EBOOK FORMATS EVOLUTION

The state of the art and future of digital publications

Gregorio Pellegrino,
Chief Accessibility Officer,
Fondazione LIA

FONDAZIONE LIA
LIBRI ITALIANI ACCESSIBILI

Fundación Germán Sánchez Ruipérez

International Publishing Distribution Association

BORN ACCESSIBLE PUBLICATION
LIA

SUMMARY

INTRODUCTION

From 2017 onwards, the field of digital publishing standards is experiencing a new phase; while the arrival of EPUB 3 gave the impression of having reached its peak in terms of the possibilities offered by the digital world for publications, the move from the International Digital Publishing Forum (IDPF) to the World Wide Web Consortium (W3C) gave new impetus to inventiveness in this area with the design of Web Publications: publications that can be used online and offline, natively in the browser.

What we are trying to find is **the Holy Grail of publishing**: a unique format for any type of publication. As we work on the specifications of new formats we must not forget that there is an important player in terms of numbers, Amazon, which uses its own digital format (.mobi) with its own rules, which inevitably affects the perception of the final consumer with respect to digital publishing.

In this paper we will try to describe the state of the art of digital publishing and the future scenarios that may arise, trying not to forget the role of Amazon.

THE MERGER IDPF AND W3C

On January 30, 2017 the body that had published and maintained the specifications of the EPUB format since 1999, the International Digital Publishing Forum (IDPF), was incorporated into the World Wide Web Consortium (W3C), the consortium that since 1994 defines the specifications for web-related languages and was founded by Tim Berners-Lee, the inventor of the www.

This was an important step for the publishing community: on one hand the EPUB format has always been based on web technologies (HTML and CSS), on the other hand the web is the publishing platform par excellence, even for non-traditional content (such as multimedia and interactive or social networks).

The unification of efforts seemed so natural, that in 2016 Tim Berners-Lee declared:

> *«We share an exciting vision for W3C and IDPF to fully align the publishing industry and core Web technology. This will create a rich media environment for digital publishing that opens up new possibilities for readers, authors, and publishers. […] Think about educational text books. The book content we know today is becoming highly interactive and accessible with links to videos and images from actual historical events and original research data. This provides greater authenticity and a more engaging learning environment for teachers and students.»*[1]

For the publishing world, being able to take full advantage of the Open Web Platform, i.e. the set of open-source technologies developed by the W3C for the web, offers new possibilities of integration between online and offline (as we will see for Web Publications) and content enrichment; moreover, the fact of having the publishing world within the consortium mainly made up of IT companies leads to a common enrichment. An example are the head letters, which have been inserted within the specifications of CSS 3, at the express request of the publishers; or the WAI-ARIA tags, which, thanks to the work of the publishers, have expanded the dictionary of words available to describe portions of text (e.g. cover page, summary, colophon, etc.).

The integration between the two organizations was not sudden, but came after a long period of engagement: since 2013 there was, within the W3C, the Digital Publishing Interest Group whose components partly overlapped those of the IDPF. The idea of Web Publication also comes from previous works: in 2014 Ivan Hermann (W3C) and Markus Gylling (IDPF) at the Books in Browsers conference in San Francisco presented their vision for the future of digital publishing entitled "Bridging the Web and Digital Publishing: EPUBWEB" (whose slides are still available online[2]), which in large part can be considered the starting point of Web Publications.

Working groups and internal organization

The world of digital publishing within the W3C is divided into three working groups, of three different types:

- **Publishing Business Group**[3], is a Business Group that is focused on a wide and specific

1 https://www.w3.org/2016/05/digpub.html.en

2 https://www.slideshare.net/ivan_herman/bridging-the-web-and-digital-publishing-epubweb

3 https://www.w3.org/publishing/groups/publ-bg/

business area, publishing; it is composed of members of the W3C or members of the Business Group only (with a smaller shareholding) and serves as a channel for feedback between the publishing ecosystem and the W3C; this type of group can publish documents and specifications, but the results are not formal W3C Recommendations.

