

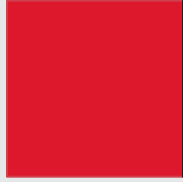


Foto: Natale Carioni



Tagged PDF

What makes it a proper PDF/UA?



René Treuber

Product Manager @ axaio software GmbH

- Why are we talking about tagged PDF / PDF/UA
- What is PDF/UA
- The Matterhorn Protocol
- Important checks

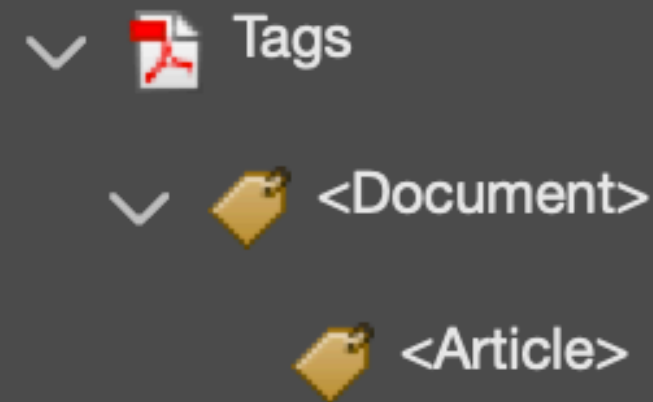
Why tagged PDF



- PDF wide spread in content delivery
- Ensure content is accessible
- Tags = Structure

Issues with “tagged” PDF

- Not tagged at all
- Tagged, but incomplete
 - ... empty
 - ... only page numbers
- Tagged, but failed semantic
 - ... use only “P”
 - ... use other “wrong” tags



all tags in ISO:32000 pdf

Why PDF/UA



- ISO (14289) standard for accessible PDF
 - current version ISO 14289-1:2014
- Description of tagged PDF
 - creators can follow
 - assistive technology can rely and assist
- *“Content shall be marked in the structure tree with semantically appropriate tags in a logical reading order.”*

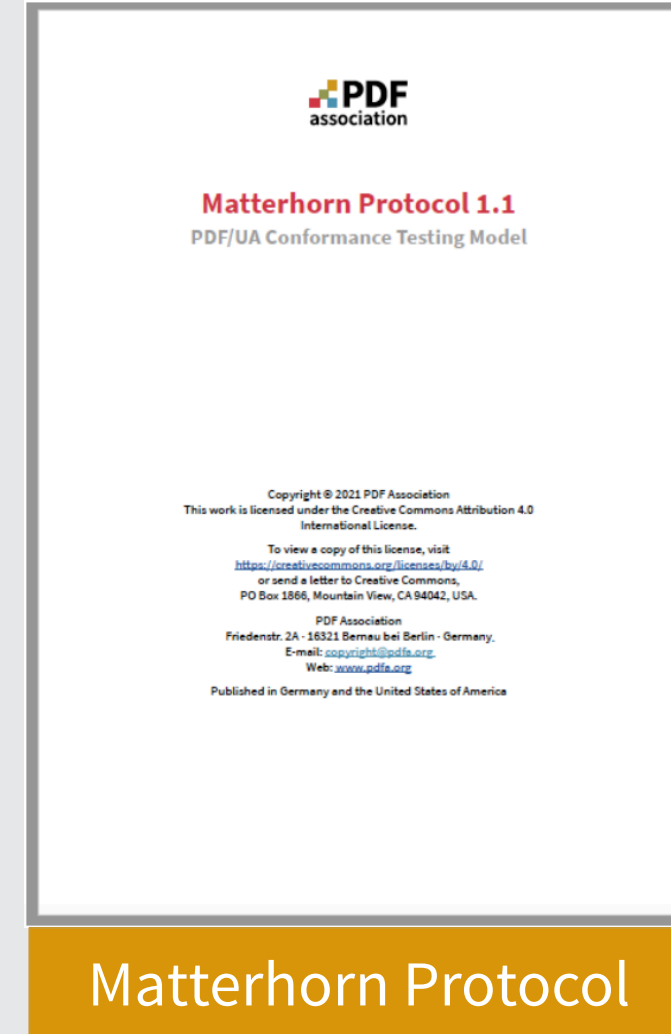


Validate

The Matterhorn Protocol



- PDF/UA Conformance Testing Model
- Created by PDF Association
 - publicly available
 - updated version in 2021



