PDF

PDF Days Online 2021

Deriving HTML from PDF lessons learned

Implementation challenges in reusing PDF content

History, Evolution of PDF

- Electronic representation of paper. Based on graphical model, accurate representation(screen and print)
  - Adopted word **rendering**
- Marked content
  - Introducing semantic into the content
- Tagged PDF
  - PDF/UA
  - First time used different presentation model for assistive technology

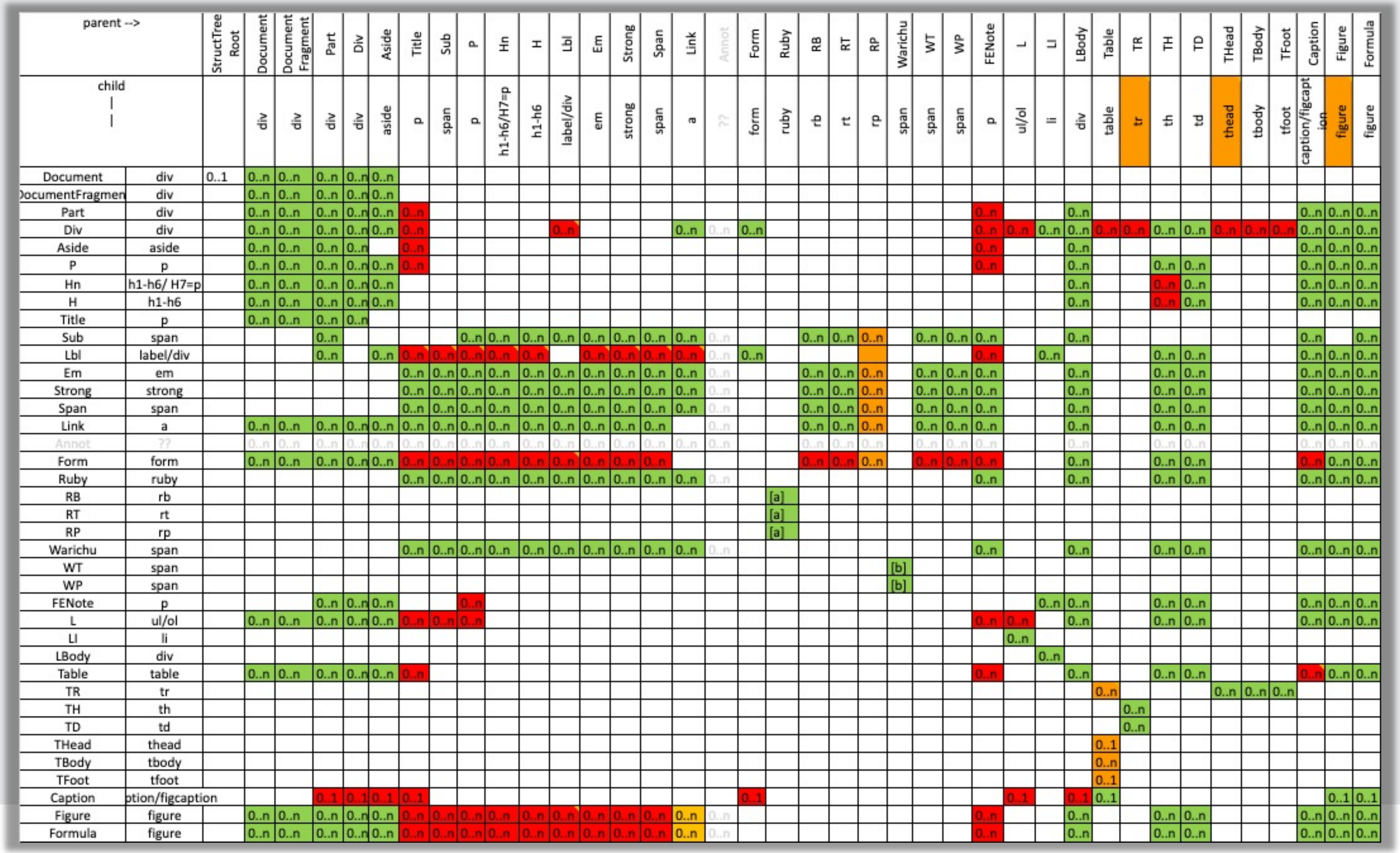
2019

- Problems – PDF in the web, in mobiles
  - Can’t control UX
  - Interactivity, navigation
  - Responsivness
- Solution – Deriving HTML from PDF
  - Algorithm that produces conforming HTML from a tagged PDF
  - Derivation = HTML rendering
  - Author controls the output not the processor

HTML Derivation - concept

- “converting” pdf structure elements to html tags
- Structure tree
  - Marked content
  - links, annotations, form fiels
- Attributes (Layout, CSS, HTML) via ClassMap too
- Associated files
- Actions
- Rolemap, Namespaces

Valid HTML



parent -->		StructTree Root	Document	Document Fragment	Part	Div	Aside	Title	Sub	P	Hn	H	Lbl	Em	Strong	Span	Link	Annot	Form	Ruby	RB	RT	RP	Warichu	WT	WP	FENote	L	LI	LBody	Table	TR	TH	TD	THead	TBody	TFoot	Caption	Figure	Formula
child			div	div	div	div	aside	p	span	p	h1-h6/H7=p	h1-h6	label/div	em	strong	span	a	??	form	ruby	rb	rt	rp	span	span	span	p	ul/ol	li	div	table	🟧 tr	th	td	🟧 thead	tbody	tfoot	caption/figcaption	🟧 figure	figure
Document	div	0..1	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n																																	
DocumentFragmen	div		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n																																	
Part	div		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n																			🟥 0..n			🟩 0..n								🟩 0..n	🟩 0..n	🟩 0..n
Div	div		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n					🟥 0..n				🟩 0..n	0..n	🟩 0..n								🟥 0..n	🟥 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟩 0..n	🟩 0..n	🟩 0..n
Aside	aside		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n		🟥 0..n																			🟥 0..n			🟩 0..n								🟩 0..n	🟩 0..n	🟩 0..n
P	p		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n																			🟥 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Hn	h1-h6/ H7=p		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n																							🟩 0..n			🟥 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
H	h1-h6		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n																							🟩 0..n			🟥 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Title	p		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n																																		
Sub	span				🟩 0..n					🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n			🟩 0..n	🟩 0..n	🟧 0..n		🟩 0..n	🟩 0..n	🟩 0..n			🟩 0..n								🟩 0..n		🟩 0..n
Lbl	label/div				🟩 0..n		🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n		🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	0..n	🟩 0..n				🟧				🟥 0..n		🟩 0..n				🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Em	em							🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n			🟩 0..n	🟩 0..n	🟧 0..n		🟩 0..n	🟩 0..n	🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Strong	strong							🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n			🟩 0..n	🟩 0..n	🟧 0..n		🟩 0..n	🟩 0..n	🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Span	span							🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n			🟩 0..n	🟩 0..n	🟧 0..n		🟩 0..n	🟩 0..n	🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Link	a		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n		0..n			🟩 0..n	🟩 0..n	🟧 0..n		🟩 0..n	🟩 0..n	🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Annot	??		0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n	0..n			0..n	0..n	0..n		0..n	0..n	0..n			0..n			0..n	0..n				0..n	0..n	0..n
Form	form		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n					🟥 0..n	🟥 0..n	🟧 0..n		🟥 0..n	🟥 0..n	🟥 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟥 0..n	🟩 0..n	🟩 0..n
Ruby	ruby							🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n									🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
RB	rb																			🟩 [a]																				
RT	rt																			🟩 [a]																				
RP	rp																			🟩 [a]																				
Warichu	span							🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	0..n									🟩 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
WT	span																							🟩 [b]																
WP	span																							🟩 [b]																
FENote	p				🟩 0..n	🟩 0..n	🟩 0..n			🟥 0..n																			🟩 0..n	🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
L	ul/ol		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n																	🟥 0..n	🟥 0..n		🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
LI	li																											🟩 0..n												
LBody	div																												🟩 0..n											
Table	table		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n																			🟥 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟥 0..n	🟩 0..n	🟩 0..n
TR	tr																														🟧 0..n				🟩 0..n	🟩 0..n	🟩 0..n			
TH	th																															🟩 0..n								
TD	td																															🟩 0..n								
THead	thead																														🟧 0..1									
TBody	tbody																														🟧 0..n									
TFoot	tfoot																														🟧 0..1									
Caption	ption/figcaption				🟥 0..1	🟥 0..1	🟥 0..1	🟥 0..1											🟥 0..1									🟥 0..1		🟥 0..1	🟩 0..1								🟩 0..1	🟩 0..1
Figure	figure		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟧 0..n	0..n									🟥 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n
Formula	figure		🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟩 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟥 0..n	🟧 0..n	0..n									🟥 0..n			🟩 0..n			🟩 0..n	🟩 0..n				🟩 0..n	🟩 0..n	🟩 0..n

