

robust open reflowable universal

# The Editable PDF Initiative

standardized collaborative  
intuitive



PDF Days Europe 2022 | Berlin

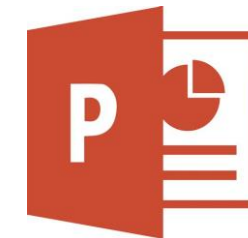
## What's holding PDF back?

Should PDF generation be a one-way street?

# The de-facto standard for ...



Final-form documents



L<sup>A</sup>T<sub>E</sub>X

Editable documents?

# Structure of the presentation



- Document exchange: Pains and opportunities
- Current solutions
- Proposal: Universal editability
- Technical challenges
- Proof of concept

# Document exchange



## Pains

- Differences in formats and versions
- Missing fonts can cause problems
- Differences in UI between apps
- Potential for typographic errors and text reflow
- “Good enough” for most use cases

## Opportunities

- Universal format for canvas-based documents (“digital paper”)
- Openness
- Robust layout
- Self-contained file
- Consistent user interface

# Unique selling proposition



- Why is PDF so popular today?

PDF guarantees layout and avoids text reflow, regardless of viewer app, system, installed fonts, screen resolution, etc.

- An editable format that offers the same guarantees **does not exist yet**

# Current solutions



- Embedding the source file
  - OpenOffice “Hybrid PDF” (content stream); PDF/A-3
- Often ignored by viewers such as Acrobat, Preview, Evince
  - Due to same file extension (.pdf) users are unaware that document is editable!
- Robust layout lost when editing document
- Not a generic solution; no portability when editing
- No guarantee that embedded file is same document (security?)

# Current solutions (2)



- Editing functionality of PDF software
  - Acrobat, FoxIt, Nitro...
  - Use AI to rediscover the PDF's structure
  - Suitable only for minor touch-up operations
  - Repeated editing and re-saving causes cumulative errors
- PDF forms

# Proposal: Universal editability



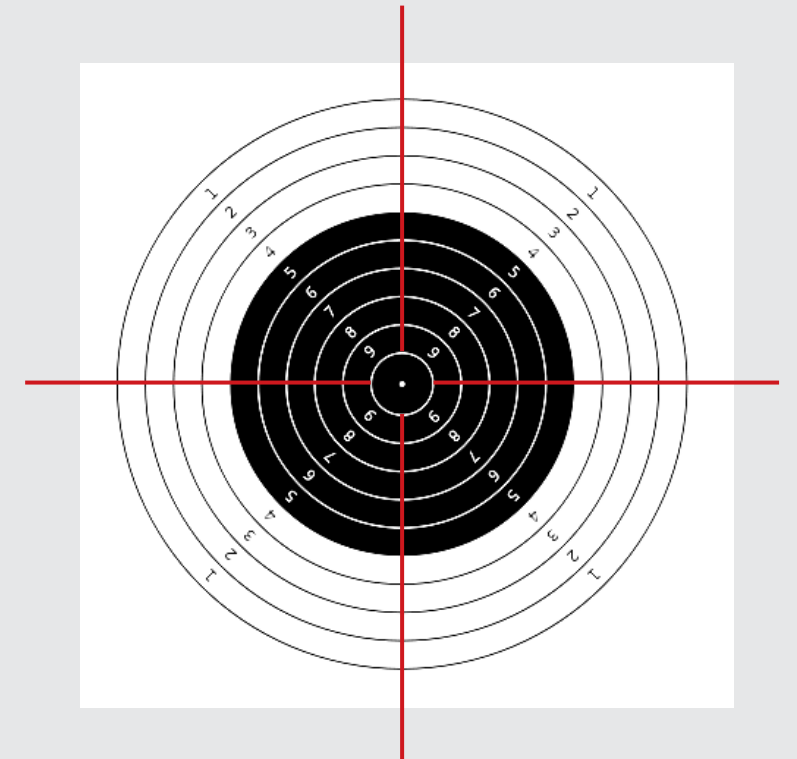
- Applications such as Word, PowerPoint and InDesign share many similarities in their model, UI, and the way that they handle text
  - PDF has the potential to unite all these models
  - Technical challenge with three major components:
    - Generic model for a canvas-based document
    - Clearly defined rules for editing and reflow
    - Robust text typesetting to account for platform and font differences
- [Hassan & Hunter DocEng 2015]



# Aside: What about web standards?



- A moving target!
- No layout robustness
- Underlying code can be a mess
- Currently a poor foundation to build upon
- However, also widely adopted
- In the long term, it is likely that these two worlds will unite





# Editing and reflow



## ■ Simple layouts

2

and other web pages with a similar structure. These rules utilize the document structure and specific attributes of the user-selected data items to locate other relevant data instances to be retrieved. The logical foundation on Web data extraction, and on the complexity and expressive power on data extraction using the Lixto approach were studied in [10, 11, 9].