The Matterhorn Protocol



- For document creators
 - test validity
- For document testers
- For software developers
 - creator software
 - validation software
- As reference
 - conforms to both PDF/UA-1 (ISO 14289-1) and to PDF/A-2a (ISO 19005-2)

The Matterhorn Protocol

- Specifying a set of tests
- 31 Checkpoints
 - 136 Failure Conditions
- Human (87) vs Machines (47)

PDF/UA-1 Failure Conditions

Checkpoint 01: Real content tagged

Index	Failure Condition	Section	Type	How	Sec
01-001	Artifact is tagged as real content.	UA1:7.1-1	Object	H	–
01-002	Real content is marked as artifact.	UA1:7.1-1	Object	H	–
01-003	Content marked as Artifact is present inside tagged content.	UA1:7.1-1	Object	M	–
01-004	Tagged content is present inside content marked as Artifact.	UA1:7.1-1	Object	M	–
01-005	Content is neither marked as Artifact nor tagged as real content.	UA1:7.1-2	Object	M	–
01-006	The structure type and attributes of a structure element are not <i>semantically</i> appropriate for the structure element. All of the following structure types must be taken into account: <Document>, <Part>, <Art>, <Sect>, <Div>, <BlockQuote>, <Caption>, <TOC>, <TOCI>, <Index>, <NonStruct>, <Private>, <P>, <H>, <H1>, <H2>, <H3>, <H4>, <H5>, <H6>, <L>, , <Lbl>, <LBody>, <Table>, <TR>, <TH>, <TD>, <THead>, <TBody>, <TFoot>, , <Quote>, <Note>, <Reference>, <BibEntry>, <Code>, <Link>, <Annot>, <Pub>, <Marichu>, <DD>, <DT>, <DD>, <WT>	UA1:7.1-2	Object	H	–

Matterhorn Checkpoints

Matterhorn Checkpoints



- 01: Real content tagged
- 02: Role Mapping
- 03: Flickering
- 04: Color and Contrast
- 05: Sound
- 06: Metadata
- 07: Dictionary
- 08: OCR Validation
- 09: Appropriate Tags
- 10: Character Mappings
- 11: Declared Natural Language
- 12: Stretchable Characters
- 13: Graphics
- 14: Headings
- 15: Tables
- 16: Lists
- 17: Mathematical Expressions
- 18: Page Headers and Footers
- 19: Notes and References
- 20: Optional Content
- 21: Embedded Files
- 22: Article Threads
- 23: Digital Signatures
- 24: Non-Interactive Forms
- 25: XFA
- 26: Security
- 27: Navigation
- 28: Annotations
- 29: Actions
- 30: XObjects
- 31: Fonts

Checking software



- Various software available
 - PDF Accessibility Checker
 - CommonLook PDF Validator
 - Acrobat Preflight
 - callas pdfaPilot
 - ...

PDF Accessibility Checker

- PAC 3 / PAC 2021
- based on PDF/UA specification and has all machine based matterhorn checks
- 2021 version includes WCAG criteria



The screenshot shows the PDF Accessibility Checker 2021 interface. The title bar reads "PAC 2021 - PDF Accessibility Checker 2021". The main header displays the logo, "PDF Accessibility Checker 2021", and "Version: 21.0.0.0". A status bar at the top indicates "This PDF file is not PDF/UA compliant." with a red X icon.

The interface is divided into two main sections: "PDF/UA" and "WCAG". The "PDF/UA" section shows a list of checkpoints with their status:

Checkpoint	Status
PDF Syntax	✓
Fonts	✓
Content	✓
Embedded Files	✗
Natural Language	✓
Structure Elements	✗
Structure Tree	⚠
Role Mapping	✓
Alternate Descriptions	✗
Metadata	✗
Document Settings	✓

The "WCAG" section is currently inactive. Below the checkpoints, there are buttons for "Results in Detail" and "Screen Reader Preview".

The "Report" section on the right displays a table of results:

Title	Count
PDF/UA	1,151
Basic requirements	0
ISO 32000-1	0
Fonts	0
Content	0
Embedded Files	0
Natural language	0
Logical Structure	1,150
Structure Elements	2
Headings	0
Notes	0
Annotations	0
Figures	2
Tables	0
Structure Tree	1,105
Role mapping	0
Alternative Descriptions	43
Alternative text for "Figure" structure elements	43
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	
Alternative text missing for "Figure" struc...	