Resources

- https://github.com/Normex/PDF-Derivation
- Sample files
  - Styling
  - Associated files
  - Forms
  - Interactive
  - Fail cases
- Implementation (commandline, GUI)

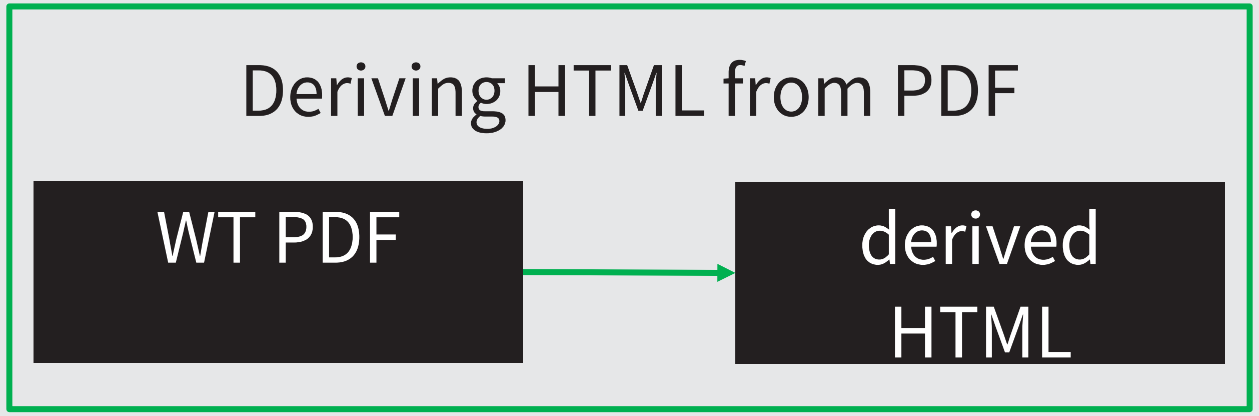
Problems - general

- Bad tagging (wrong attributes, content vs. struct tree)
- No PDF 2.0 support yet
- PDF/UA isn’t Well Tagged = what is Well Tagged?
  - reusable pdfs also accessible ?
- No PDF/UA-2 yet
- Export doesn’t produce Tagged PDFs (changing !!)
- No tools for manual tagging

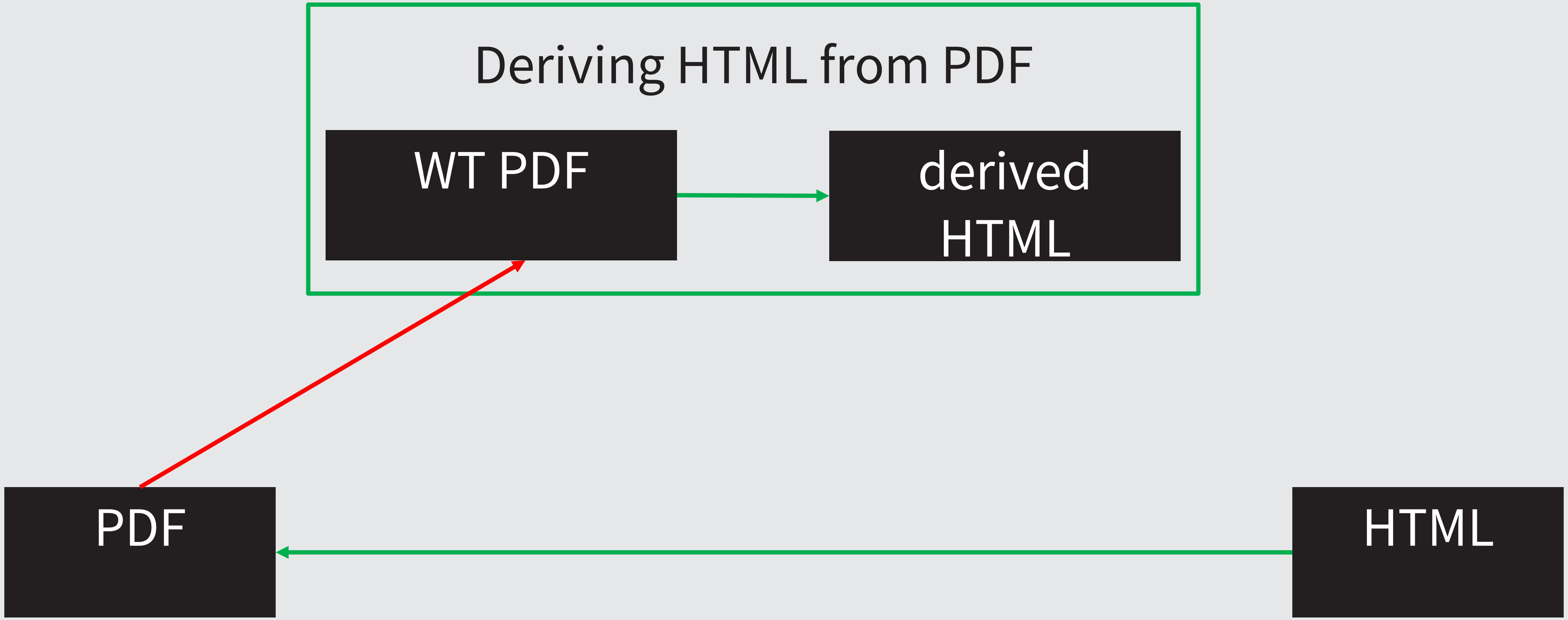
Problem - technical

- Elements (TOC)
- Attributes (Layout vs. CSS)
- Links across pages
- Forms, Javascript
- Annotations