In the first part of this paper, we will report on a recent addition within this Web data extraction framework to employ a learning strategy to select an optimal set of attributes for identifying the data items on a Web page. In the next part, we will discuss our approach to unsupervised data extraction from Web pages. Another topic that we are currently investigating is the extraction from non-HTML formats such as PDF. We will use a use case, i.e., extracting data in the domain of digital cameras, to illustrate the techniques that we have developed in these various fields of Web data extraction. Furthermore, we will report on industrial applications of Web data extraction and highlight the benefits that these applications can generate in a real-world setting.

### 2 Supervised Wrapper Generation

We will use the digital camera domain to illustrate typical use cases for Web data extraction. Let us imagine the application of monitoring camera prices. Assume that we have several competitors and we want to continuously monitor their websites for price development of the listed goods. We store the prices collected from their websites into a local database. We can then use business intelligence tools to analyse the aggregated prices, and this will allow us to react to market changes with more effective pricing strategies and advertisement campaigns. Figure 1 shows a sample of Web pages from the Dell online shop that serves as an example for similar online shops from which price information could be extracted.

This use case favors the usage of a supervised approach due to the following requirements:

- if the wrapping algorithm does not work on the given website, we cannot go to another site and collect the prices there;
- accuracy (precision/recall) must be high, because business decisions rely on the gathered data, and the solution therefore must guarantee the quality of the results obtained with the wrapping service;
- deep Web navigation is required before the actual data can be wrapped, e.g., it is required to fill Web forms or handle JavaScript (AJAX) execution.

In the following example, we need to collect prices of digital cameras from the Web site [www.dell.com](http://www.dell.com). To obtain prices for some of the cameras we have to navigate to the detail pages of the shopping cart. Informally, the problem we are trying to solve is: given a Web site (or a set of Web pages) as input and a user knowing what should be extracted from the Web site, we need to construct a wrapper to extract exactly the required information items.

## ■ Complex layouts

# AKTUEL

OKTOBER 2009

## VG WORT

VERWERTUNGSGESELLSCHAFT WORT · RECHTSFÄHIGER VEREIN KRAFT VERLEIHUNG

**Liebe Leserinnen und Leser,**

beim Google-Vergleich gibt es überraschende Entwicklungen. Am 18. September 2009 wurde bekannt, dass nach Auffassung des amerikanischen Justizministeriums der Vergleich in seiner gegenwärtigen Form nicht gebilligt, sondern von den Parteien nachverhandelt werden sollte. Daraufhin haben die Kläger des Ausgangsverfahrens (Authors Guild und Association of American Publishers) in Übereinstimmung mit der Beklagten (Google) beantragt, den Termin für die Gerichtsanhörung am 7. Oktober 2009 aufzuheben. Dem ist der zuständige Richter Danny Chin am 24. September 2009 nachgekommen. Stattdessen wird jetzt am 7. Oktober 2009 eine „status conference“ vor dem Gericht stattfinden, bei der entschieden werden soll, wie das Verfahren fortgesetzt wird.

Auf nationaler Ebene befinden sich die Verwertungsgesellschaften noch immer mit der Geistindustrie in schwierigen Verhandlungen über die Höhe der Vergütungssätze für bestimmte Geräte und Speichermedien. Betreffend ist davon insbesondere der audiovisuelle Bereich, in dem die Ausschüttungen bereits erheblich zurückgegangen sind.

Aufgeführte Informationen zu diesen und weiteren aktuellen Themen finden Sie in unserem Newsletter.

Mit freundlichen Grüßen

Rainer Jast                      Dr. Robert Strauß

Geschäftsführende Vorstände

**Aktuelle Entwicklungen im Google-Vergleich**

Beim Google-Vergleich ist wieder alles offen. Bekanntlich bedurfte der von Authors Guild und Association of American Publishers mit Google im Herbst 2008 abgeschlossene Vergleich noch einer endgültigen Bestätigung des zuständigen Gerichts in New York. Dabei war es möglich, gegen den Vergleich bis zum 8. September 2009 Einwände einzunreichen. Am 7. Oktober 2009 sollte ein so genanntes „Fairness Hearing“ stattfinden. Bis zum Tag des Fristablaufs waren zahlreiche Einwendungen geltend gemacht worden. Unter anderem hatte der Börsenverein des Deutschen Buchhandels gemeinsam mit