Common Look PDF Validator

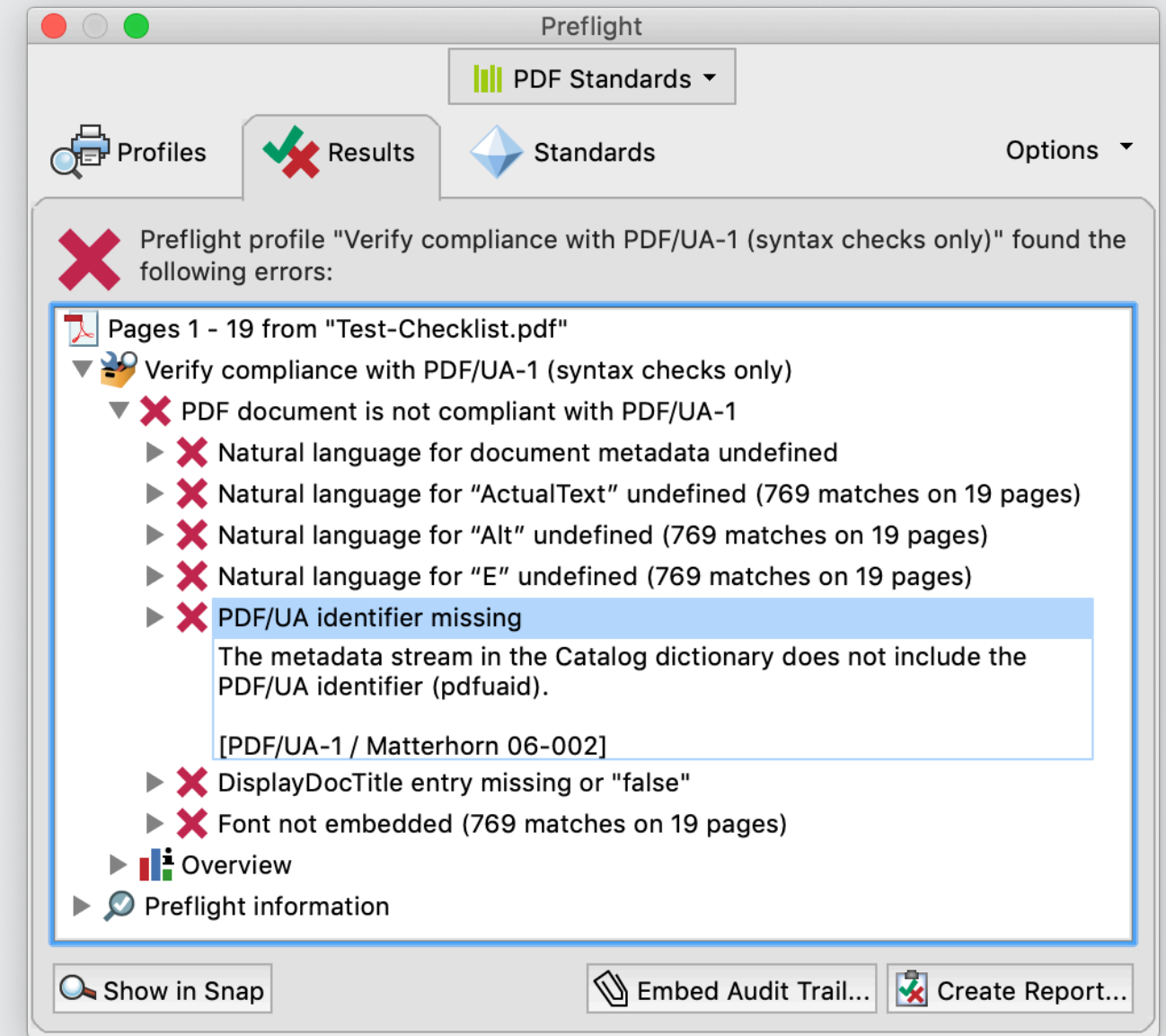


- Acrobat Plug-In (Windows)
- Checks based on
 - PDF/UA
 - Section 508
 - WCAG
 - ...