HTML vs. PDF


Deriving HTML from PDF

HTML vs. PDF



Deriving HTML from PDF

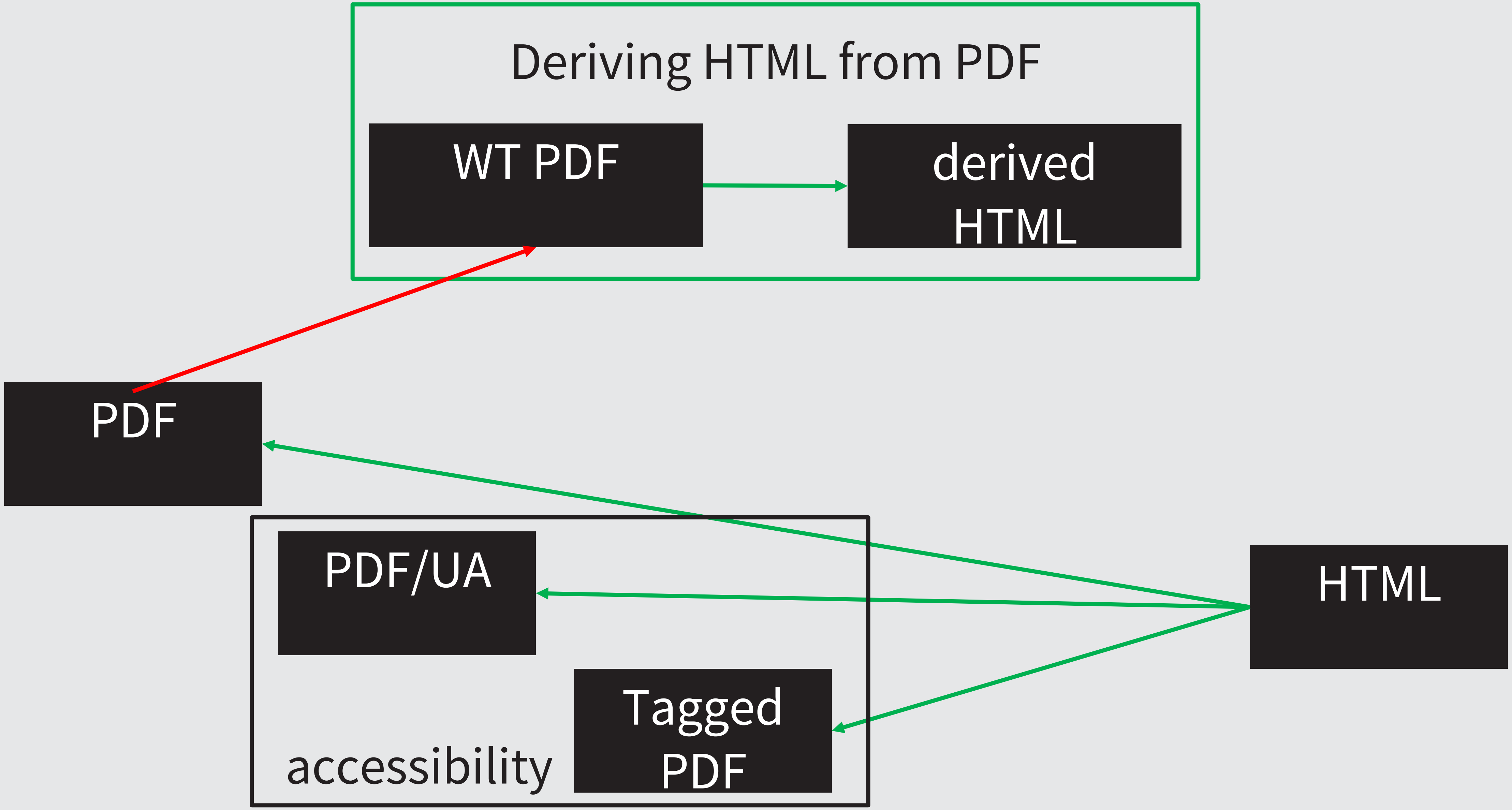
WT PDF

derived HTML

PDF

HTML

HTML vs. PDF



Deriving HTML from PDF

WT PDF

derived HTML

PDF

PDF/UA

Tagged PDF

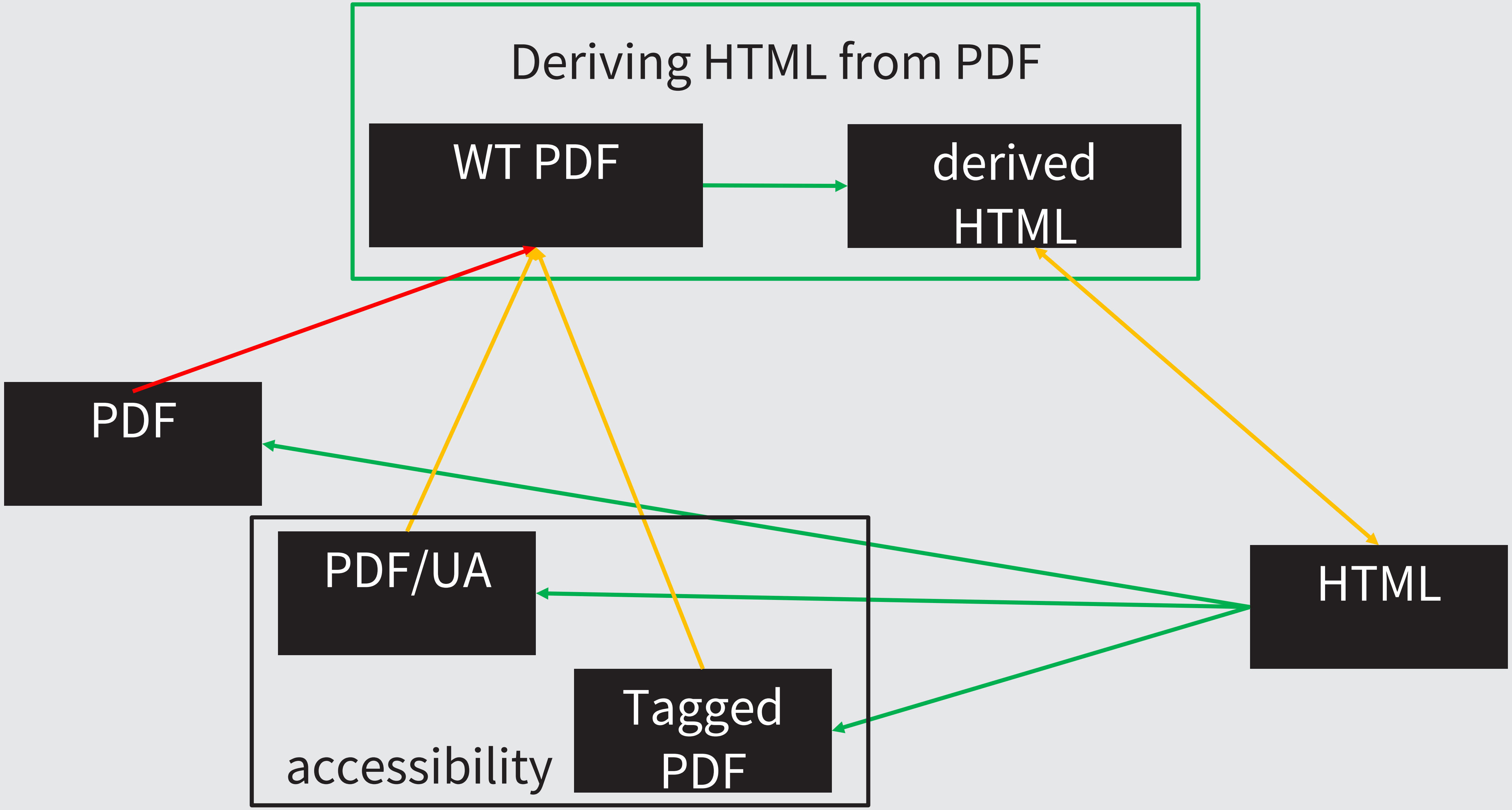
accessibility

HTML

Copyright © 2021, PDF Association

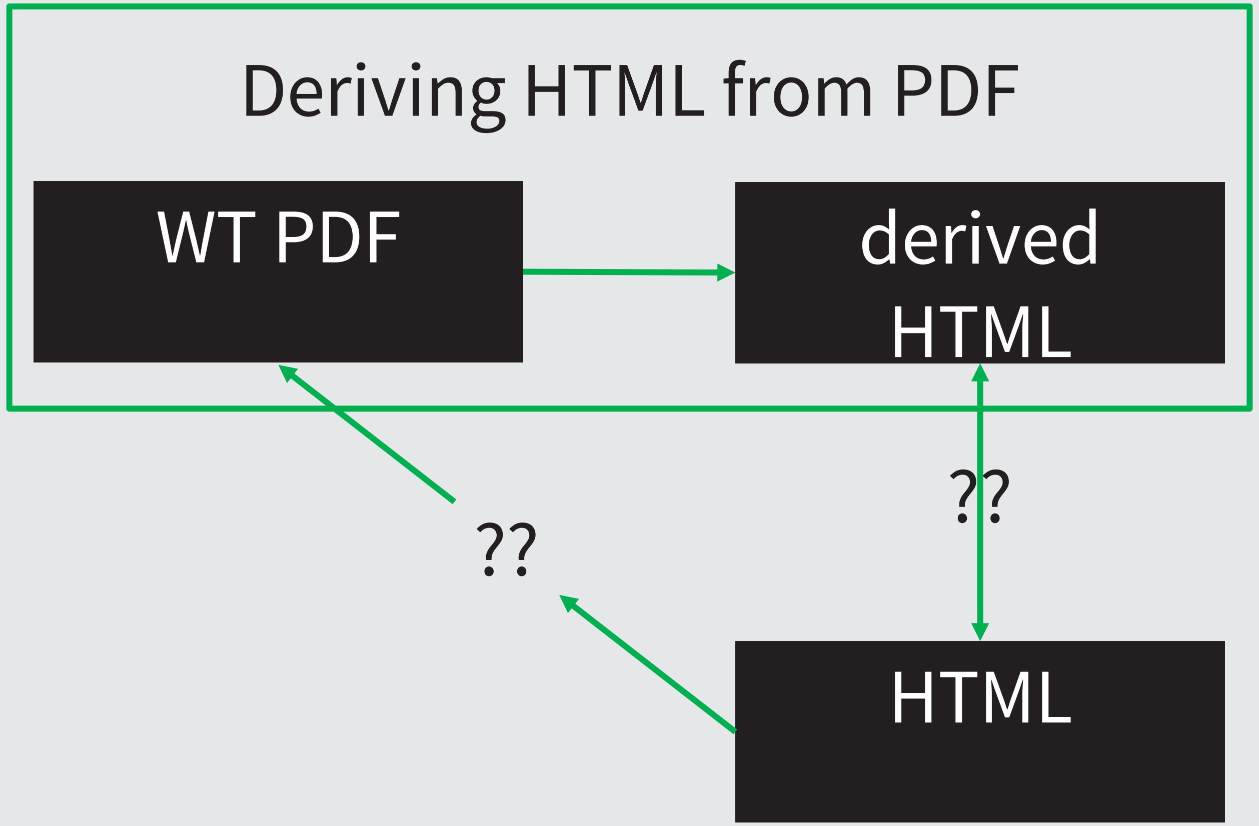
HTML vs. PDF

PDF Days Online 2021



HTML vs. PDF

PDF Days Online 2021



