Alternative text for images
- Missing or bad language tag
- Headings hierarchy
- Colour contrast
- Document title
- Semantics

Check your tool functionalities

Evaluating a tool is a complex activity and involves multifaceted aspects. One of these is represented by the expectations of those who evaluate the instrument, which largely depends on their specific needs. A tool could be evaluated based on the results it allows to obtain even when used by non-expert users. At the same time, what is sought may be a tool capable of integrating effectively and fluidly into existing workflows and which therefore potentially requires advanced coding and programming skills. The ideal solution may also be platforms or programs that allow making batch modifications to large quantities of files. Combining such broad and varied expectations would make a complex checklist. Hence, the checklist we prepared focuses on the features related to the most recurring accessibility issues highlighted by the gap analysis.

The checklists reported in part 4 of the *Guidelines*[28] can be used as methodological support that allows one to obtain a clearer view of the current state of a tool. They are provided to help stakeholders, remediation tools producers and developers check the state of the art of their tools and to plan future developments.

However, they do not claim to be a complete and definitive list to assess the quality or completeness of a remediation tool, nor does meeting all the points in the list guarantee that the remediated files will be fully accessible and compliant with the requirements of the EAA and international accessibility guidelines.

It's also important to highlight that these guidelines might not be maintained after the end of the project. Therefore, they can not be used to claim conformity with the latest level of required accessibility and must not be used as a marketing or commercial argument.

The points of the checklists and the content of the guidelines are high-level indications that can be relevant for remediation tools regardless of the input file format.

[27] For a complete and detailed list of the most recurrent accessibility issues detected, see *Recurrent accessibility issues detected*, in "Gap analysis", ABE Lab, https://www.abelab.eu/outcomes/gap_analysis/#recurrent-accessibility-issues-detected
[28] *Guidelines for remediation tools producers*, Released on 2024, April the 25th. Diffusion: Public. Available as HTML, EPUB and PDF at https://github.com/ABELaboratory/publications/deliverables/guidelines-remediation-tools-producers

Leveraging impacts

The expected impacts of the ABE Lab project detailed in the Activities section are considered over three periods: short term (before 2025), medium Term (2025 – 2026) and long Term (2026 – 2030). They follows three axes:

- analysis of the composition of the ebook backlist
- classification of the remediation needs of the ebooks of the backlist
- assessments of the associated costs.

We hope that the outcomes of this project will be useful in achieving significant impacts in the publishing industry. However, we recognize that there are still challenges related to consolidating these impacts over the long term. As a result, we propose additional activities that we believe would be valuable for all stakeholders in the publishing chain. These activities should address the needs of a diverse ecosystem, including smaller actors, by providing tools and resources. However, it's important to consider the sustainability of every actor involved, beyond just technical solutions. This includes the work of experts on standardisation and advocacy, as well as efforts to raise awareness and build confidence among readers who would benefit from accessible ebooks. Ultimately, these activities should aim to create a sustainable and inclusive publishing ecosystem that benefits everyone involved.

The consolidations to consider are following four axes:

- efficient decision-making at organisation and national levels, enabled by a clear picture of the state of the ebook backlist;
- publishers level evaluation of the composition of their collections and associated remediation costs based on expert-established analysis methodology;
- a greater number of ebooks made accessible with a high level of quality made possible by enhanced technical workflows and tools;
- an easier process for future production and a better repartition of knowledge over the value chain thanks to the sum of information provided in the project communications[29].

The next sections detail our propositions to provide ecosystem-level mechanisms and help with shaping the future of accessible ebooks.

Consistent data collection and analysis methodology

ABE Lab represents a very new picture, providing as good work as possible per today's state of data collection (more than expected at the beginning). Publishers, unions for the blind and visually impaired, and public services have, for the first time, a picture of the status of the European ebook backlist, but it's a quick snapshot; available data will not allow for yearly comparisons in the foreseeable future and the view presented here might soon become obsolete.

Another aspect is that the picture is clear, but the content is noisy, and clarifications could only be done by yearly aggregations and refining. There is a need for a longer time view to highlight and valorate the industry efforts and progress made.

[29] ABELab Communication are available from the Activities section of the abelab webpage: https://www.abelab.eu/activities/

Classification wizard service

ABE Lab has developed a methodology for classifying ebooks in a way that supports efficient remediation. This methodology can be used by more technically oriented publishers to evaluate the composition of their own ebook backlist.

To address this gap, there is a need for tools that can help publishers calculate the cost of remediation for each segment of their backlist and support them in defining their priorities. These tools should be designed to be user-friendly and accessible to publishers of all sizes.

However, creating and maintaining such tools requires specific investments and a sustainable economic model. It's important to ensure that these tools are not only effective in the short term but also viable over the long term. This may involve exploring different funding models, such as subscription-based services or partnerships with industry stakeholders.