Results						
Filter by : All		Report	Status	Column Options	Clear f	
Resul	Checkpoint	Test	Standard N	Results	Description	C
1	7.1 General	Document	PDF/UA 1.0	Passed	Document elemer	^
2	7.1 General	Correct Structure - RP, RB and	PDF/UA 1.0	Not Applicable	No RP, RB or RT el	
3	7.1 General	Correct Structure - Ruby	PDF/UA 1.0	Not Applicable	No Ruby elements	
4	7.1 General	Table Cells	PDF/UA 1.0	Passed	All Table Data Cell	
5	7.1 General	THead, TBody and TFoot	PDF/UA 1.0	Passed	All THeads, TFoots	
6	7.1 General	Table Rows	PDF/UA 1.0	Passed	All Table Rows pas	
7	7.1 General	Table	PDF/UA 1.0	Passed	All Table elements	
8	7.1 General	TOCI	PDF/UA 1.0	Not Applicable	No TOCI elements	
9	7.1 General	TOC	PDF/UA 1.0	Not Applicable	No TOC elements	
10	7.1 General	Correct Structure - Warichu	PDF/UA 1.0	Not Applicable	No Warichu eleme	
11	7.1 General	Correct Structure - WT and WP	PDF/UA 1.0	Not Applicable	No WP or WT eler	
12	7.1 General	Valid Role Maps	PDF/UA 1.0	Passed	Passed Role Map 1	
13	7.1 General	Metadata - Title and Viewer Pr	PDF/UA 1.0	User Verification	Please verify that i	
14	7.1 General	No Suspects in document.	PDF/UA 1.0	Passed	The Suspects entr	
15	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
16	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
17	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
18	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
19	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
20	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
21	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
22	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	
23	7.1 General	Role Maps - Semantic appropri	PDF/UA 1.0	User Verification	Please check that	v
Number of results: 685						

Acrobat Preflight

- Acrobat Preflight / callas pdfaPilot
 - Verify compliance with PDF/UA
 - Checks based on Matterhorn protocol
 - provide checkpoint index number,
e.g. [PDF/UA / Matterhorn 06-002]



Common Pitfalls



- **01: Real content tagged**
- **02: Role Mapping**
- *03: Flickering*
- *04: Color and Contrast*
- *05: Sound*
- **06: Metadata**
- 07: Dictionary
- **08: OCR Validation**
- **09: Appropriate Tags**
- 10: Character Mappings
- 11: Declared Natural Language
- 12: Stretchable Characters
- **13: Graphics**
- **14: Headings**
- **15: Tables**
- 16: Lists
- 17: Mathematical Expressions
- 18: Page Headers and Footers
- 19: Notes and References
- 20: Optional Content
- 21: Embedded Files
- 22: Article Threads
- 23: Digital Signatures
- 24: Non-Interactive Forms
- 25: XFA
- 26: Security
- 27: Navigation
- 28: Annotations
- *29: Actions*
- 30: XObjects
- 31: Fonts

Checkpoint 01: Real content tagged



- Real content vs. artifacts
 - decision has to be made
 - page numbers
- Semantically appropriate tags
 - use all available