Deriving HTML from PDF

WT PDF

derived HTML

??

??

HTML

NORMEX

Foxit

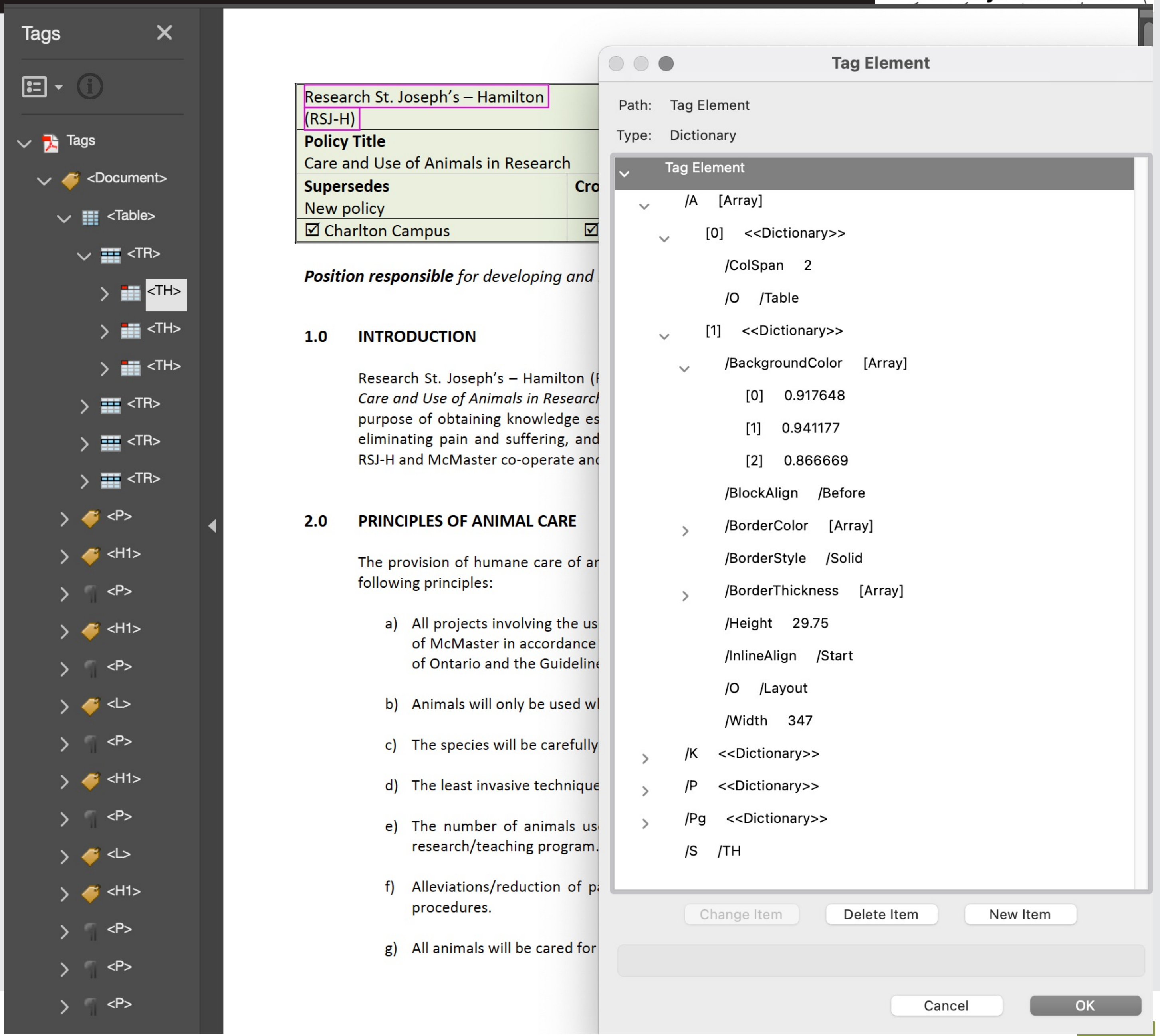
HTML 2 PDF

- Not a one to one conversion.
  - <div>, <span> → Div, Span
  - <b>, <i> → Span and loosing information about tag
  - <ul>, <ol> → L (ListNumbering) and we need to add Lbl that doesn’t exist in html
  - <header>, <time> → loosing semantic, no attributes to represent such
- Complex problems
  - Interactive elements
  - Javascript
  - css