anderen europäischen Verlegerverbänden und Verlagen beim zuständigen Gericht in New York einen Schriftsatz mit „objections“ eingereicht. Auch das deutsche Bundesministerium der Justiz hatte einen so genannten „amicus curiae“ - Schriftsatz übermittelt und ebenfalls erhebliche Bedenken gegen den Vergleich vorgetragen. Das US-Gericht hatte außerdem das amerikanische Justizministerium gebeten, zu den kartellrechtlichen Fragen des Vergleichs Stellung zu nehmen. Am 18. September 2009 veröffentlichte das Ministerium seine Stellungnahme. Dabei ging es nicht nur um kartellrechtliche Fragen, sondern um eine rechtliche Bewertung des Vergleichs insgesamt. Das Ministerium kam zu dem klaren Ergebnis, dass das Settlement in seiner derzeitigen Fassung gerichtlich nicht gebilligt werden sollte. Vielmehr sollten die Parteien des Vergleichs neu verhandeln, um die Regelungen in Übereinstimmung mit Urheber- und Kartellrecht zu bringen. Damit war klar, dass der Vergleich in der bisherigen Fassung kaum aufrechterhalten bleiben würde. Wenige Tage später haben Authors Guild und Association of American Publishers in Übereinstimmung mit Google beantragt, dass das Fairness Hearing vertagt und Neuverhandlungen aufgenommen werden sollten. Das Gericht ist diesem Antrag nachgekommen. Am 7. Oktober 2009 findet jetzt stattdessen eine „status conference“ statt, um das weitere Verfahren festzulegen.

Was bedeutet diese überraschende Entwicklung für die VG WORT? Zunächst ist es ein Erfolg, dass die berechtigten Einwände von Autoren und Verlagen dazu geführt haben, dass der bisherige Vergleich nicht aufrechterhalten wird. Aufgabe der VG WORT war es dabei stets, sicherzustellen, dass auch bei einer Genehmigung des Vergleichs die Rechte der deutschen Autoren und Verlage bestmöglich gewahrt werden können. Die VG WORT hatte sich deshalb bekanntlich die Vergütungsansprüche für die bereits vorgenommenen Digitalisierungen sowie das Recht, die Werke aus dem Digitalisierungsprogramm von Google zurückzurufen („removal“), übertragen lassen. Gleichzeitig war ihr das Recht eingeräumt worden, digitale Nutzungen von vergriffenen Werken zu lizenzieren, wenn die

