More than just digital paper—hidden features of the PDF format

Dietrich von Seggern
Vice Chair, PDF Association
callas software
Berlin, Germany
d.seggern@callassoftware.com

Klaas Posselt
Member, PDF Association
digital Prepress & ePublishing Consulting
Berlin, Germany
klaas.posselt@einmanncombo.de

Tamir Hassan
Member, PDF Association
Round-Trip PDF Solutions
Vienna, Austria
tamir@roundtrippdf.com

Thomas Zellmann
Managing Director, PDF Association
Foxit Europe
Berlin, Germany
thomas.zellmann@pdfa.org

ABSTRACT

PDF has long been established as the *de facto* format for the exchange of print-oriented documents and is known for its robust visual presentation across a variety of operating systems and platforms.

However, relatively few users are familiar with the format's newer features, such as tagging, forms and security. This tutorial aims to give an overview of the most important of these features and demonstrate the benefits of creating and exchanging PDF files that make use of them.

CCS CONCEPTS

• **Applied computing** → **Publishing**; **Document preparation**;

KEYWORDS

PDF, Tagging, Machine-readable structure, Workflows, Document formats

ACM Reference Format:
Dietrich von Seggern, Klaas Posselt, Tamir Hassan, and Thomas Zellmann. 2019. More than just digital paper—hidden features of the PDF format. In *ACM Symposium on Document Engineering 2019 (DocEng '19), September 23–26, 2019, Berlin, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3342558.3351873

1 SUMMARY

PDF is the most common file format on the Web after HTML, and everyone in the Document Engineering community without exception has to deal with the format in one way or another. Just about everyone is familiar with PDF being "digital paper," after all, that is really how it started out. But PDF has become "smarter" in the last few years, yet it is still struggling to shake off its reputation as a purely end-of-line, dumb file format that is unsuitable for further machine processing. Comparatively few people—and this includes even many DocEng participants who do not work directly with PDF—are not familiar with PDF's additional features, which have grown over the past few years.

The aim of this tutorial is to introduce the audience to the most important of these features and give practical examples on how they can benefit from generating and exchanging PDF files that go beyond digital representations of the printed page.

The following section lists the topics that were prepared for this half-day tutorial. However, we also plan to ask the audience about what interests them the most and spend more time on those topics.

2 TOPICS

- Tagged PDF: Embedding logical structure in PDF for:
  - accessibility
  - repurposing of content
- Hidden, selectable text for scanned documents via OCR
- Structured data (measurements, statistical data) as embedded files
- Digital, fillable forms in PDF
- Object-level metadata; keeping source information for history and license enforcement
- Commenting workflows (annotations)
- Security: Encryption and signatures
- Compression algorithms enabling small file sizes for high-quality images
- Colour reproduction (if requested by audience)

3 ABOUT THE ORGANIZERS

Dietrich von Seggern received his degree as a printing engineer, and in 1991 started his professional career as head of desktop prepress production in a reproduction house. He became involved in research projects for digital transmission of print files, and moved to the German Newspaper Marketing Organization (ZMG), where he was responsible for a project to enable the digital transmission of newspaper ads. In 2002 he joined callas software as Director of Product Management, and subsequently introduced callas' PDF/A related products for the archiving industry. Today, Dietrich is Managing Director at callas software, and Vice Chair and ISO Liaison Officer at the PDF Association.

DocEng '19, September 23–26, 2019, Berlin, Germany

ACM ISBN 978-1-4503-6887-2/19/09.
https://doi.org/10.1145/3342558.3351873

Klaas Posselt is a graduate engineer in printing and media technology who, following a number of lines of inquiry, eventually landed on the subject of universally accessible PDF documents. He trains, assists and supports clients as they implement and optimize publication processes and move towards new digital output channels including ebooks, accessible PDFs and web platforms. As a member of the PDF Association's PDF/UA Competence Center, he is involved in the ongoing development and growth of the PDF/UA standard for universally accessible PDF documents.

Tamir Hassan has over a decade of experience in the area of document engineering. After writing his doctoral thesis on the topic *User-Guided Information Extraction from Print-Oriented Documents,* he worked as a researcher in academia, and more recently in the Printing and Content Delivery Laboratory of HP Labs. He now works as an independent researcher and consultant offering solutions for round-trip processing of PDF documents (i.e. both the conversion of formatted documents to structured form and the automated creation of of formatted documents from structured content or data).

Tamir is a regular contributor to DocEng. In 2016, he was the Program Chair and has been active in the Steering Committee since then. He has been a member of the PDF Association since 2017.

Thomas Zellmann has been working in electronic data processing (EDP) for more than 30 years and has extensive experience with classic and modern IT solutions. Prior to joining LuraTech/Foxit in 2001, he worked for Softmatic AG, Software AG and Nixdorf among others. Thomas focuses on the banking/insurance and archives/libraries segments. As Managing Director of the PDF Association, Thomas coordinates and executes many of the organization's activities.