- **Publishing Working Group**[4], is a Working Group that defines specifications related to digital publishing: these specifications are formal W3C Recommendations and there is an IPR protection on the results; only members of the W3C can be part of the group; **it is in this group that the new standards for the digital publishing of the future are defined**: the aim of the group is to allow all types of publications, with all their peculiarities and traditions, to become primary entities of the web (therefore accessible directly through browsers); the purpose of the group is also to provide the necessary technologies on the Open Web Platform to complete the integration between traditional and web publishing in terms of accessibility, usability, portability, distribution, archiving, offline access to content and reliable cross-references.
- **EPUB 3 Community Group**[5], is a Community Group whose objective is to maintain and improve the specifications of EPUB 3, and everything that revolves around it (such as the EPUBCheck, the guidelines for EPUB Accessibility, etc.); it is, basically, a working group that inherits and maintains the work done by the IDPF. **Anyone can be part of the group** (without having to pay any registration fee). The group may publish documents and specifications, but those are not formal W3C Recommendations.

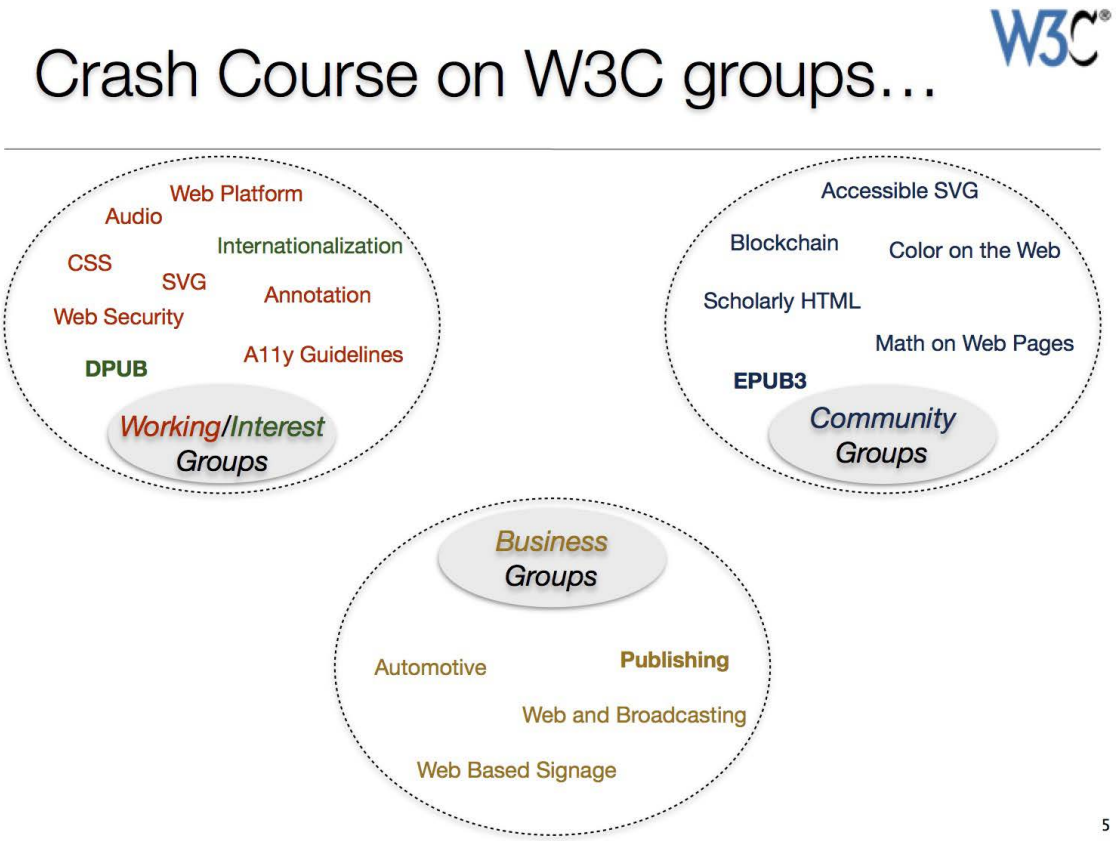

Figure 1 Different types of W3C groups, from Ivan Herman (W3C) presentation at EPUB Summit 2017, Bruxelles[6]

