Digital Publishing Salon

Results and Notes

September 13, 2022

TPAC 2022, Vancouver

# Introduction

The Digital Publishing Salon was held with members of the community to brainstorm and discuss the future of digital publishing. EPUB 3.3 is nearing completion and is expected to become a full W3C recommendation in Q1 2023. Once published, EPUB 3.3 will enter maintenance mode, meaning that a working group will form to address any errata or issues raised after publication.

This opens us up to an opportunity to explore what comes next.

During the salon, discussions covered a broad range of topics and concerns. In the [appendix](#_q4qzs9wwutpi) for this document, we have included the notes from some of these discussions. We have also included the notes gathered from the whiteboard during the session, you can find them in the sections on [Summarized Problem Statements](#_kryh09hd0l1a) and [Proposed Solutions and Areas of Exploration](#_rteukoxfkv0c). There were a lot of great ideas that fall outside of the technical specification realm, but can be explored in other venues, like the business group.

If you have any questions about this summary, what comes next, or about anything to do with the publishing activity, please feel free to reach out to us at public-publishing-sc@w3.org.

## Goals

Going into the salon, we wanted to focus the discussion on four areas:

* Accessibility
* Portability
* Offline-ability
* Use Web Technologies

From those constraints, we wanted to determine what areas could be brought into a new working group charter. The charter is a guiding document that provides structure to what a working group will explore. In general, the expectation of a charter is for a group to produce a technical specification.

In order to frame and organize the results of this workshop, as well as any future discussions or planning, we used the [GIST Framework](https://itamargilad.com/gist-framework/).

As part of the framework, we need to begin with goals, and we have elaborated on the goals we presented in the beginning of the salon, as well as including another one that arose during discussion:

* Accessibility > Make digital publications more accessible for all users, on all platforms
* Portability > Make digital publications portable through space and time
* Offline-ability > Make digital publications independent of internet connectivity
* Web Technologies > Continue to modernize and develop the publishing ecosystem
* Security > Make digital publications secure, verifiable, and protective of user privacy

These goals will help us guide the charter discussions, as well as what ideas we should explore as we begin those discussions.

# Summary of Discussion at the Salon

Discussion at the salon was focused on the goals stated above, as well as the imperative to “dream big” about the future of digital publications. We also want to recognize where we are coming from, and much of the discussion did cover what has been done in the past, or where we are seeing challenges in the ecosystem.

Other themes arose during conversation as well:

* No more “if we build it, they will come” as an approach to specification; the spec should respond to real users’ problems with solutions whose success can be measured with confidence
* Security as an underspecified concern for the publishing industry
* Best Practices are not standards
* Need for better tooling, communications, and skills development
* Evolving user needs, expectations, and interactions

## Summarized Problem Statements

These statements were gathered after the first breakout session, and have been grouped into 6 main topic areas, which were developed after receiving all of the feedback. The first session focused on the problems or challenges participants faced with digital publications in the market today.

### Designing Complex and Inclusive Content

* Challenge of designing for digital as its own form of reading, the design of the content and user interface, its affordances
* Complex content -- Sheet music display, mathml, interactive tables, etc.
* As a publisher, I want to produce content with a rich layout that communicates information from Spatial relationships.
* Pressure to have accessible documents by 2025 (European standards)
* Gaps between designing for print vs designing for digital -- author vs. user control
* Horizontal spatial relationships are difficult to communicate (difficult for reflowable content)
* movement away from epubs -- publications that are between epubs and websites; divergence from standards
* As a publisher, I also want spatial relationships to be accessible, and the communication of the Spatial relationships is something I cannot yet accomplish.
* Accessibility has fallen through the cracks
* As an end-user, I want product manuals and corporate documents to use modern digital publishing standards that deliver accessibility and navigability.

