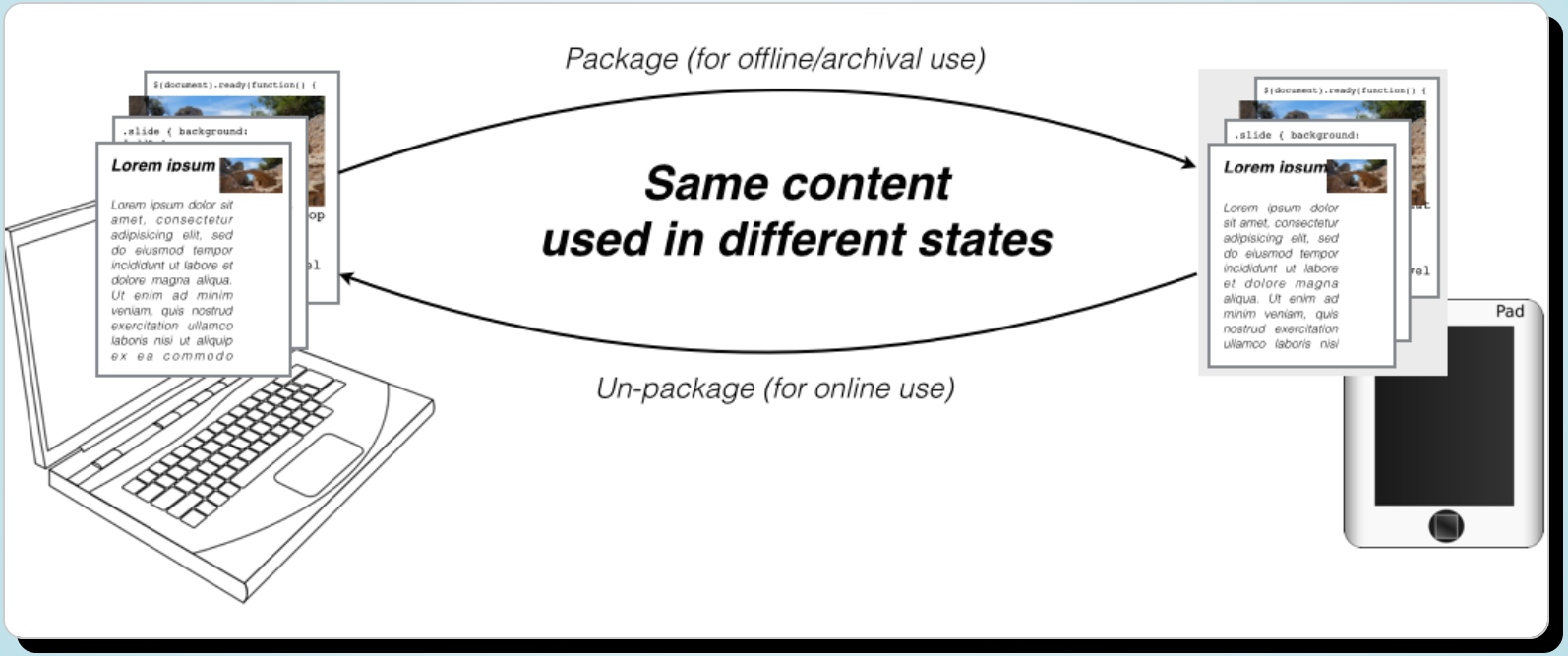
BRIDGING THE WEB AND DIGITAL PUBLISHING: EPUB-WEB

Markus Gylling, IDPF, and Ivan Herman, W3C

2014-11-26

EPUB-WEB IS A VISION FOR THE FUTURE



THE VISION

- Portable documents are fully native citizen of the Open Web Platform
- Separation between online (i.e., the “Web”) and portable (i.e., “EPUB”) is diminished to zero
- This means:
  - Content authored for primarily offline use can be used online by loading it into a browser
  - Content authored for primarily online use can be easily saved as a portable document for offline use
  - These should be doable smoothly, solely based on the user’s interaction

THE VISION

- Publishers can choose to utilize either or both of these publishing modes
- Users can choose either or both of these consumption modes
- Essential features flow seamlessly between on-line and off-line modes, like
  - cross-references, user annotations, access to on-line databases
  - licensing and rights management
  - etc.

WHY BOTHER?