4 https://www.w3.org/publishing/groups/publ-wg/

5 https://www.w3.org/publishing/groups/epub3-cg/

6 https://www.w3.org/2017/Talks/EPUBSummit-IH/

The importance of being there (even for book sellers and distributors)

At the moment, the work within the W3C is the avant-garde in terms of digital publishing, but the innovation concerns the entire digital publishing chain and not only the content producers, so it is important for distributors, aggregators and retailers of ebook to be part of the working tables and not just be spectators: only in this way you can be actors of change, offering your contribution or your doubts, and not be overwhelmed.

As we will illustrate in the following pages, inside the publishing dedicated groups within the W3C, professionals are working on many new features, some of which disruptive for the world of distribution and sale of ebooks (as we know it now). It is therefore essential to understand how the publishers are moving, in order to anticipate the business choices. It's important also to be aware that some small and large players involved in aggregation, distribution and sales are already more or less actively part of working groups such as Google, Rakuten Kobo, Microsoft, VitalSource.

THE EPUB FORMAT

The EPUB format is the digital format most used by publishers o publish ebooks in the trade channel (both fiction and non-fiction); originally developed and maintained by the International Digital Publishing Forum (IDPF), it is now maintained by the World Wide Web Consortium (W3C).

The specifications of EPUB 2, the most widely used version on the market, date back to 2007 and have been declared obsolete since June 2014 and will therefore no longer be maintained. The most recent version, EPUB 3, initially published in 2011 as an evolution and enrichment of EPUB 2, represents today the state of the art of what can be achieved as a digital publication: its specifications make it in fact an optimal format for the creation of both simple texts (narrative and non-fiction), and complex texts (school and professional), also offering the possibility of including interactive and multimedia elements.

From a technical point of view, the EPUB 3 is based on the technologies of **the Open Web Platform**:

- (X)HTML latest version, i.e. the XML inflection of **HTML 5**: in simple terms, each opening tag must match a closing tag;
- CSS snapshot, the latest version of **CSS 3** with some prefixed-properties specific to the EPUB mainly related to the management of the layout[7] and with the ability to take advantage of the media queries;
- **SVG** latest version, for using vector images;
- a subset of **MathML** to represent mathematical formulas;
- **JavaScript** to create interactive and dynamic content.

Some important elements of the specifications are:

- support for **audio and video** (mp3, mp4 formats);
- the possibility of producing **fixed layout** publications (e.g. for children's books);
- the possibility of inserting in the plugs (the list of documents to be shown to the user) graphic files (images or vectors), which make the format very convenient for the publication of comics, such as Japanese Manga.

At the time of publication of this document the latest public version is the EPUB 3.1[8], whose technical specifications are available on the IDPF website, although this version has not been very adopted by publishers, also because the Epubcheck, currently, does not support the control of these files.

EPUB 3.2

With the passage of the IDPF into the W3C the maintenance of the EPUB has been taken over by the EPUB 3 Community Group[9] (open to anyone wishing to participate); this working group is dealing with the publication of the EPUB 3.2 version, whose release is scheduled for the last quarter of 2018. The new version is mostly a maintenance version that will add few new elements, without breaking the backward compatibility, rather going to reintroduce some elements present in EPUB 3.0.1, but removed in EPUB 3.1.