HTML 2 PDF

PDF Days Online 2021

- As many Layout attributes as possible



Tags

```
Tags
  <Document>
    <Table>
      <TR>
        <TH>
        <TH>
        <TH>
      <TR>
      <TR>
      <TR>
    <P>
    <H1>
    <P>
    <H1>
    <P>
    <L>
    <P>
    <H1>
    <P>
    <L>
    <H1>
    <P>
    <P>
    <P>
```

Research St. Joseph's – Hamilton (RSJ-H)		
Policy Title Care and Use of Animals in Research		
Supersedes New policy	**Cro**	
☑ Charlton Campus	☑	

***Position responsible** for developing and*

1.0 INTRODUCTION

Research St. Joseph's – Hamilton (I *Care and Use of Animals in Researc* purpose of obtaining knowledge es eliminating pain and suffering, and RSJ-H and McMaster co-operate an

2.0 PRINCIPLES OF ANIMAL CARE

The provision of humane care of ar following principles:

a) All projects involving the us of McMaster in accordance of Ontario and the Guidelin

b) Animals will only be used wl

c) The species will be carefully

d) The least invasive technique

e) The number of animals us research/teaching program.

f) Alleviations/reduction of p procedures.

g) All animals will be cared for

Tag Element

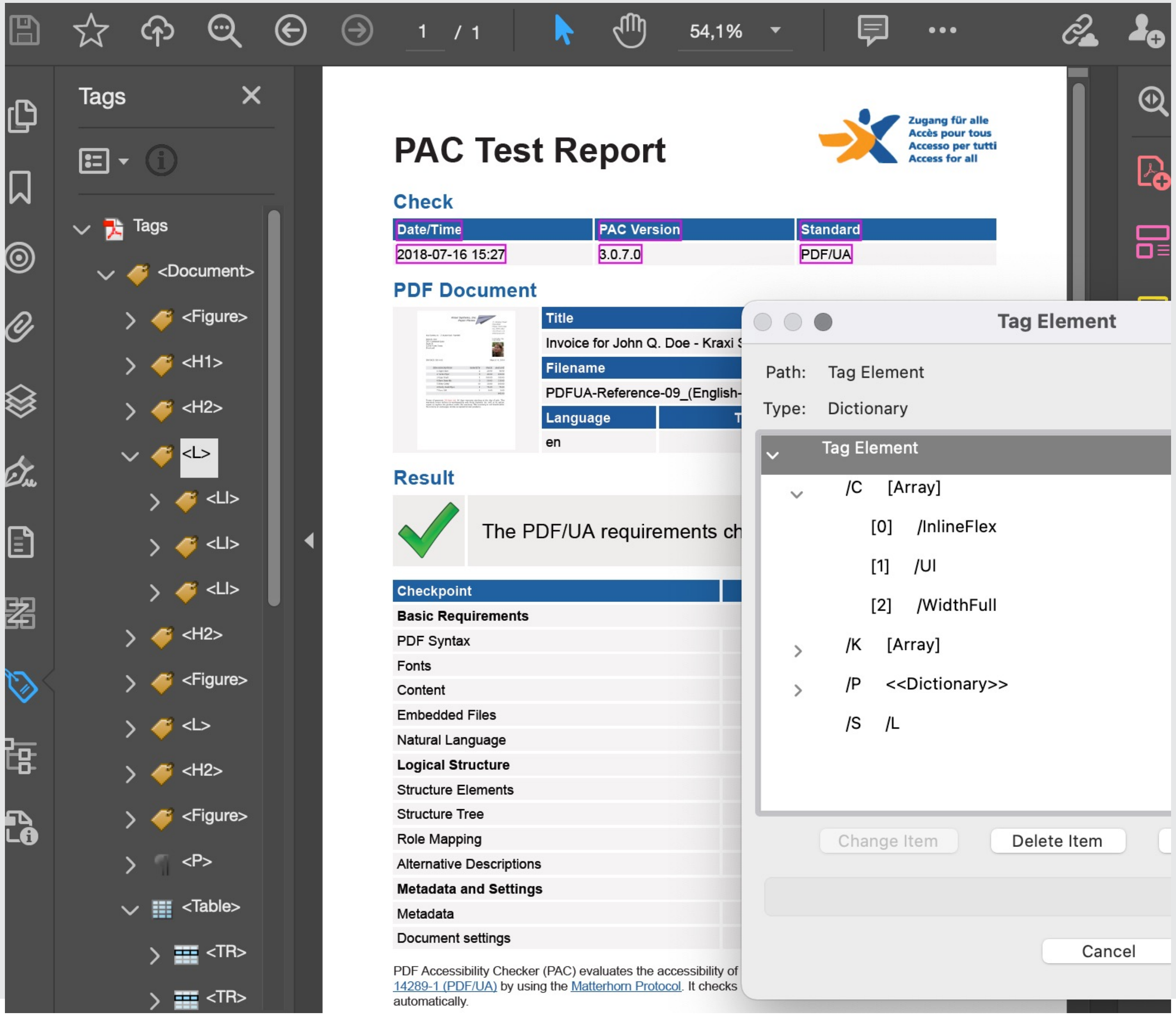
Path: Tag Element

Type: Dictionary

```
Tag Element
  /A  [Array]
    [0]  <<Dictionary>>
      /ColSpan  2
      /O  /Table
    [1]  <<Dictionary>>
      /BackgroundColor  [Array]
        [0]  0.917648
        [1]  0.941177
        [2]  0.866669
      /BlockAlign  /Before
      /BorderColor  [Array]
      /BorderStyle  /Solid
      /BorderThickness  [Array]
      /Height  29.75
      /InlineAlign  /Start
      /O  /Layout
      /Width  347
  /K  <<Dictionary>>
  /P  <<Dictionary>>
  /Pg  <<Dictionary>>
  /S  /TH
```

HTML 2 PDF

- As many Layout attributes as possible
- 2.0 allows adding CSS, HTML attributes



```
Tags
  <Document>
    <Figure>
    <H1>
    <H2>
    <L>
      <LI>
      <LI>
      <LI>
    <H2>
    <Figure>
    <L>
    <H2>
    <Figure>
    <P>
    <Table>
      <TR>
      <TR>
```

PAC Test Report

Zugang für alle
Accès pour tous
Accesso per tutti
Access for all

Check

Date/Time	PAC Version	Standard
2018-07-16 15:27	3.0.7.0	PDF/UA

PDF Document

Title
Invoice for John Q. Doe - Kraxi S
Filename
PDFUA-Reference-09_(English-
Language
en

Result

✓ The PDF/UA requirements ch

Checkpoint
Basic Requirements
PDF Syntax
Fonts
Content
Embedded Files
Natural Language
Logical Structure
Structure Elements
Structure Tree
Role Mapping
Alternative Descriptions
Metadata and Settings
Metadata
Document settings

PDF Accessibility Checker (PAC) evaluates the accessibility of 14289-1 (PDF/UA) by using the Matterhorn Protocol. It checks automatically.