### Community Engagement, Cross-Industry Communication

* Need more librarians participating in these discussions
* Lower technical barriers to entry, not relying on solo tech evangelists in each organization
* Difficulty of getting standards organizations talking with industry and vice-versa
* No technical knowledge in-house
* Encourage more young people, new generations into this work
* Divergence of standards across industry (LTI, Caliper, etc.)
* Getting end-user participation

### Digital First Practices

* Digital books that can serve multiple versions from same file, depending on context
* Serialization, delivering content as multiple pieces within a single publication
* Archiving ebooks for long-term preservation
* Books that can respond to ereading environment and present as either reflowable or fixed-layout
* Books that adapt themselves easily to the device/mode they are in
* Updating books as delta updates instead of the whole book
* Interoperability
* No software or downloads required
* Versioning standards; will be hard to communicate updates to readers whether they have latest versions

### Tools & Production

* Remediation of older formats into new formats
* Move large amounts of backlist into new, accessible formats that are long-living
* Better shared best practices and tools to validate against best practices
* Better tools for building EPUBs
* Interactivity in educational publications
* Managing large file sizes
* No good tools for authoring ebooks
* Difficult to author content as ebooks
* Use the open web to deliver content that is interactive, can be changed, and is accessible.

### Addressability, Discoverability, Data

* Potential of correlating cross-media, using more robust metadata to connect all the forms media might take (books, movies, VR, metaverse)
* Data consistency, standardizing data to be useful and consistent
* How to identify a publication and communicate the content and functionality to the user
* Addressability -- being able to speak precisely about the content (refer to a specific place in the book)
* Prevent link rot -- decentralized ids and redundant systems to keep links live during life of an EPUB
* Better search results sorted by hierarchy of the results in the text
* Primary vs secondary metadata, primary generated by the publishers, secondary generated by users (learning objectives, trigger warnings, genre, etc.)
* Privacy and user identity (especially in combination with evolving rental model)
* public vs private annotations
* clunky experience of annotations

### Rights and Distribution

* Shift from purchase to rental models
* Rights and inability to sell content worldwide simultaneously, differences of logistics in digital vs print
* ebooks have to be purchased and repurchased; costly for libraries

## Proposed Solutions and Areas of Exploration

These statements were gathered after the second breakout session, where we asked participants to think of solutions or areas of exploration for the 6 topics we developed after the first session.

### Designing Complex and Inclusive Content

* Challenges facing publishers and organizations about creating content and in reader’s hands
* Importance of designing for mobile devices (e.g., economies where phones are the primary means of connection/communication)
* Expertise of HTML authoring is going away
* Previous attempts to bring books to browsers failed due to lack of industry (namely publisher) buy-in
* Modern technology industry that is focused on differing publication standards (markdown vs. markup)
* Possibilities of EPUB vs limitations of reading systems
* Solution: involving the browser as a host for reading systems

### Community Engagement, Cross-industry Communication

* Focus on increasing knowledge and engagement (e.g., lunch and learns?) about what’s going on in digital publishing
* Educational programs, outreach to train next generation of publishing professionals (e.g., in schools that host publishing programs) -- supporting digital & a11y curriculums
* More engagement with tech (why aren’t they participating in our community?)
* Reality in our industry is that new people are not being paid what they deserve to do this kind of work
* Need engagement with other W3C groups. How can we make that easier? Increase knowledge of pain points across groups.
* Involve other players in publishing (platforms like wattpad, LMSs, etc.)

### Tools & Production

* Brainstorm on book authoring: generation of EPUBs
* Divided workspace, multiple authoring environments
* EPUB generation is secondary in most prevalent tools (e.g., InDesign)
* Complicated native formats in authoring tools; options are usually import/export towards lowest common denominator
* Analyze the gaps in reading systems, identify what’s actually implementable
* Developing specification around linking FXL content to reflow presentation
* Generation of example EPUB content to “show off” features may be helpful
* Extend the specification
* Simplification of complex content (adaptive to context or platform)
* Tools to bridge existing tool chains, InDesign to other tools
* Technical expertise, HTML/XML first workflows, etc.

