iPRES2018 - September 24 - 27, 2018 - Boston & Cambridge, MA, USA

Formats (paper presentations and panel)

Session Chair: Bonnie Gordon
Speakers: Maureen Pennock, Juha Lehtonen, Lisa Sisco, Duff Johnson, Frederic Grevin, Yan Han, Leonard Rosenthol

27 SEPTEMBER 2018 / 9:00 - 10:30 / PECHET ROOM 1

LINKS TO PRESENTATIONS / PAPER

Session organizers! If you have any materials you would like to link to, you can do so here.

Additionally, this is the place for proceedings entry, once uploaded to OSF.

NOTE TAKERS

Please add your names here: Erwin Verbruggen, Micky Lindlar, Juha Lehtonen ...

PAPER 1 : ADVENTURES WITH EPUB3

Author(s): Maureen Pennock, Michael Day
Presenter(s): Maureen Pennock

NOTES

Good session choice: adventures and mayhem! 🙌

Whistlestop tour - BL “pretty big” with “a lot of ebooks”

ePUB = standard maintained by the International Publishers Foundation

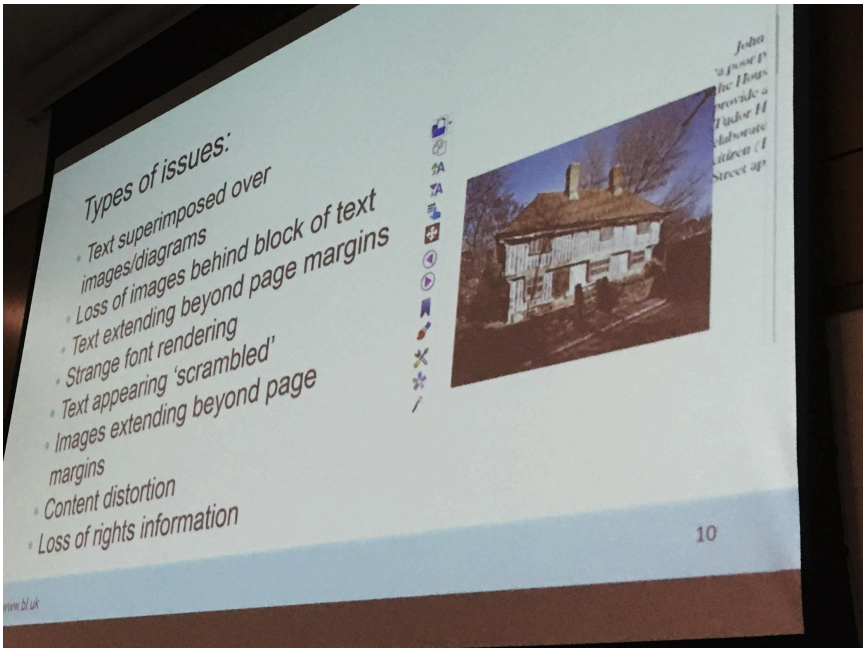
Format increasingly being used by publishers and therefore more frequently reaching libraries / archives

Epub designed for Reflowable text but some content benefits from fixed layout (ie. graphic novels, comics, children's books, cookbooks). Use cases exist for both - reflowable and fixed - e.g. comic books / graphic novels need fixed layer. Different viewing modes gave diff results on screen. ePub3.1 spec had embedded fixed lauyout. Used epubcheck to create sample.

Sample selected of 20 books. 5% of sample contained fixed layout declaration. 33% reflowable. Tested with Calibre (not supporting fixed) and Readium (which does).

- Majority of reflowable ePUB3s worked quite well in both. File constrcution methods caused some problems, readium worked marginally better.
- Fixed layout content results differed: Readium worked well, Calibre 3.14 struggled

At time of testing/writing: Calibre did not indicate supporting ePub3 - research intended as general overview of rendering issues.



Lessons learned: rendering matters! Files in the sample are intact & well formed. Different apps render files diff, dif views too, not all applications support all formats equally welcome. Lesson for wider collection - suitability

of rendering application can change when format is updated. Using later sw versions doesn't necessarily mean they support your material.
Characterisation: need to consider including format versions + format features.