The ideas behind the new specifications are:

---

7 http://www.idpf.org/epub/31/spec/epub-contentdocs.html#sec-css-prefixed

8 http://idpf.org/epub/31

9 https://www.w3.org/community/epub3/

- any publication now available in EPUB 3.0.1 version will be compatible with EPUB 3.2 without changes (great news for the publisher's catalogue);
- any Reading System that supports EPUB 3.0.1 will be able to render an EPUB 3.2, although sometimes you will need a fallback (great news for developers of reading solutions).

Major changes from EPUB 3.0.1

Given the attention to make the new version backwards compatible, there are few differences with the previous version: every valid EPUB 3.0.1 is a valid EPUB 3.2, without the need for changes, which means that the specifications add new possibilities, without removing the previous ones, at most making them deprecated.

The main changes from a technical point of view are:

- possibility to insert WOFF 2.0 and SFNT fonts (now they are Core Media Type);
- some old elements have been deprecated such as: bindings, epub:trigger, epub:switch, epub-sc, display.seq, portrait (within rendition:spread); these elements are not prohibited, their use is not recommended;
- ability to link to remote resources such as fonts and resources used by scripts, which may then be outside the EPUB package.

A detailed list of the changes can be found in the EPUB 3 Community Group[10] document.

On EPUB 3 worldwide patchily adoption

On the international scene, Japan was one of the first markets to massively adopt EPUB 3, whose specifications offer the possibility of using it to natively publish digital versions of traditional Japanese comics, such as manga, a process that was much more complex with EPUB 2.

In France too, there has been a move towards the use of the new format: the Hachette group, an important international publisher, for example, has begun to produce all its digital innovations in EPUB 3 format; the move took place after a careful study of the backward compatibility of the new format with the oldest reading devices and with the aggregation and distribution solutions that until then were based on EPUB 2, as well as with the DRM applications available on the market.

But the global transition to the EPUB3 format currently shows a gap in several countries, such as Italy, South America, etc., where most of the new features are still published in the EPUB 2 version whose specifications, from June 2014, have been declared by the W3C as obsolete and will therefore no longer be maintained, with all the consequences of the case. The transition from EPUB 2 to EPUB 3 should not create particular problems in the management of production processes: most of the authoring softwares fully manage EPUB 3, presenting this default saving option. The latest version of Adobe Indesign allows to create files in a few clicks.

Evidently the slow adoption of EPUB 3 is not only due to technical problems, but more to the fact that, for some types of very popular publications (for example fiction and non-fiction), the EPUB 2 format is good enough; this is Dave Cramer's thesis, co-chair of the EPUB 3 Community Group of the W3C, who at the beginning of 2018 published the article "Good Enough: A Meditation on the Past, Present and Future of EPUB"[11] in which he says:

---

10 https://w3c.github.io/publ-epub-revision/epub32/spec/epub-changes.html#sec-epub32-a11y

11 http://epubsecrets.com/good-enough-a-meditation-on-the-past-present-and-future-of-epub.php

> *We work in publishing because we love books. We work with ebooks because we're both idealists and gluttons for punishment. We are full of frustration with the present and hope for the future. We want things to be better; we want change; we need change. But does anyone else want change? […] After more than six years of EPUB 3, EPUB 2 is alive and well. Even my employer, one of the largest publishers in the world, makes EPUB 3s which are as close to EPUB 2 as possible.*

The transition to EPUB 3 is becoming a global concern: in August 2018, four personalities from the world of digital publishing published an article in which they urged the book industry to abandon EPUB 2. The article "EPUB 2 sunset"[12] written by Ric Wright (Director of Engineering at Readium Foundation), Luc Audrain (Head of Digitalization Support at Hachette Livre, France), Dave Cramer (Senior Digital Publishing Technology Specialist at Hachette Book Group, USA) and George Kerscher (Chief Innovations Officer at DAISY Consortium) says that a general effort is needed to get the whole industry to adopt EPUB 3. This is not a request by technology fanatics who want to adopt the latest standard at all costs: the EPUB 3 is much better, compared to the EPUB 2, for everything concerning the accessibility of content by people with visual disabilities, but also for the possibilities offered by CSS 3 to make publications richer graphically and the possibility of including mathematical formulas in the text.