### Addressability, Discoverability, Data

* Locators have been solved (EPUB CFI)
* Web Annotations
* Testing existing solutions
* Address versioning directly in file metadata (signatures); convey differences between versions within the publication
* Archiving materials - workshop proposal from 2019 (Tzviya, Deborah, Ivan) -- https://www.w3.org/2019/archival-workshop
* Version control of files and getting people to implement depends on annotations (due to loss of anchors)
* Signature.xml and digital signing
* Lean more on fuzzy matching (e.g., to solve annotation problems)
* File System API, resource classes for links

# Conclusion and Takeaways

After the discussion, we reviewed the suggested solutions and areas of exploration to determine how they might fit into the publishing activity and future planning. One of the challenges of this exercise is determining what can be done within the publishing activity, and what requires broader participation. In addition, some areas discussed were not topics that we felt fit into technical standards. In our Takeaways section, we have organized the ideas into what can be standardized within a working group, what is out of scope, where we can collaborate with other web standards groups, and what we should explore within the publishing activity.

## Takeaways

### Areas to standardize:

* Digital books that can serve multiple modalities from same file, depending on context
	+ I.e. Media Queries, different stylesheets for fxl vs reflow
* Serialization, delivering content as multiple pieces within a single publication
	+ Chapter level metadata, purchase the package, it updates as content is released
	+ Updating content standards
* Different levels of metadata
	+ Global, title level
	+ Chapter/item level
	+ Secondary streams (user generated, social media, etc.)
* Digital Signing and Content Verification
	+ Use existing web signing methods (HTTPS, TLS), verify the source and ownership

### Out of Scope:

* Tooling, simplifying production process may make tooling easier (see Audiobooks implementation)
* Remediation of older formats
* Rights
* Blockchain/NFTs
* DRM

### Areas to collaborate within W3C:

* Annotations
* Addressability
* Archiving
* Accessibility

### Areas to explore within the Publishing Activity:

* Divergence of Publishing Standards
* Communications about standards
* Privacy and security education for Publishers
* End user participation
* Rights

## What Comes Next

Now that we have areas to explore and proceed from, what happens next?

The Publishing Activity will take the information we have and work with our goals to determine what work items can be derived from them and form a charter. This will include looking at use cases, current implementations, gap analyses, and discussions within the various members of the publishing activity.

Before we proceed to the charter writing phase, we want to be confident with our goals and findings. We plan to keep communication about this process in the open, so please keep in touch and we will do our best to update the community on our efforts.

Thanks to everyone who participated, we appreciate your time and commitment to open discussion and publishing standards. If there are any questions or comments on the session or this report, please reach out to us at public-publishing-sc@w3.org.

# Appendix

## Breakout #1

### Collected Breakout Group Notes:

Breakout #1 Digital Publishing Salon

9/13/22

Goldman Sachs publishes in PDF, consumers not use to tablets kindles, mostly use phones and blackberries.

Big 10 Public Universities / Libraries. - Founding member of HathiTrust – Google Books scanning project. ½ came from big ten.

U of Mich. Directory of IT built HathiTrust. Engaged with problem with books published in physical to digital.

How do you preserve books digital? EPUBs come to HathiTrust, the only way Library knows how to preserve that is to strip it down to XML and it all gets destroyed to preserve the raw content for the future. We can’t do it with EPUB. The Standards are not there to preserve that container. Best we can do is to buy a physical copy and to save a physical copy. How do we engage with longterm preservation.

3 pillars accessibility / offlineability & portability in space/time

DRM – making users lives worse. Inability to reference books. Only link to book on the store, and ecosystems don’t talk to each other.

I can give you name / title if you are in a different ecosystem, you may not be able to find the exact same book in your ecosystem. I can’t tell you the book let alone the chapter / page.

Addressability is the main issue here. ISBN does not work DOI does not work, how about versioning.

Finding page numbers in your are in a classroom that hits all the criteria, simple an clear, but ISBN/DOI not simple or clear.

Company A has an educational platform ane want to specify book A. I can do this with our classroom experience, but there may be ways to get Company B’s books not very easy to get.

For example Google Classroom may need to put engineering efforts to work with another group to support their books but it would be a one off.

Proprietary solutions don’t work.

Single Use reading device, not sustainable.

4.7 Billion people have access to internet.

EPUB only really accessible to mostly western civilization with $$

The barriers outside the W3C. technology exists the problem is a financial / technology access.

Are there technology obstacles we can address?