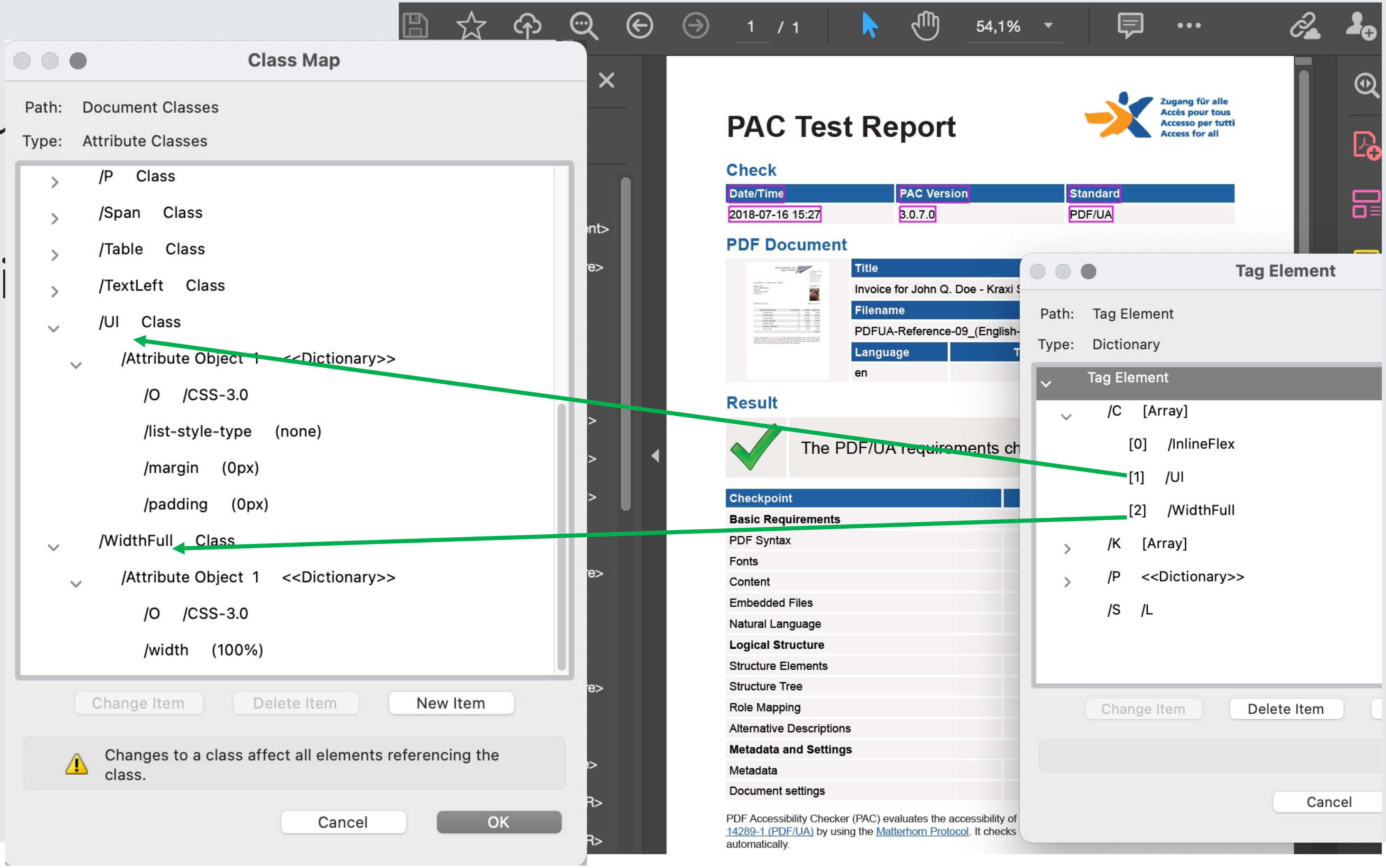
Tag Element

Path: Tag Element
Type: Dictionary

```
Tag Element
  /C   [Array]
       [0]  /InlineFlex
       [1]  /UI
       [2]  /WidthFull
  /K   [Array]
  /P   <<Dictionary>>
  /S   /L
```

HTML 2 PDF

- As many Layou
  possible
- 2.0 allows addi
  attributes



Class Map

Path: Document Classes

Type: Attribute Classes

```
/P  Class
/Span  Class
/Table  Class
/TextLeft  Class
/UI  Class
  /Attribute Object 1  <<Dictionary>>
    /O  /CSS-3.0
    /list-style-type  (none)
    /margin  (0px)
    /padding  (0px)
/WidthFull  Class
  /Attribute Object 1  <<Dictionary>>
    /O  /CSS-3.0
    /width  (100%)
```

⚠ Changes to a class affect all elements referencing the class.

PAC Test Report

Check

Date/Time	PAC Version	Standard
2018-07-16 15:27	3.0.7.0	PDF/UA

PDF Document

Title: Invoice for John Q. Doe - Kraxi S

Filename: PDFUA-Reference-09_(English-

Language: en

Result

✓ The PDF/UA requirements ch

Checkpoint

Basic Requirements
PDF Syntax
Fonts
Content
Embedded Files
Natural Language
Logical Structure
Structure Elements
Structure Tree
Role Mapping
Alternative Descriptions
Metadata and Settings
Metadata
Document settings

PDF Accessibility Checker (PAC) evaluates the accessibility of 14289-1 (PDF/UA) by using the Matterhorn Protocol. It checks automatically.