EPUBcheck

A fundamental element for the world that gravitates around the EPUB is the EPUBcheck. This open-source software is responsible for checking the validity of digital publications: each element of the EPUB file is checked to ensure that it follows all the specifications dictated by the format.

The software is very important for the publishing chain: many publishers have incorporated it into their production flows, digital distributors use it to check the quality of the files under management, large digital stores use it to assess the quality of files and discard those that could create problems during reading. It is therefore a fundamental element in the adoption of the EPUB 3 worldwide.

The software needs an update to validate the latest versions of the EPUB, the 3.1 and soon the 3.2, as well as code maintenance work to make it more powerful.

With the unification of the IDPF and the W3C, the Publishing Business Group undertook the maintainance and supports the EPUB standard. To this end the group has recently launched a major upgrade project for the EPUBCheck to ensure that publications created and distributed by all publishers meet the EPUB specifications and are therefore interoperable.

The work has been entrusted, after a selection competition, to the DAISY Consortium, an international body specialized in accessible digital publishing, which over the years has developed several production and control tools for digital publications. The W3C has advanced $150,000 for this work and launched an international fundraising campaign to raise the full amount through sponsorship.

Quickly updating the EPUBcheck to align it with the latest specifications of the EPUB format is essential to speed up the transition to the EPUB 3.

[12] https://github.com/readium/readium.github.io/wiki/EPUB-2-sunset

WEB PUBLICATIONS

The Web Publications are the result of the union between IDPF and W3C: it is a new standard, currently under development, for the creation of publications usable both online and offline. This is an important step for digital publishing: the specifications are completely new (even if they are written by the same community that gravitates around the EPUB) and most likely they will not be backwards compatible with the EPUB 3.

The group in charge of defining the specifications within the W3C is the Publishing Working Group led by Tzviya Siegman (Information Standards Lead at John Wiley and Sons) together with Garth Conboy (Engineering manager for Play Books Clients at Google); the aim of the group is to publish the final specifications of the Web Publications in 2020, meanwhile Working Drafts have already been published and are available on the W3C website[13].