Q&A

- Were findings reported back to calibre developers?
  Yes, but developer currently not interested in fixing the problem.
- Do you have a preference on how you receive files on a policy level?
  We have preferred formats list.

PAPER 2 : PDF MAYHEM

Author(s): Juha Lehtonen, Heikki Helin, Johan Kylander, Kimmo Koivunen
Presenter(s): Juha Lehtonen

NOTES

National digitisation services. Long path between creating digital files and preserving them. Try to fix errors before files arrive in preservation service. PDF complexity: from VW beetle to VW beetle. Most sw not specifically created for creating valid pdfs. 2 test sets: JHove PDF Test & Govdocs. Used veraPDF for validating PDFs. Acrobat reader regularly used but performs variety of fixes when opening a file.

Q: When you have PDFs und you have errors and they are still unreadable, how do you handle that?
A: that is negotiated with the partner institution. Primary target is that we won't get broken PDF files.

Q: Automated QA of migration? "Process that you are doing, your reconstruction is referred to as refrying and has been actively discouraged in the PDF development community for over 20 years now?"
A: Look at it as a great example of what not to do. What else should we do?
A: Restructure instead of refrying
(https://www.prepressure.com/pdf/basics/refrying)

The authors would like to denote that our presentation was a pilot study of publicly available broken PDF files to understand the PDF field better. By the name "PDF Mayhem" we want to say, that we were not presenting any strong

answers to the various PDF problems that exist. We are aware that our method that was used in the pilot study may result content loss or even new errors. We actually showed various examples of this kind of behaviour in our paper and presentation. We do not use this kind of method in our production system by default. Instead, in our paper we said that "further studies are needed". Our key message is that in the future the best way to fix problematic files is to focus on the process that produces these files. Still, we believe that some common errors in the already existing (at least simple) PDF files could possibly be fixed with a model similar to our pilot. In such case, we also reminded that a careful QA model for comparing the original and resulted files is needed.

PANEL : PDF/A:UNPACKING THE STANDARDIZATION PROCESS

Panel Moderator: Lisa Sisco
Speaker(s): Duff Johnson, Frederic Grevin, Yan Han, Leonard Rosenthol

Session chairs / paper authors! If you would like to share your agenda or any other information here ahead of time, please do so!

NOTES

Leonard: Atomic Energy Regulation in USA required to store data for the half-life of the isotopes - came to Adobe → ISO international committee to work on an archival standard PDF/A. Good about ISO process - every country gets to vote and says yay or nay for new standards.

DUff: Industry coordination with digital preservation
The PDF Association is a meeting-place for vendors and “investors” in PDF technology (like digital preservationists). It has ISo connections and Technical Working groups that employ an open, neutral and consensual process. A great success story from the PDF Association is veraPDF - the PDF/A validators expressly designed to meet preservationist AND industry needs. The product has been “industry supported” from the beginning, and, therefore, already widely adopted. Key message here is that aligning preservationst and industry needs drives preservation-minded software development.
http://verapdf.org/

https://www.pdfa.org/

Yan: “Why are there so many PDF/A options?” a frequently asked question when looking at PDF/A 1 - 4 with different conformance levels within each spec. Which one should I pick? Can be a confusing question. Answer is that not one is better than the other, but that they are different profiles that support different requirements and needs. One should not think that the PDF/A classes are read like version numbers where the highest number supersedes the previous ones.
“Preservation with PDF/A” - dpc technology watch report (2017)
https://www.dpconline.org/docs/technology-watch-reports/1707-twr17-01-revised/file

Fred: Get involved in the working group stuff - it’s fun. You get to travel.

Q: can’t you make the different conformance levels within the profiles simpler? E.g. U and UA?
A: we’ve heard the comments by the community and are streamlining this with PDF/4 - there will only be one conformance level. The only schism that couldn’t be brought together is “attachments” vs. “no attachments” (i.e., embedding arbitrary documents).

Q: what’s the most joyful & most frustrating experience about development?
A: most joyful is camaraderie - like fred mentioned. Most challenging is not to get frustrated in the long duration iterative standardization process.