Tag Element

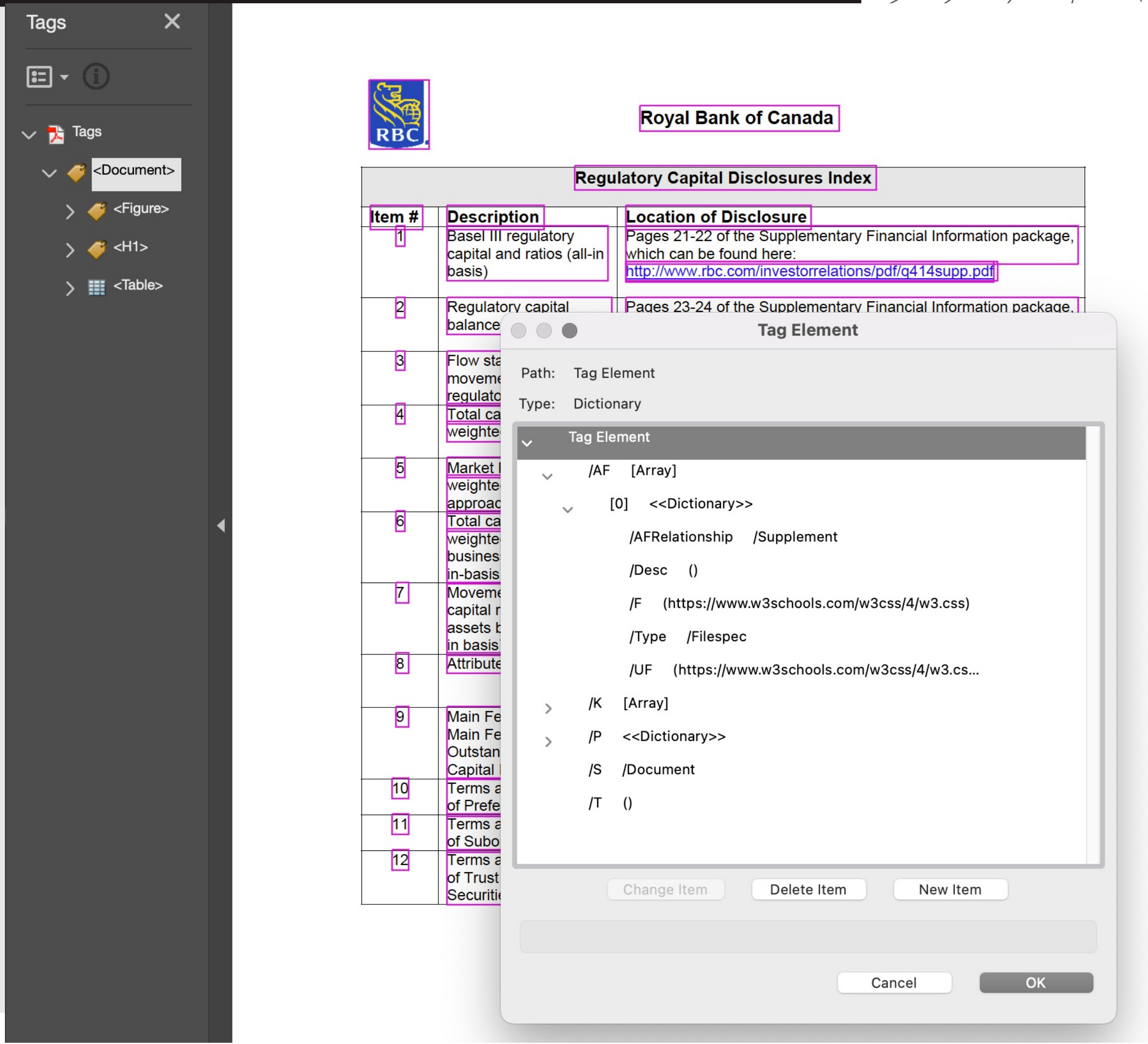
Path: Tag Element

Type: Dictionary

```
Tag Element
  /C  [Array]
    [0]  /InlineFlex
    [1]  /UI
    [2]  /WidthFull
  /K  [Array]
  /P  <<Dictionary>>
  /S  /L
```

HTML 2 PDF

- As many Layout attributes as possible
- 2.0 allows adding CSS, HTML attributes
- URL binding



Tags

- Tags
  - <Document>
    - <Figure>
    - <H1>
    - <Table>

Royal Bank of Canada

Regulatory Capital Disclosures Index		
Item #	Description	Location of Disclosure
1	Basel III regulatory capital and ratios (all-in basis)	Pages 21-22 of the Supplementary Financial Information package, which can be found here: http://www.rbc.com/investorrelations/pdf/q414supp.pdf
2	Regulatory capital balance	Pages 23-24 of the Supplementary Financial Information package.
3	Flow sta moveme regulato	
4	Total ca weighte	
5	Market weighte approac	
6	Total ca weighte busines in-basis	
7	Moveme capital assets b in basis	
8	Attribute	
9	Main Fe Main Fe Outstan Capital	
10	Terms a of Prefe	
11	Terms a of Subo	
12	Terms a of Trust Securiti	

Tag Element

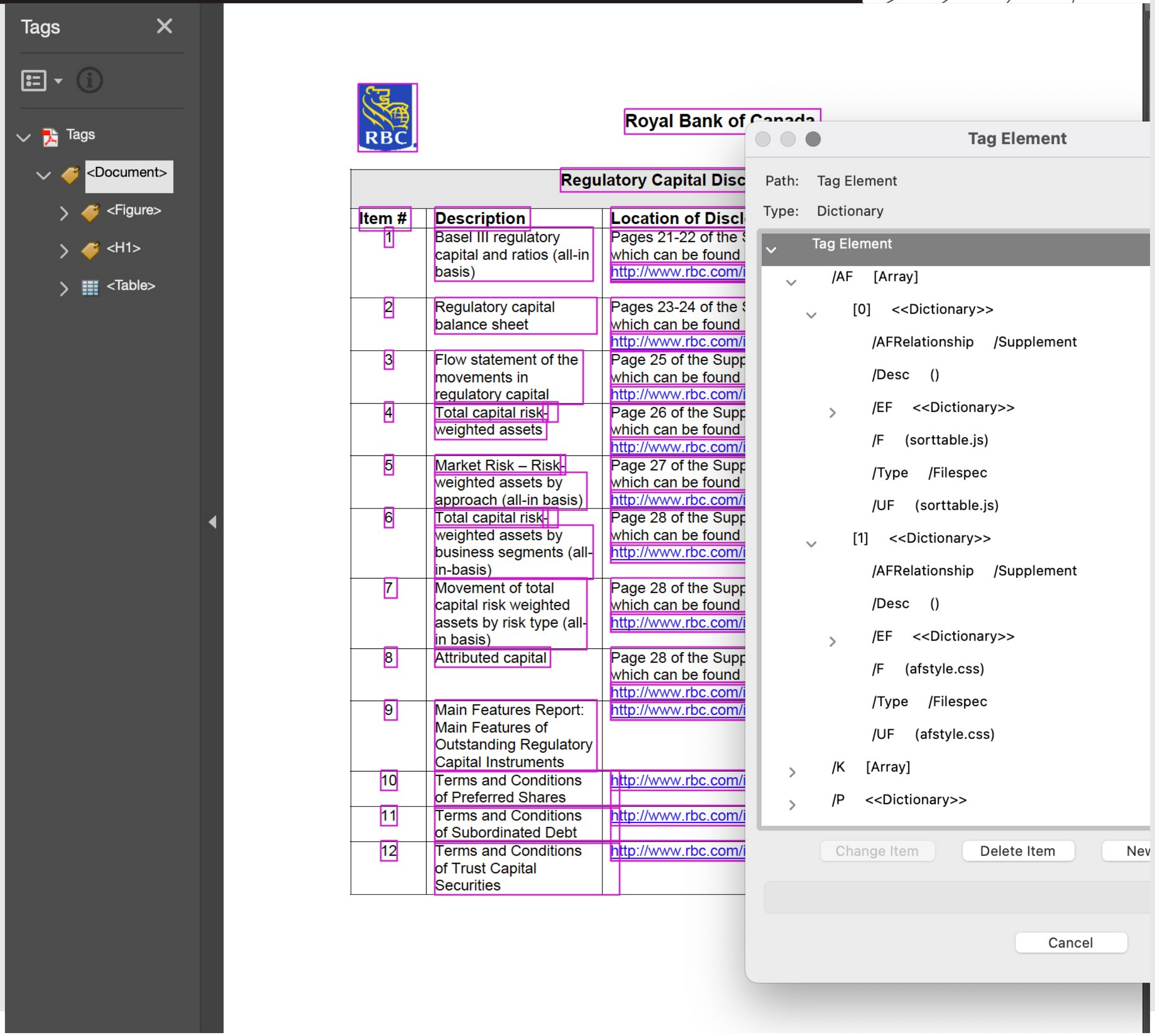
Path: Tag Element

Type: Dictionary

```
Tag Element
  /AF     [Array]
    [0]   <<Dictionary>>
      /AFRelationship   /Supplement
      /Desc    ()
      /F    (https://www.w3schools.com/w3css/4/w3.css)
      /Type    /Filespec
      /UF    (https://www.w3schools.com/w3css/4/w3.cs...
  /K    [Array]
  /P    <<Dictionary>>
  /S    /Document
  /T    ()
```