Ultimately, the goal is to provide publishers with the resources they need to make informed decisions about ebook remediation and prioritize accessibility in a way that is sustainable and cost-effective. By doing so, we can work towards creating a more inclusive publishing ecosystem that benefits everyone involved.

Enhanced workflow documentation

One of the stronger reasons for the permanence of old formats is their long-time integration in production and reading environments. Facilitating the production, conversion, enhancement and manipulation of file formats is a key factor in helping with their adoption.

The role of the Readium Foundation in facilitating the building of reading systems has been decisive in the diffusion of the standardised EPUB format and created the environment for the rise of economic model opportunities, liberating the industry from the silos built by big tech companies. Because open-source SDKs rely on standards, it also opened the way to more widely adopted accessibility features in reading systems.

There is an equivalent need for EPUB production: more authoring tools are needed, and to open the way and shape a viable economic environment, documentation and open source SDKs are a way to pave the road. Helping those developments is as necessary as the consultancy provided to mainstream products.

Research on AI

Artificial Intelligence, Machine Learning and other forms of automation can speed up remediation work. However, these technologies are far from perfect and sometimes lead to unexpected, if not incorrect, results. Since this technology is rapidly evolving, studying and monitoring future developments will be crucial. A publisher-oriented solution is also becoming necessary to work on editorial products efficiently.

Publishers and other actors in the publishing chain have been using automation, including Natural Language Processing and Artificial Intelligence technologies, for several years already. Now facing the revolutions of Large Language Models (LLM) and Generative AI technologies (GAI), they are observing, testing and defining positions towards this technology and its use in editorial workflows. This has led, in some cases, to express concerns related to the uncertainty about how their contents and files are used by such systems.

In the current state of the art, Machine Learning systems cannot replace the competence and evaluation ability of a human being. If it is possible, for example, to automatically generate alternative descriptions for

images with AI, it has yet to be proved that these alternative descriptions can provide information equivalent to that conveyed by the images themself or that they are comparable to those produced by a subject expert. Human decision-making is still key to determining whether such descriptions are appropriate for a given context, and there are probably ways to facilitate this validation operation.

More accurate results can probably be achieved by additional actions to help conduct machines to the most probable correct and pertinent textual alternative and also to help with automated evaluation of this pertinence or detect probable wrong or insufficient proposals. These actions can be built only on data mining over large curated collections and through multiple iterations. To make sure that such experiments are conducted taking into account publishers' concerns and the results are assessed in relation to the real needs of readers, we think it is important that professional membership organisations and accessibility experts are involved in monitoring research activities.

Appendix

The two following subsections establish industry events and key points to collect to establish a good knowledge of the collections. All of them will probably not apply to each collection; they are to be assessed and selected based on the country, the publishing house, and the ebook collection.

Industry facts to know

There are some dates that have affected the ebook production workflow widely. They are good to know and may help identify, understand and classify ebooks. However, they did not affect all files, and certainly not simultaneously.

EPUBCheck is a validation tool for EPUB files, and it has undergone several versions since its inception. Here are the main EPUBCheck versions[30] and their corresponding EPUB specifications:

- EPUBCheck1.x: this version was released around the time of the EPUB 2.0.1 specification, published in 2010. EPUB 2.0.1 was an update to the original EPUB 2.0 specification, released in 2007;
- EPUBCheck3.x: this version corresponds to the EPUB 3.0 specification, which was published in 2014. EPUBCheck 3.0.1 was released in 2015, and EPUBCheck 3.2.1 was released in 2021. These versions of EPUBCheck are used to validate EPUB 3.0 and later files;
- EPUBCheck 4.x: this version corresponds to the EPUB 3.2 specification, which was published in 2017. EPUBCheck 4.0.0 was released in 2018, and EPUBCheck 4.2.1 was released in 2021. These versions of EPUBCheck are used to validate EPUB 3.2 and later files;
- EPUBCheck4.1.0 (November 2018) was the first release under the DAISY Consortium maintenance role. It notably introduces better support for the EPUB Accessibility Vocabulary, reports duplicate landmarks and not unique landmarks anchors and reports for CSS absolute positioning as well as font size validation;
- EPUBCheck5.0.0 (January 2023) is the version testing EPUB 3.3 rules. It notably includes checking *epub:type* restrictions, considering HTML form elements as scripted content, reporting empty title elements in XHTML Content Documents, allow remote resources in scripted content and fonts, disallowing remote links in TOC, landmarks and pagelist navigation, allowing *epub:type* on all HTML elements as well as extend authorised ARIA roles.

It is also important to notice that EPUBCheck includes checks from the W3C markup validator[31].

ACE 1.0 was released in January 2018. Two intermediary versions are to consider: 1.1 (April 2019), 1.2 (April 2021). Since then, one maintenance update has been released almost every year. ACE's full changelog can be consulted on the ACE GitHub repository.