EPUB AT AN INFLECTION POINT

- EPUB 3 is getting adopted by publishers
- But rendering EPUB 3 requires:
  - browser extensions; or
  - significant development on top of “browser cores” for software/hardware solutions
- EPUB as “first class Web citizen” would mean an easier and quicker deployment

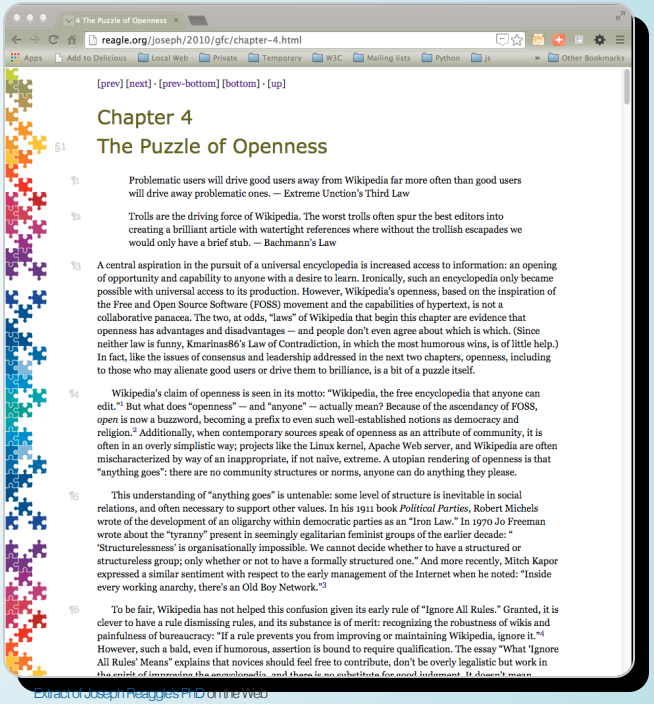
WEB PLATFORM AT AN INFLECTION POINT

- Usage of (mobile) apps come to the fore (although often based on the same “browser cores”)
- EPUB-WEB development would go in parallel with the development of general issues like packaging, metadata, or offline support

A NUMBER OF USE CASES…

FOR EXAMPLE: BOOK IN A BROWSER

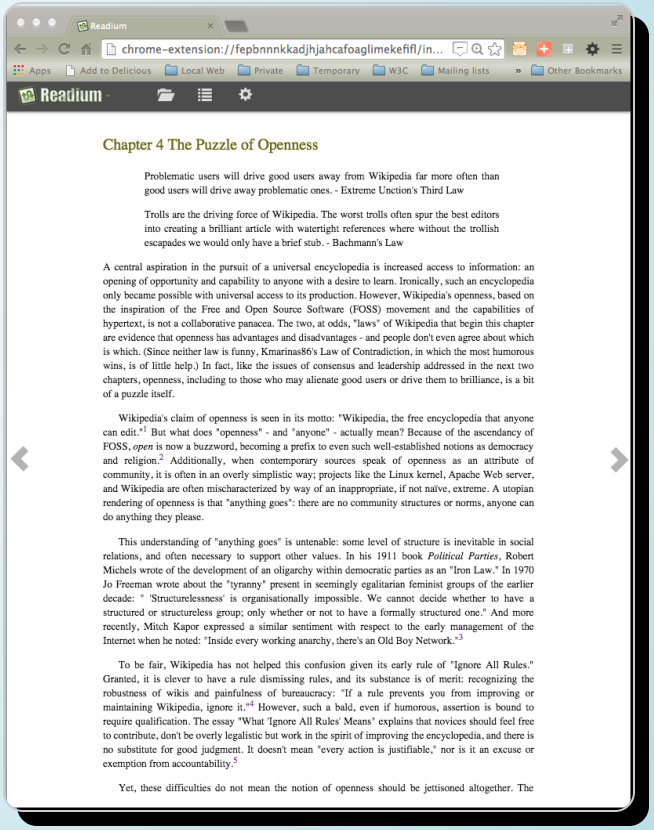
- On a desktop I may want to read a book just like a Web page:
  - easily follow a link “out” of the book
  - create bookmarks “into” a page in a book
  - use useful plugins and tools that my browser may have



Extract of Joseph Reagle's PhD on the Web

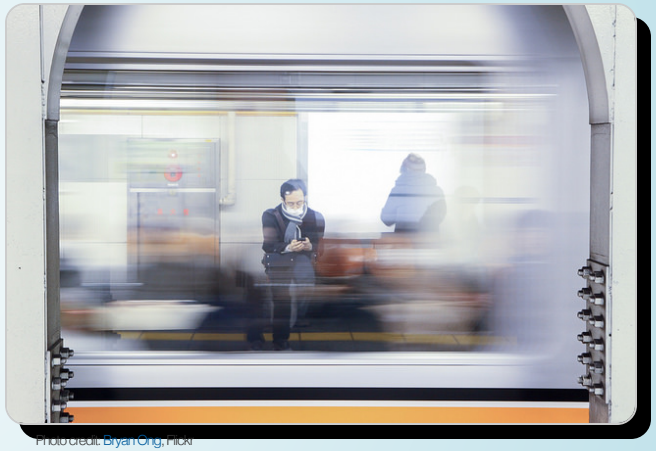
FOR EXAMPLE: BOOK IN A BROWSER (CONT.)