Checkpoint 09: Appropriate Tags

- Logical reading order
- Inappropriate nesting
- Table for non table data

And you will read this last

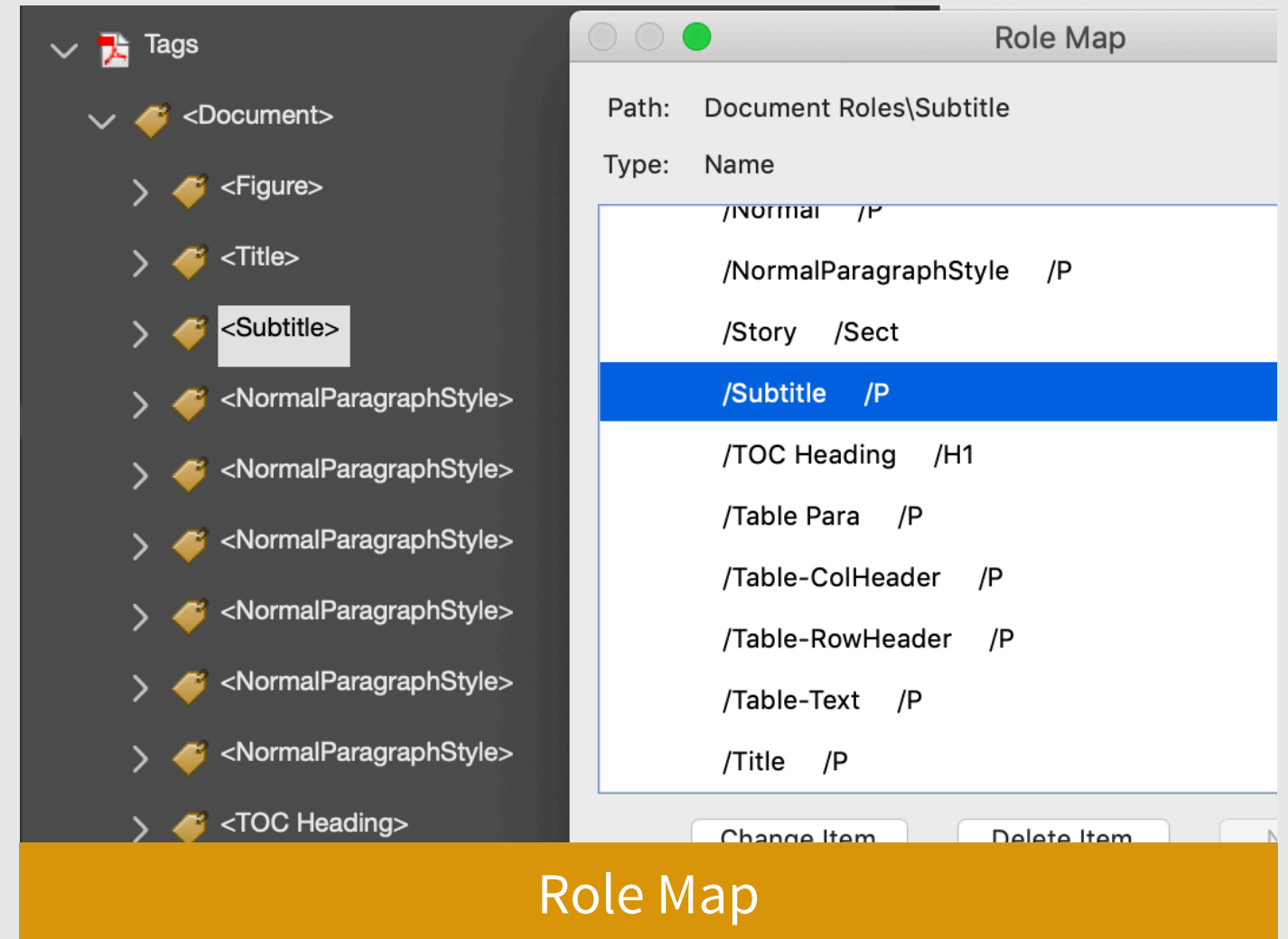
**You will read
this first**

And then you will read this

Then this one

Checkpoint 02: Role Mapping

- All tags must map to a standard structure type
- Appropriate mapping
- No remapping of standard structure type



The screenshot displays the 'Role Map' application. On the left, a tree view under 'Tags' shows a hierarchy: <Document> (expanded) contains <Figure>, <Title>, <Subtitle> (highlighted), and six instances of <NormalParagraphStyle>, followed by <TOC Heading>. On the right, a table lists mappings for the selected <Subtitle> tag. The table has columns for 'Path' and 'Type'. The 'Path' column shows 'Document Roles\Subtitle' and the 'Type' column shows 'Name'. The table lists various roles and their corresponding structure types, with '/Subtitle /P' highlighted in blue.