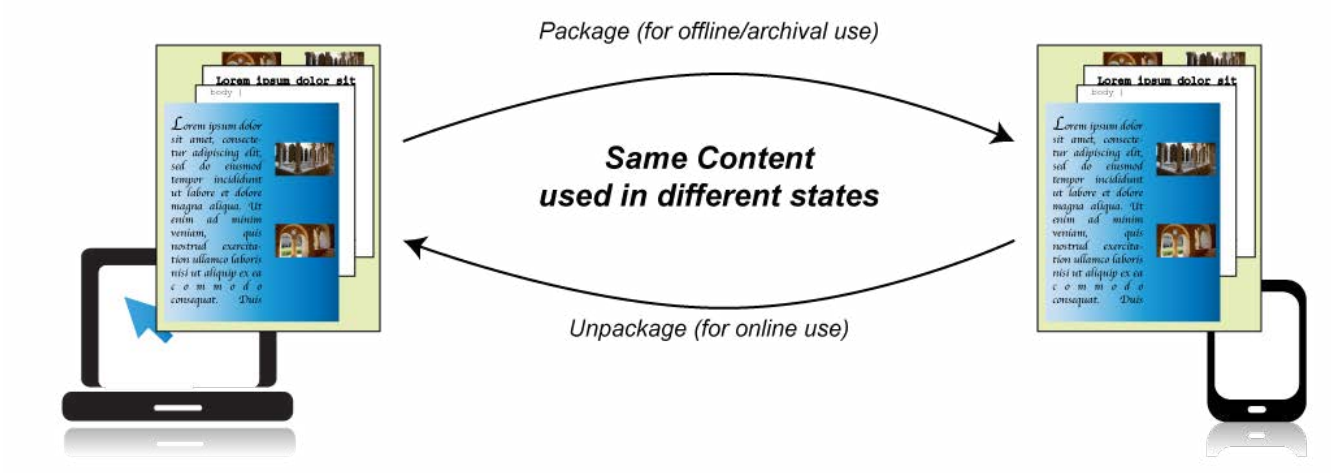

Figure 2 A scheme that summarizes the operation of the EPUBWEB, taken from the first published document, W3C[14]

As mentioned previously, the idea of Web Publications is not recent: in 2014 Ivan Hermann (W3C) and Markus Gylling (IDPF), then belonging to two separate bodies, at the Books in Browsers conference in San Francisco, presented their vision: the EPUBWEB.

This scheme, taken from one of the first published documents related to the EPUBWEB, illustrates well the idea behind it: the same content that can be used online (through a browser), or used offline wrapped in a single file that contains all the resources needed for reading. In 2014 it was a visionary momentum, even if the technical basis was consistent: the EPUB 3 and the Web are based on the same technologies (the Open Web Platform), why not integrate the two, making the transition from one state to another very simple?

The first document, drawn up on 21 November 2014 by Gylling and Herman, began like this:

> *Our vision for EPUB-WEB is that portable documents become fully native citizens of the Open Web Platform. The current format- and workflow-level separation between offline/portable (EPUB) and online (Web) document publishing is diminished to zero. These are merely two dynamic manifestations of the same publication: content*

13 https://www.w3.org/TR/wpub/

14 https://github.com/w3c/epubweb

> *authored with online use as the primary mode can easily be saved by the user for offline reading in portable document form. Content authored primarily for use as a portable document can be put online, without any need for refactoring the content.*

After the merge of the IDPF in the W3C, the work on the specifications has intensified to define different entities to work on:

- **Web Publications**: the online representation of a publication, accessible through a browser;
- **Portable Web Publication**: the offline representation of a Web Publication, packaged in such a way as to contain all the resources necessary for the fruition of the content;
- **EPUB 4**, which will be a specific profile for Portable Web Publications.

In the following paragraphs we will analyse some technical aspects of these different specifications. Please note that the documentation is not definitive and therefore the elements reported may be subject to change, so we recommend to always check the latest versions available on the site of the W3C.[15]

Technicalities of Web Publications

In the world of the web there is no concept of publication: normally we speak of individual web pages, hosted on a website. A Web Publication, abbreviated as WP, is defined as "a discoverable and identifiable collection of resources"[16] or a set of elements organized by a poster (in JSON-LD format), which contains all the information related to the publication: the order in which the elements are read, the metadata (bibliographic and technical) and the list of all the resources that are necessary to use the content (even offline).