Single use industry, if we move beyond binaries fixed layout ; reflowable. Responsive books no matter what the screen, platform, device etc.

Manga is an issue which is not responsive.

Build in responsiveness.

Is this a book specific question? Yes.

Any website should work the same.

We should remove those constraints.

Sections of the code base were developed to a trajectory of the EPUB. Once we go past the single use reading environment. We are going to move to a Browser. How do you build a book that

Responsive as well as accessible, and argues that a book that isn’t accessible is NOT a book.

Western context is a linear concept. Story telling how they convey their knowledge / history not in the way books are represented digitally. They want to show/ tell multimedia. Bringing in other cultures changes the game.

Visually recorded the stories, / art, fully responsive accessible you can see the land, hear their voices which were preserved. If you write a story down that is a lesser way to carry the knowledge on you must hear that within the book.

Communication must be preformed / encountered, listened to , heard, seen.

Are there are kinds of story telling the fall into books? Interactive ways, seeing comics more the just storybooks.

MAUS book.

Another example some cultures telling a story is only complete if you hear it text that you speak vs. a song that you sing, if you song/chant which we concentrate the text or the music books on a song, but the two together that counts. Should be part of publishing which isn’t there.

Manga for instance, being able to follow the comic strip down into the panels that may be streaming that video within the panel, an if you turn the phone sideways it goes into landscape.

Wiki page vs. book EPUB vs. Website is a profound difference.

We did have a Publishing on the Web CG. Security was a concern. Non of the players / device manufactures were outside of their interest. Was more of a dream.

Two Italian publishers EPUB specs are fine for them, they find help making comics, no spec, everyone does their own way. Industry is not educated for digital books in Italy. Authoring tools are out of sync with the specs.

For a11y remediation of backlists. EPUB tools to make them a11y EPUB3. Huge for the Accessibility European Act in 20225

Born Accessible Authoring Tools.

Digital Publishing has held the same place of movie days, theater, Digital book as the primary tool of the future, this is how billions will read in the future. Spent 25,000 for the print book and only 500 towards the digital book but when this switches this will turn the tide.

Challenges:

* Things coming up, need to solve
* Paul: Not a surprise: K12 client, 2 big issues
	+ How to present graphical content, geared to pre-K early-K, with words. On a digital/mobile format, with accessible components.
* John: ever increasing file sizes of EPUBs, not a problem in developed countries with good infra scale, but connectivity and bandwidth is an issue for the rest of the world. Make it work on 5yo+ devices. Graceful degradation
* Paul: Higher ed space, integrating into courseware wanting to have assessment items synchronized to gradebooks, data integration.
* Liisa: For trade, standard methods for interactivity or data sync is still a real limitation for many kinds of content, fill-in, self-help books, children’s titles.
	+ Ever increasing filesize, because its compressed, still has limitations on what reading systems can support, image limits, HTML size issues.
	+ John: **Above 250MB starts being problematic**, less practical, imagine older devices handling large files.
* John: The PDF ask: To reference page numbers, academics want this. Consistent to print material, hybrid teaching materials. Not including page anchors, surprising number
* Juan: Problem: Getting new software developers to join the industry, and deal with the esotericism of the industry, XML over HTML as an example.
	+ Take notes from developer advocate programs by big tech. Help devs be comfortable to progress through the domain specifics and cut through to the core web tech that is familiar and flexible.
	+ For me as a dev/software company, what do I best invest my time in? My personal anecdote, I spent quite a bit of time on CFIs early on and it was interesting to learn about its shortcomings from the other sides.
* **Preflight checks:** EPUB check++, (Ace checker)
	+ Lighthouse for EPUB, this content is AA+ grade not just for accessibility for but for other concerns like the ones discussed so far
	+ Shared understanding of best practices.
	+ Checks/guidances that are targeting the use cases
* Kim: PDF to EPUB, having speech technology, accessible content. Cognitive load for users to enable/seek out different things.
	+ Having page numbers consistent for instance, all the little things that help add up to remove fragmentation. **To reduce fragmentation & encourage a little increase in consistency.**
* Paul, Theme here: better education on what the best practices are, what are the expectations, share knowledge, build up shared understanding, make issues visible. “We never knew, no one really asked us to do this”
	+ Shared navigational concern: Better environment for reading systems.
	+ Know where you are headed towards the standard.
	+ Info is out there: Gathered, vetted, made accessible
		- **Locked up in the brain of some veteran in the industry**
		- **Vetted modules you could grab and plug in, not exactly a template, but like a software package/repo.**
	+ Kim: Allow users to affect this and share. See how users organize things, allow them to do this. **Allow the teacher to split the book or the package.** Talking about the packaging here, “Here’s a pile of books, they are next to each other”
	+ John: **Allow to sell the individual chapters, but allow the package to be visible too.**
* Liisa: Scale up or down the EPUB package. This has volumes, you paginate this this way, or this smaller chunk is treated like this. Industry wide taxonomy to tag those chunks.
* Liisa: Big problem: Reading systems were limited, coding was limited, small file size, very narrow box to develop with originally. From the readers perspective they have no sense about this box, limitations, or the variables. “Why can’t we have this, why can’t you do X,Y,Z?” We are still dancing around issues, “Kindle for iOS bug” case bug; SVG path on a curve.
* Liisa: Another struggle, update things to make things more accessible.
	+ No ability to seamlessly update a piece of content, and allow the reader to track the change of their content. Versions. **Seamless upgrading of content for both the industry and the user.**
	+ **I got this version of the book.**
	+ John: Can make annotations on the content, content updates, annotations go offset. Gracefully move towards the latest version.
	+ Allow users to see what’s been updated, what’s different, and let the user tell what they think is broken. Get them excited about the updates. It’s a pain to figure out what happened.