# Editing and reflow

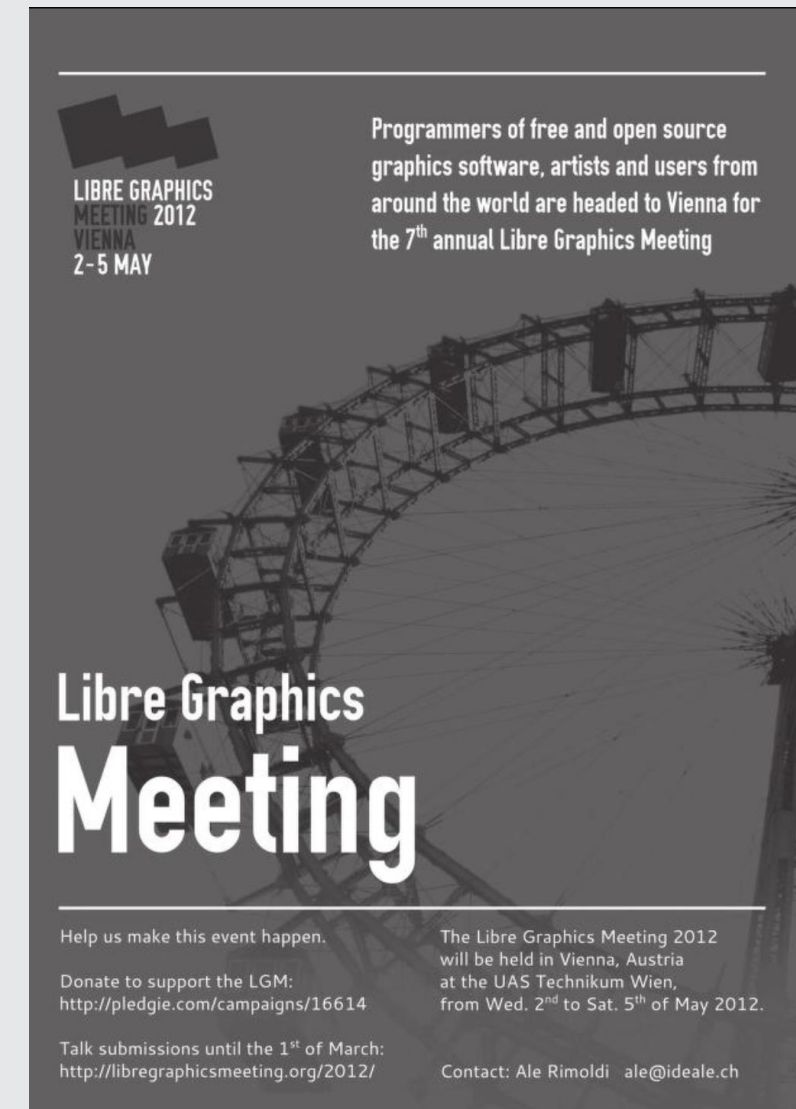


- Simple layouts
  - Automatic reflow trivial
- Complex layouts
  - May contain multiple columns, floats, etc.
  - Automatic reflow possible within limitations according to defined model
  - Float repositioning manually initiated; app-dependent

# Editing and reflow



- Free-form layouts
  - “Mostly graphic” documents
  - Text positioned in frames, like a DTP
  - Ultimate freedom
  - Reflow possible between frames; frame dimensions may need adjusting afterwards





# Robust text typesetting



- We can maintain line breaks, even if minor typographical changes or even edits occur
- [Hassan & Hunter DocEng 2015] shows that the area of a text column can be varied by up to 10% without significantly altering typographic appearance
- More drastic changes (e.g. font substitution) necessitate a change in character width

In old times when wishing still helped one, there lived a king whose daughters were all beautiful, but the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face.

In old times when wishing still helped one, there lived a king whose daughters were all beautiful, but the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face.

# Aside: Fonts



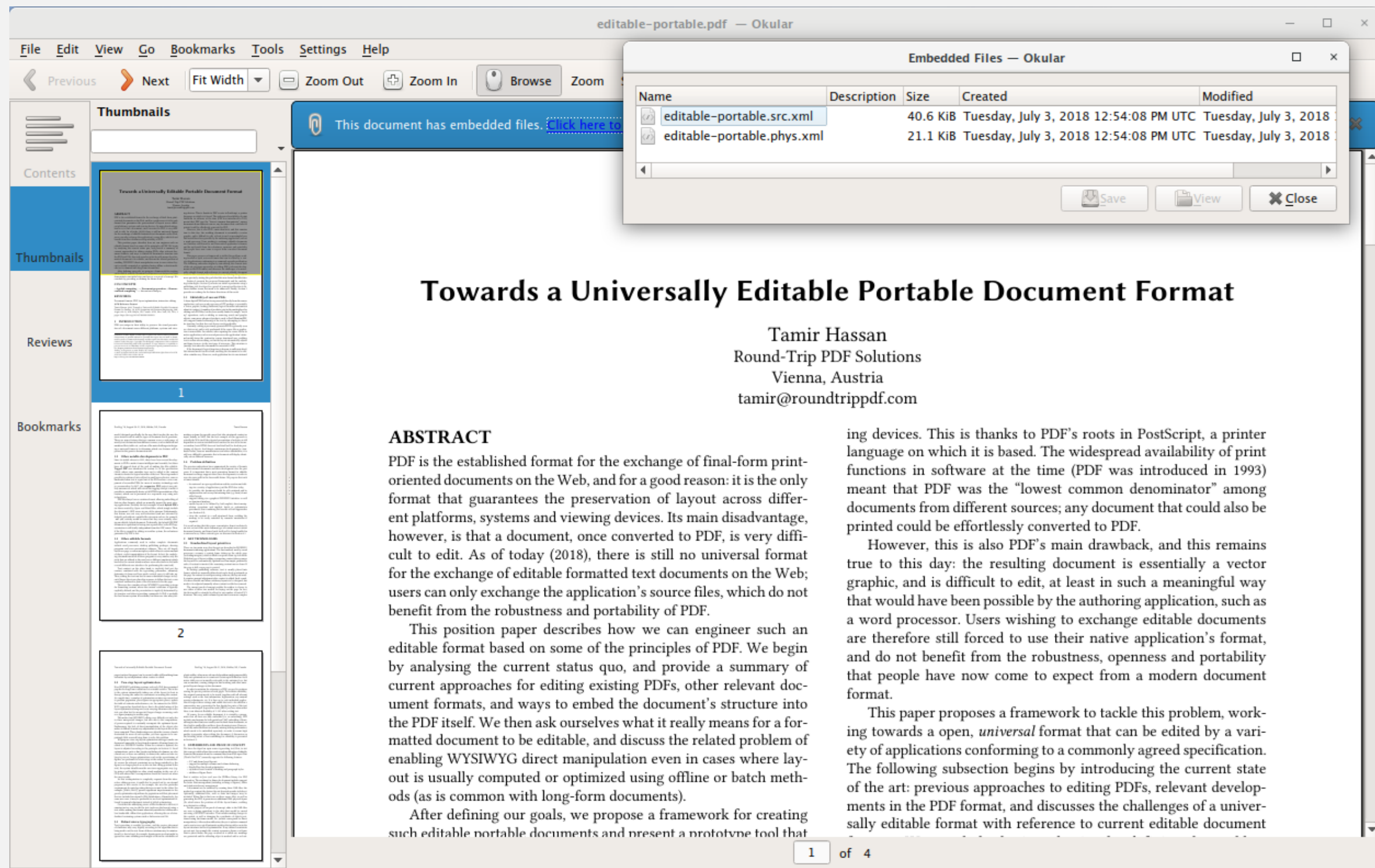
- Fonts are complicated, technically and legally
- Many commercial fonts are licensed only for embedding as a subset
- Conservative editing replaces the fonts only where edits take place
- By storing the metrics, editing is possible on a different computer, even though a different font is being displayed