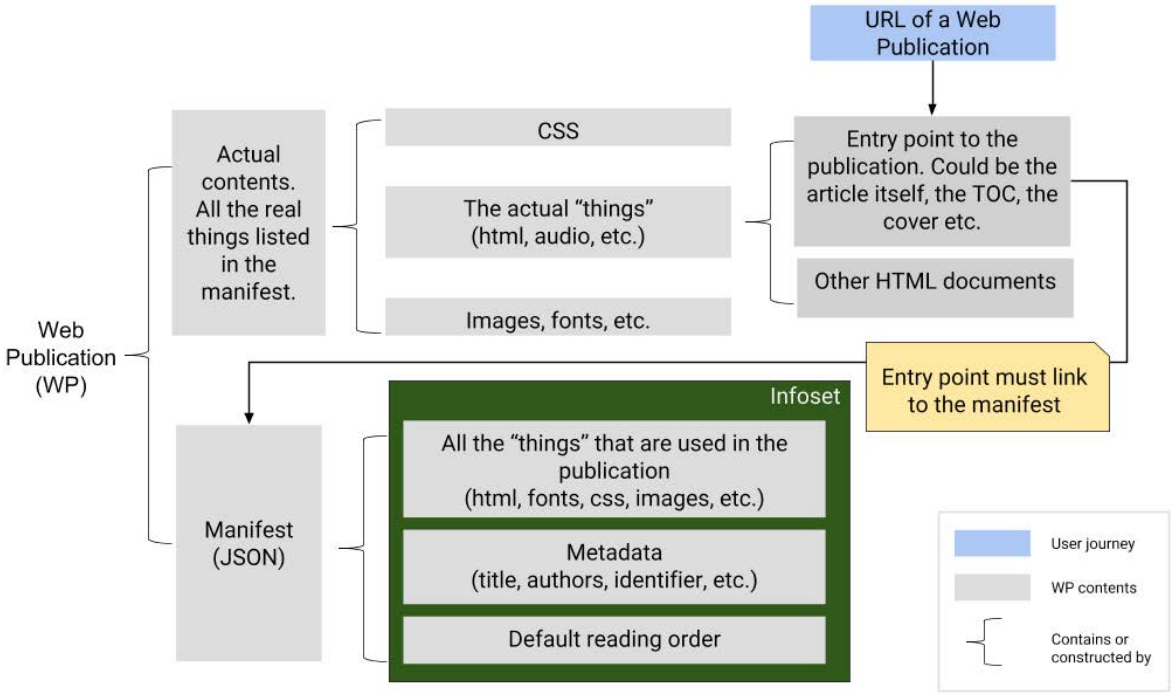

Figure 3 Simplified Diagram of the Structure of Web Publications, from Web Publications Working Draft, W3C[17]

15 https://www.w3.org/publishing/groups/publ-wg/PublStatus

16 https://www.w3.org/TR/wpub/#what-is-wp

17 https://www.w3.org/TR/wpub/#what-is-wp

The diagram describes the organization of the content and structure of a Web Publication and the path of the user, who can access the WP from any of the resources contained in it (TOC, cover, text, etc..) through a web browser. The strength of the WP lies in the fact that any user, with a normal web browser, can enjoy the content: the specifications invite the browser to develop ad hoc interfaces for reading (called "publication mode"), but even without these developments the WP will still be readable (being resources of the Open Web Plaform, then natively interpreted by the browser). The specifications hope that the reading software, called more generally User Agents in the documents (which will mainly be web browsers), develop some features typical of reading digital books in native mode, in order to offer the best possible user experience. In particular, the following functions necessary for the use of an ebook have been identified, which at the moment are not present in the most used browsers:

- **user settings** for: font size, font, display mode (night, high contrast, etc.), playback speed (for audio and video resources);
- possibility to read the content in **scrolling or paginating mode** (divided into virtual pages);
- to show the **table of contents**;
- show what **percentage of** the text has been read;
- possibility to **read the text offline**;
- possibility to **search** all the documents that make up the publication.

These features should be within the publication mode that the browser should activate when it encounters a WP.

Portable Web Publications

The specifications of Portable Web Publications, abbreviated to PWP, have not been defined ye. At the moment there is only a basic description that says that a Portable Web Publication is a Web Publication that has been packaged into a single file, so it can be saved, archived and transported regardless of he internet connection. Inside the PWP you will find the content to be displayed, the poster that identifies the metadata and the order in which the content is read, as well as all the resources that allow a correct use of the content, even offline. An interesting aspect is that a PWP can be easily unpacked and published online making it a Web Publication, even though with some technical limitations and n relation to copyright.

It should be noted that it is not necessary for PWPs to be generated from a Web Publication published online, but they can be done completely offline through production processes similar to those normally used for the creation of EPUB files by content producers.