Authoring

- Authoring pipeline continues to be an issue

- Lowest-common denominator approach to content creation (ie, no widgets/special CSS/etc.) is safest due to poor support levels

Relationship with print

- Because print is still 80% of the market, the authoring pipeline is neglected (Accessible books consortium working with Adobe to improve EPUB export from InDesign)

- User overrides not familiar to publishing - authors used to having total control of print output. Striking right balance between user and author control is a hard problem

Education-specific

- In education/academic - industry is moving from purchase to rental model, associated privacy issues

- In education/academic - PDF still widely used, standards like LTI are bypassing EPUB

- In education/academic - creation and management of derivative products is an issue - lack of a work identifier makes this challenging

Standards & education

- Difficult to get broad involvement with and adoption of standards across industry (this is an issue with both very large and small players)

- Many publishers outsource technical expertise, so knowledge is missing in-house

Portability

- Portability of files between platforms an idealistic early goal which was never achieved. Difference between goals of technical folks and their employers

- Legal interoperability mandates may change this in the long term

-The need for more librarians and scholarly publishing staff involved in these discussions.

-The need for a better landscape of tools and resources around EPUB that would allow:

--Lay people to open files with ease in browser with no additional knowledge or software downloads needed.

--Tools to facilitate EPUB development so publishers can create better EPUBs with a lower technical barrier to entry, so publishers don't need to compete with other higher-dollar industries for engineers and technical staff and the industry can be self-sustaining without relying on solo digital evangelists in each org.

--Better interaction with tools that need to interact with digital books, ie: learning management systems

--EPUBs that can serve multiple versions from a single file, ie: a fixed layout or reflowable EPUB from the same file dependent on context.

--BETTER SEARCH results weighted based on hierarchy - ie: a result found in a heading is ranked higher than results found elsewhere.

We also discussed sheet music display, project MUSE's work converting EPUB to WEB with custom CSS, MathML, Hooks and sharing, interactive tables, and living books that can be updated in deltas vs. sending a full file.

Problem statement: I would like to use the full OWP stack - thus being interoperable and independent of any OS, in contrast to a file that is distributed (EPUB), the contents stay under control of the publisher who will be able to control and update, and add new features ETC.

Problem statement: As an end-user, I want product manuals and corporate documents to use modern digital publishing standards that deliver accessibility and navigability.

Problem statement: As a publisher, I want to produce content with a rich layout that communicates information from Spatial relationships. I also want this to be accessible, and the communication of the Spatial relationships is something I cannot yet accomplish.