- But:
  - my book may be 2000 pages long
  - conventional Web browsing may not be the right way to view content, a paginated view may be better
  - I may also want to use a small dedicated reader device to read the book on the beach…
- All this should happen using *the same* book, and not a conversion from one format to the other!



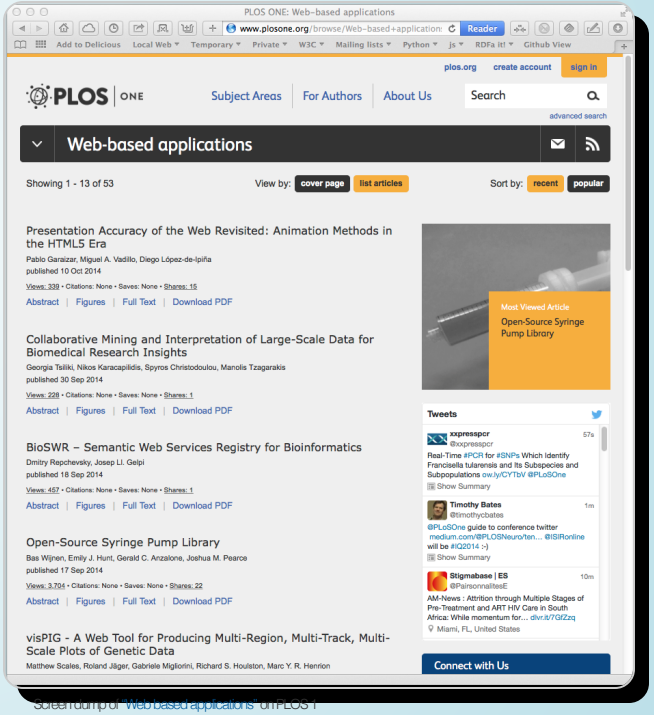
FOR EXAMPLE: I MAY NOT BE ONLINE…

- I may find an article on the Web that I want to review, annotate, etc., while commuting home on a train
- I want the results of the annotations to be back online, when I am back on the Internet
- Note: some browsers have an "archiving" possibility, but they are not interoperable
  - the content can definitely no be read on a dedicated reader



FOR EXAMPLE: SCHOLARLY PUBLISHING

- My paper is published, primarily, on-line, but people may want to download it for offline use
- The format of the paper should be adaptable to my reading environment
  - do not want a two column, fixed layout file that I cannot handle on my iPad…
- My "paper" may also contain video, audio, data, programs…
  - scholarly publishing is not text only any more!



Screendump of "Web based applications" on PLOS ONE

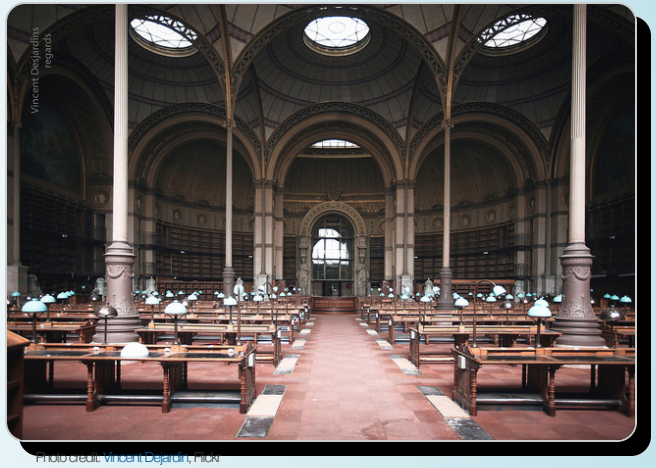
FOR EXAMPLE: IN-HOUSE PUBLISHING

- Major companies (IBM, Intel, Boeing, FAO, Renault,…) are specialized publishers through the publication of huge amount of documentation
- Delivering it on paper is not an option any more
- Fast refresh time is needed
- The *same* document should be available offline (e.g., in the cockpit) or online (e.g., on the work floor): there should be no difference between the two

Photo credit: Pedro Agulla, Flickr

FOR EXAMPLE: ARCHIVAL AND PRESERVATION

- Archiving digital assets (i.e., Web pages with *all* dependencies) is a major problem
- There is a need to produce, easily, a *complete* version of a page to be stored through archival facilities



FOR EXAMPLE: EDUCATIONAL MATERIALS

- What is an educational publication?
  - *A book* of possibly long texts that requires offline access on dedicated devices?
  - *A packaged application* with built-in interactive tests, animated examples?
  - *A Web client* reaching out to Web services for assessing test results, to encyclopedia, …?
  - *An interactive data container* storing various data for, e.g., demonstrations?
- The borderline between a “book” and a “(Web) Application” are becoming blurred!