# Proof of concept




- Prototype based on Pint Formatter
  - “Pint Is Not TeX”
  - <https://github.com/tamirhassan/pint-publisher>
  - Published in [Hassan DocEng 2018]







Open





editable-portable-new.flex.xml  
~/workspaces/publisher/Publisher

Save









editable-portable-new.phys.xml

×

editable-portable-new.flex.xml

×

```
<!-- Author information -->
<h4 id="15465229">Tamir Hassan</h4>
<h5 id="7828650">Round-Trip PDF Solutions</h5>
<h5 id="11372245">Vienna, Austria</h5>
<h5 id="3756058">tamir@roundtrippdf.com</h5>

</col>

<multi-col id="680775" num-cols="2">

  <h2 id="11771699">ABSTRACT</h2>
  <p id="9198523">
    PDF is the established format for the exchange of final-form print-oriented
    documents on the Web, and for a good reason: it is the only format
    that guarantees the preservation of layout across different platforms,
    systems and viewing devices. Its main disadvantage, however, is that
    a document, once converted to PDF, is very difficult to edit. As of
    today (2018), there is still no universal format for the exchange
    of editable formatted text documents on the Web; users can only exchange
    the application's source files, which do not benefit from the robustness
    and portability of PDF.
  </p>

  <p id="12072298">
    This position paper describes how we can engineer such an editable
    format based on some of the principles of PDF. We begin by analysing
    the current status quo, and provide a summary of current approaches
    for editing existing PDFs, other relevant document formats, and ways
    to embed the document's structure into the PDF itself. We then ask
    ourselves what it really means for a formatted document to be editable,
    and discuss the related problem of enabling WYSIWYG direct manipulation
    even in cases where layout is usually computed or optimized using
    offline or batch methods (as is common with long-form documents).
  </p>

  <p id="12159182">
    After defining our goals, we propose a framework for creating such
```

XML

▼

Tab Width: 8

▼

Ln 437, Col 29

▼

INS

n Java and uses the PDFBox  
F generation. The roadmap  
opment includes support for  
management (including resiz-  
tables and citation/reference

system is the concept of a  
*ment*. A document can be  
eating three XML files: the  
ent, the layout description and  
tionally, additional files, such  
ages, may be included. When  
on these source files, it gener-  
document; this can be either  
n additional *layout file* in the



Fig. 1

called out, warnings are generated and the  
offending object is marked in red on the PDF  
(see Figure 1). In such a case, the command  
pint recompute globally optimizes the layout  
to meet these constraints.

## ogies of the proposed

cribes three key technologies  
of the core framework for  
id have been implemented in  
ncept prototype described in

### ayout primitives

main ways that layouts are  
WYSIWYG document authoring  
e first method, used by word  
mes a content frame taking up  
(excluding margins), which is



Fig. 1

Non-WYSIWYG publishing systems, such as  
LaTeX, have remained popular for long-form  
content such as scientific articles. This is due  
to the system automatically taking care of the  
layout (at least in theory), leaving the author  
to concentrate on writing the content. At com

# Conclusion



- A universal, editable format is a great opportunity for PDF
- The technical challenges are solvable; market penetration requires strong partners and a showcase product
- Looking for collaborators
- References
  - The Editable PDF Initiative, <https://editablepdf.org>
  - [Hassan DocEng 2018] Towards a Universally Editable Portable Document Format
  - [Hassan & Hunter DocEng 2015] Knuth-Plass Revisited