Adobe InDesign[32] has supported EPUB 3 export since the release of InDesign CC 2015 in June 2015. InDesign CC 2017 introduced support for audio and video embedding, and InDesign CC 2018 introduced PDF Accessibility enhancements and added support for MathML. Fondazione LIA is leading a group of experts working to provide insights to Adobe and guide the InDesign development team on missing or incorrect InDesign export features for accessible EPUB. The group is the result of a common request done to Adobe by Fondazione LIA, Daisy Consortium, IPA (International Publishers Association) and FEP (Federation of European Publishers)[33].

[30] Full changelogs of epubcheck can be found at https://www.w3.org/publishing/epubcheck/releases/
[31] W3C markup validator change logs can be found at https://github.com/validator/validator/releases/
[32] InDesign changelogs are not centralised. The *what's new* page points the last updates and has a section pointing to most recent previous versions: https://helpx.adobe.com/indesign/using/whats-new.html
[33] Exemples and discussions can be found at https://github.com/ways2read/InDesignA11Y

Amazon's Kindle Direct Publishing (KDP) platform has accepted EPUB files for conversion into the Kindle format (.mobi or .kpf) since 2016. Amazon does have its own validation process to ensure that the EPUB files uploaded to KDP meet certain quality standards. This validation uses a custom validation tool based on the EPUB specification to check for potential issues and inconsistencies in the uploaded files.

Kobo has been using EPUBCheck as a requirement for ebooks published through their platform, Kobo Writing Life, for several years. Although there is no specific date mentioned in their public resources, it is generally assumed that this requirement has been in place since the early days of Kobo Writing Life, which was launched in 2012.

Apple Books has been accepting EPUB files since the launch of the iBooks app in 2010, and it remains the standard file format for publishing ebooks on the platform. Apple Books does not explicitly impose EPUBCheck as a requirement for EPUB files. However, Apple strongly recommends using EPUBCheck to validate your EPUB files before submitting them for publication on Apple Books.

While EPUB is an open standard for ebooks, some ebook vendors and publishers use proprietary technologies within EPUB files to add enhanced features or to protect their content. It's important to note that while these proprietary technologies can add enhanced features or protect content, they can also limit interoperability and compatibility between different ebook vendors and devices. As a result, it's generally recommended to use open standards and avoid proprietary technologies whenever possible.

Key points to assess the knowledge of collections

Ebooks have been produced over more than twenty years. In this timelapse, many things may have happened, from responsible people shifting to stakeholders' decisions and passing by distributors' and vendors' policies.

To be able to identify groups of ebooks in the backlist with potentially similar remediation requirements, first of all, it is important to reconstruct the production history of the backlist files. The following elements may help separate ebooks in stacks per year of production. They are presented as an ordered list for easy reference, and the order reflects their importance.

1. *Was there a file quality control established, what did it address and with which tools?* File quality control is usually a checklist that helps eliminate some common errors. If industry-recognized tools like EPUBCheck or ACE were used, it would help identify some patterns (see section *Industry facts to know*).
2. *Was there, and since when, one or more specifications, and did it have a dedicated accessibility requirement?* The presence of contractual documents certainly gives strong indicators that can be used to separate collections.
3. *Is there internal documentation of the choices made in production?* In complement or by default of the precedent structured documents, editorial and production choices may have affected the quality of the produced ebooks. This may include graphical resources policies, decisions to convert tables as images, addition of page markers, etc. Such documentation can be rebuilt by collecting and dating events.
4. *Is there, and since when a person responsible for the digital production activities?* When responsible for digital production, people usually have a long-time view that implies minimum consistency in the choices made.
5. If the production has been outsourced, *has the relationship with service providers been maintained? Are they still suppliers?*
6. If the production has been made in-house, *how many people have been involved, and how long did they stay on?*

Secondly, it is necessary to assess the knowledge of the collections.

1. How many titles are there in the backlist?
2. Are the digital versions available in different formats? Is there a need to maintain all formats?
3. What software was used for production?
4. Are the source files still available in a usable format?
5. What resources are included in the ebook?

DOCUMENT CONTROL INFORMATION

Settings	Value
Document Title:	Public report
Project Title:	ABE Lab
Project Manager (PM):	EDRLab
Author	EDRLab
Doc. Version:	1.0
Sensitivity:	Public — fully open
Date:	30/04/2024

Document Approver(s) and Reviewer(s):

NOTE: 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*	15th may 2024
Fondazione LIA	Project partner & co-author	*approved*	15th may 2024
KB	Project partner & Author	*approved*	15th may 2024

Document history:

The Document Author is authorised to make the following changes to the document without requiring that it 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).

Revision	Date	Created by	Short Description of Changes
Initial publication	17th may 2024	EDRLab, Fondazione LIA, KB	Initial publication

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 in: https://github.com/ABELaboratory/publications/