SYNERGY EFFECTS OF CONVERGENCE

ADVANTAGE FOR PUBLISHERS' COMMUNITY

- Publishers want to concentrate on what *they* know better: how to produce, edit, curate, etc, great content
- Publishers are *not* technology companies, nor do they intend to be; they want instead to rely on the vibrant Web community!

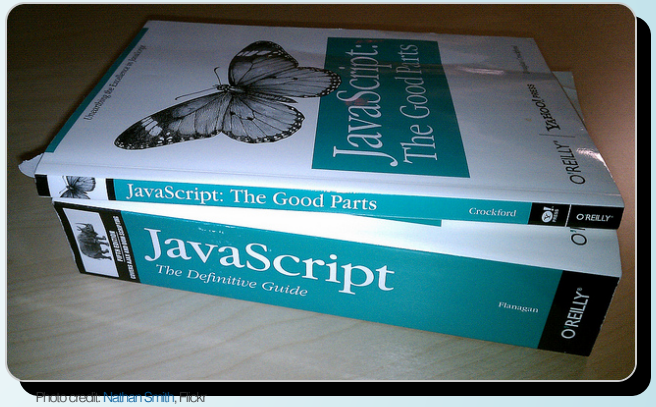


Photo Credit: Nathan Smith, Flickr

ADVANTAGE FOR THE WEB COMMUNITY

- Publishers have a long experience in ergonomy, typography, paging, complex layout, etc.
- Publishing long texts, with the right aesthetics, readability, structure, etc., is an expertise the Web community can profit from
- Experience of publishers in the complete curatory workflow for producing content may become important for Web design

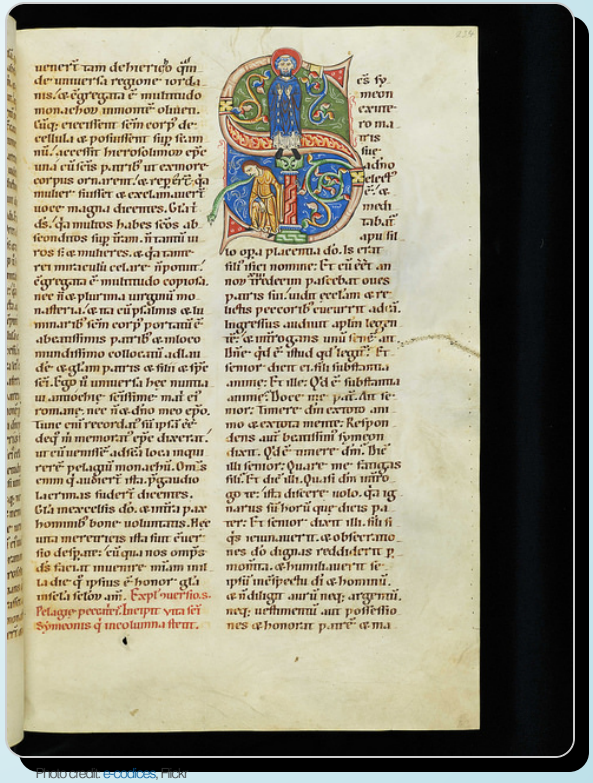


Photo credit: e-codices, Flickr

SOME COMMUNITIES THAT MAY BE AFFECTED

(Trade) Publishers
Browser vendors
Governmental bodies
Consumers of ebooks
Web designers
Web Developers

STM Publishers
Large companies
International institutions
Scholarly authors
Archivists
Publishing workflows

HOW DO WE GET THERE? (TECHNICALLY)

HOW DO WE GET THERE?