Path	Type
/Normal /P	
/NormalParagraphStyle /P	
/Story /Sect	
/Subtitle /P	
/TOC Heading /H1	
/Table Para /P	
/Table-ColHeader /P	
/Table-RowHeader /P	
/Table-Text /P	
/Title /P	

Buttons at the bottom: Change Item, Delete Item

Role Map

Checkpoint 14: Headings



- 6 Levels defined (H1 – H6)
- Always start with H1
- Don't skip levels
- Use appropriate levels
 - Typography ≠ Semantic

Checkpoint 06: Metadata



- XMP metadata
 - PDF/UA identifier
 - dc:title
 - clearly identify document

Checkpoint 08: OCR Validation

- Needs to be tagged
- Check for significant errors

xxiv

INTRODUCTION

Shakespeare's task. They often speak as if the poet started with some central idea of which Hamlet was to be the exponent. "Shakespeare," wrote Goethe, "sought to depict a great deed laid upon a soul unequal to the performance of it." "In Hamlet," wrote Coleridge, "Shakespeare seems to have wished to exemplify the moral necessity of a due balance between our attention to the objects of our senses and our meditation on the working of our minds—an equilibrium between the real and the imaginary worlds." I prefer to think of Shakespeare as setting to work with the intention of rehandling the subject of an old play, so as to give it fresh interest on the stage; as following the subject given to him, and as following the instinctive leadings of his genius. The traditional Hamlet was distinguished by intellectual subtlety, by riddling speech, by a power of ingeniously baffling his pursuers, and, at the same time, by a love of truth. But the subtlety of Saxo's Amleth—and we may be sure the same is true of Kyd's Hamlet—was what Burke happily describes, in a different connection, as a "clumsy subtlety." If he would be taken to be mad, he affects unclean and brutal habits, or crows like a cock, or rides a horse with his head towards the tail. Shakespeare was attracted by the intellectual subtlety of Hamlet, and was inevitably led by his genius to refine this subtlety, and to diversify its manifestations. He was caught in the web of his own imaginings, and became so absorbed in his work that he forgot to keep it within the limits suitable for theatrical representation; the tragedy has, perhaps, never been presented in its entirety on the English stage.

OCR'd text

Checkpoint 13: Graphics



- `<figure>` tag for graphical content
 - not only pixel data
- Need alternative representation
 - Alternate Text
 - Actual Text
 - For figures intended to be consumed as text

Checkpoint 15: Tables




- Use table for table data (only)
 - Rows + Columns
- Need to have headers
 - via Scope
 - via Headers and IDs
- In header cells, use <P>


Several Checkpoints

- Checkpoint 03: **Flickering**
 - Prevent flickering
- Checkpoint 04: **Color** and **Contrast**
 - Information not solely provided by color (e.g. in tables)
- Checkpoint 05: **Sound**
 - Audio must be available in another form, cf. figures
- Checkpoint 29: **Actions**
 - Don't require specific timing for keystrokes

- Matterhorn Protocol: no clear definition
 - Refer to WCAG 2.0
- Use sufficient contrast
- Contrast ratio depends on text size
 - regular: 4.5:1 (AA) / 7:1 (AAA), large/bold: 3:1 (AA) / 4.5:1 (AAA)
- Conflict with Corporate Identity

- <https://webaim.org/resources/contrastchecker/>
- checks against WCAG criteria
- for AA and AAA and different text sizes

Foreground Color
#8A8A8A
Lightness


Background Color
#FFFFFF
Lightness


Contrast Ratio
3.45:1
[permalink](#)

Normal Text
WCAG AA: **Fail**
WCAG AAA: **Fail**


The five boxing wizards jump quickly.