EPUB 4

Among the various specifications which the Publishing Working Group will soon be working on, there are those related to EPUB 4. The news circulating about EPUB 4 at the moment are rumors and say that the new version of EPUB will be a profile of the Portable Web Publications: that is, a particular way of packaging resources (for example within a .zip file) that depends directly on the specifications of the PWPs. Given this direct correspondence between EPUB 4, PWP and WP, we are not sure that EPUB 4 will be backwards compatible with previous versions at the moment (and therefore with reading software, DRM, etc.).

Web Publications for audio content

The audiobook supply chain does not have a global standard for the distribution of editorial products: each aggregator and shop uses different methods for uploading audio files and related metadata (bibliographic information, but also the table of contents and the overall duration of the audio book).

The Publishing Working Group of the W3C has activated a Task Force that is defining the basic requirements of audiobooks, both in technical terms (file formats, bitrates, etc.), and in editorial terms (metadata, table of contents, etc.); the idea is to exploit the specifications of Web Publications (and PWP) to create and distribute content in the form of audio.

The distribution dilemma

After analysing the specifications of the Web Publications, the question that arises is: what will be the role of the intermediaries of digital publications?

As the specifications are oriented, most likely, the future will lead even more to the polarization between two ways of selling and enjoying digital books.

On the one hand, large web platforms that will provide content directly through a web browser (behind a paywall), a bit like other sectors of the media industry such as cinema and music.

On the other hand, large publishers (e.g. publishers of scientific journals) who will offer content to users through their own websites, for a subscription or for the purchase of a licence.

Even the specifications of Portable Web Publications seem to be oriented firstly on the B2B exchange of content (for example between publishers and sales platforms), secondly on the offline saving by the user of his own copy of the publication, so the concept of selling the file, to which we are now linked, is lost; so also the issue of DRM is overcome, offering the user access only to resources for which he has paid.

In this scenario, what will be the role of the intermediaries who are currently between content producers and large sales platforms?

THE ELEPHANT IN THE ROOM: AMAZON

After the analysis and presentation of the state of the art and the future of the standards for digital publishing, we cannot but talk about the elephant in the room: Amazon Kindle.

Amazon is the largest operator in the field of digital book sales, yet operates outside the EPUB ecosystem. Kindle uses, for the distribution and sale of ebooks, a proprietary format based initially on the format of Mobipocket (a company acquired by Amazon, which had developed a proprietary format for reading digital books), later expanded with the format KF8 which is mainly based on EPUB 3, albeit with some limitations, such as the use of interactive elements. This format can only be created using software released by Amazon and read only through Kindle applications.

It is important to note that nowadays, the best way to create a Kindle-compatible file is to start from an EPUB file and convert it; in fact, there is no professional authoring software capable of natively export files to Amazon. Amazon therefore does not use the EPUB format, but is strongly dependent on the IDPF standard, to the point that the Kindle Direct Publishing[18] guidelines have direct references to the EPUB 3 specifications and make direct use of part of the specifications, for example to indicate how to set up the ebook cover image, or the table of contents.

It should be noted that, at the moment, Amazon, despite being a member of W3C, does not participate in any working group related to digital publishing; maybe the work being done with Web Publications could awaken the ecommerce giant?

[18] https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf

FONDAZIONE LIA
LIBRI ITALIANI ACCESSIBILI

Version

Version 1.0 – 6th November 2018

Contacts

www.fondazionelia.org
https://catalogo.fondazionelia.org

Corso di Porta Romana 108,
20122 Milano
Tel. +39 02 89280808
Fax. +39 02 89280868

Gregorio Pellegrino
Chief Accessibility Officer
gregorio.pellegrino@fondazionelia.org

LIA

Born Accessible Publication
Made to be read by people with print disabilities

FONDAZIONE LIA
LIBRI ITALIANI ACCESSIBILI

Fundación Germán Sánchez Ruipérez

International Publishing Distribution Association

Actividad subvencionada por el Ministerio de Educación, Cultura y Deporte