- A strong cooperation between the different communities should be ensured
- Technical challenges must be identified
- A new generation of EPUB (“EPUB-WEB”) has to be specified

In what follows *some* of the main technical issues will be highlighted

ARCHIVAL FORMAT

- EPUB is based on ZIP
- There is no standard packaging format for browsers yet…
  - although there is a need for, e.g., applications or data sets
- … but ZIP may not be the right approach on the Web
  - Multipart Mime may be an alternative
- There is a new work item at W3C on packaging standard whicht may affect EPUB-WEB

OVERALL DOCUMENT STRUCTURE

- A complete, offline content needs additional information
  - list of all necessary content, default reading order, etc.
  - in EPUB these are stored in additional, auxilliary files (“spine”, etc.)
  - these may have to be adapted to EPUB-WEB (e.g., usage of JSON instead of XML)
- But these data may not be necessary for a simple Web page with a few CSS files
  - i.e., some sort of a default structure should be defined
- User interaction paradigms should also be developed to create documents from more complex Web sites easily

IDENTIFICATION

- A consise and unique identification for the *document* (e.g., the "work") under various usage patterns is necessary
  - i.e., what is the URI for…
    - Shakespeare's Hamlet?
    - its digital edition published by Publisher XYZ?
    - the copy I own and annotate?
- This is necessary to make a book a first-class citizen on the Web

B.t.w., this is already the topic for debates in the publishing and library communities…

IDENTIFICATION (CONT.)

- Unique identification of the work is not enough
- A fragment identification framework is also necessary to link *into* the document
- There are fragments defined for various media, but a universal model, workable for browsers, is still missing
  - these should be agnostic to offline vs. online state, to media type, etc.

Works are already ongoing within the framework of activities around annotations

METADATA

- EPUB-WEB must make use of browser friendly metadata formats (RDFa, JSON-LD, μdata) to carry metadata
- EPUB-WEB should also provide means to link to external metadata
- A typical area where a solution for identification (including fragment identification) is essential

IMPROVEMENT ON STYLING

- Books usually need higher quality typesetting than average Web pages
  - control over fine aspects of drop caps, alignments, justification, hyphenation, bleeds, etc.
  - cf. http://www.w3.org/TR/dpub-latinreq/

PAGINATION, NUMBERING, INDEXING

- What is a *"page"* for an electronic content?
  - is this a new CSS concept? Do we need an extension to the DOM?
- In general, content may be placed into linked *regions*
- References (page numbers, references to pictures or diagrams, bibliographic references, etc.) should have robust mechanism across several documents/regions
  - e.g., if the book's chapters are stored in different files
  - page numbers can dynamically change as a result of user interaction, e.g., larger fonts
  - numbering of items should be automatically continuous across documents

SECURITY AND PRIVACY

- The Web/browser mechanism is based on the control per “site”, or a ”web page”, largely based on user/password
- Current publishing practices are very divergent (strict DRM, watermarking, social DRM, etc.)
- A consensus solution will be needed

PRESENTATION CONTROL

- What is the level of user control of the presentation?
- The Web and eBook traditions are vastly different:
  - In a browser, the Web designer is in full control
    - CSS alternate style sheets are hardly in use
    - some user interface aspects can be controlled but only for the browser as a whole
  - In an eBook reader, there may be more user control
    - foreground/background color
    - choice of fonts
- There is a need to reconcile these traditions

HOW DO WE GET THERE? (PRACTICALLY)

WHAT IS NEXT?

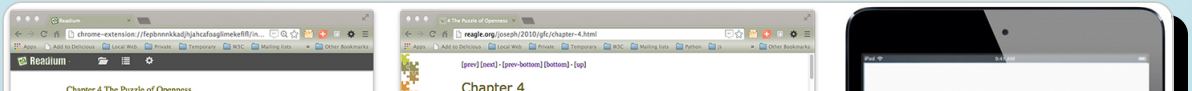
- We collect comments on this vision
  - collecting (public or private comments) to the White Paper
  - discussions at various events with the communities at large
  - internal discussions at IDPF and W3C
- Develop a more specific roadmap
  - what are the specific technical specifications that have to be developed?
  - what other work should be “influenced”?
  - what groups should be set up at IDPF and/or W3C?
- Regular updates of the White Paper based on the comments

IF THERE IS A CONSENSUS

- Work with existing IDPF and W3C groups, where necessary, on specific details
- Set up a new group (or groups) to define the EPUB-WEB specific issues

CONCLUSION

- There is a great potential in a convergence between the Open Web Platform and Portable Documents
- It will require a common effort and cooperation of both communities
- But it is an exciting prospect!



SOME POINTERS

White paper:

http://w3c.github.io/epubweb/

Issue list:

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

These slides:

http://w3c.github.io/epubweb/presentation

Direct contacts:

Markus Gylling, IDPF
mgylling@idpf.org
Ivan Herman, W3C
ivan@w3.org

THANK YOU FOR YOUR ATTENTION