Large Text
WCAG AA: **Pass**
WCAG AAA: **Fail**

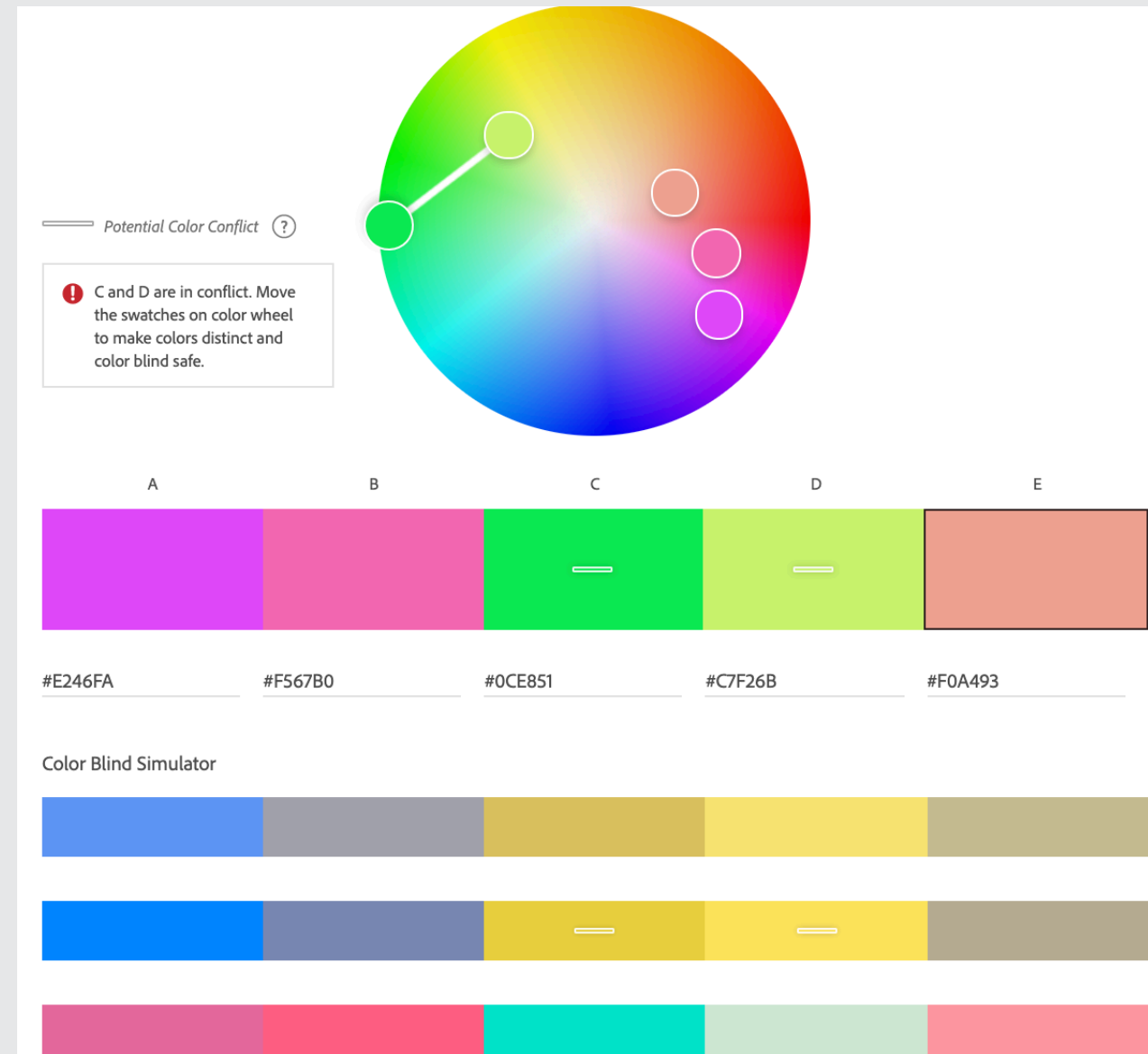
The five boxing wizards jump quickly.

Graphical Objects and User Interface Components
WCAG AA: **Pass**

Text Input



- Adobe Color Tools
 - Contrast Checker
 - <https://color.adobe.com/create/color-contrast-analyzer>
 - Color Blind Safe
 - <https://color.adobe.com/create/color-accessibility>



Conclusion



- Start!
- Keep educating yourself
- Use validation based on Matterhorn Protocol
- Do as much as possible in your authoring tools

Additional Technical Days Sessions

Lessons from implementing a PDF/UA-centric mobile reader

What we've learned while making a better mobile experience

Accessibility

Oct. 7, 17:00
1100 ET / 0800 PT



Matt Kuznicki
AbleDocs

WCAG or PDF/UA: what's the difference

Gap analysis and challenges in validation

Accessibility

Oct. 12, 16:00
1000 ET / 0700 PT



Boris Doubrov
Dual Lab sprl

Accessible PDF – How to tag content the right way

Finding appropriate Tags for content

Accessibility

Reuse

Oct. 12, 17:00
1100 ET / 0800 PT



Klaas Posselt
Einmanncombo

Our Solution Days Session

axaio MadeToTag – The tool
for proper accessible PDF
out of Adobe InDesign

Oct. 19, 16:30 1030 PT / 0730 ET



Questions?

■ Title Slide

- Source: https://commons.wikimedia.org/wiki/File:Matterhorn_from_the_south.jpg
- Author: [Cortomaltese](#)
- License: [CC BY 2.0](#)