HTML 2 PDF

- As many Layout attributes as possible
- 2.0 allows adding CSS, HTML attributes
- URL binding
- Associated files



Royal Bank of Canada

Regulatory Capital Disc

Item #	Description	Location of Discl
1	Basel III regulatory capital and ratios (all-in basis)	Pages 21-22 of the S which can be found http://www.rbc.com/i
2	Regulatory capital balance sheet	Pages 23-24 of the S which can be found http://www.rbc.com/i
3	Flow statement of the movements in regulatory capital	Page 25 of the Supp which can be found http://www.rbc.com/i
4	Total capital risk-weighted assets	Page 26 of the Supp which can be found http://www.rbc.com/i
5	Market Risk – Risk-weighted assets by approach (all-in basis)	Page 27 of the Supp which can be found http://www.rbc.com/i
6	Total capital risk-weighted assets by business segments (all-in-basis)	Page 28 of the Supp which can be found http://www.rbc.com/i
7	Movement of total capital risk weighted assets by risk type (all-in basis)	Page 28 of the Supp which can be found http://www.rbc.com/i
8	Attributed capital	Page 28 of the Supp which can be found http://www.rbc.com/i http://www.rbc.com/i
9	Main Features Report: Main Features of Outstanding Regulatory Capital Instruments	
10	Terms and Conditions of Preferred Shares	http://www.rbc.com/i
11	Terms and Conditions of Subordinated Debt	http://www.rbc.com/i
12	Terms and Conditions of Trust Capital Securities	http://www.rbc.com/i

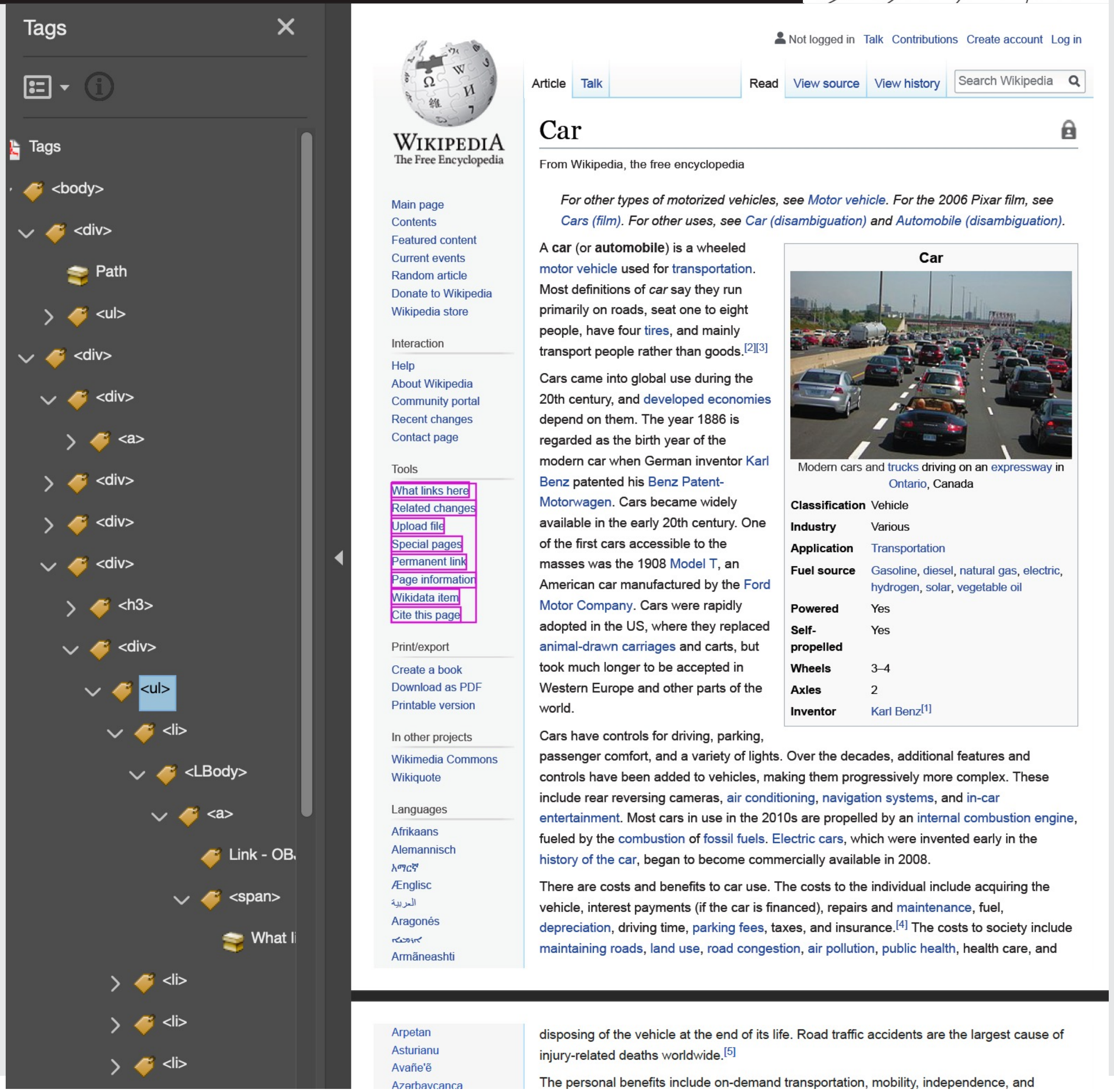
Tag Element

Path: Tag Element
Type: Dictionary

```
Tag Element
  /AF    [Array]
    [0]    <<Dictionary>>
      /AFRelationship    /Supplement
      /Desc    ()
      /EF    <<Dictionary>>
      /F    (sorttable.js)
      /Type    /Filespec
      /UF    (sorttable.js)
    [1]    <<Dictionary>>
      /AFRelationship    /Supplement
      /Desc    ()
      /EF    <<Dictionary>>
      /F    (afstyle.css)
      /Type    /Filespec
      /UF    (afstyle.css)
  /K    [Array]
  /P    <<Dictionary>>
```

HTML 2 PDF

- As many Layout attributes as possible
- 2.0 allows adding CSS, HTML attributes
- URL binding
- Associated files
- HTML Namespace



Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
Wikimedia Commons
Wikiquote

Languages
Afrikaans
Alemannisch
አማርኛ
Ænglisc
العربية
Aragonés
ܐܪܡܝܐ
Armãneashti
Arpetan
Asturianu
Avañe'ẽ
Azərbaycanca

Car

From Wikipedia, the free encyclopedia

For other types of motorized vehicles, see Motor vehicle. For the 2006 Pixar film, see Cars (film). For other uses, see Car (disambiguation) and Automobile (disambiguation).

A **car** (or **automobile**) is a wheeled motor vehicle used for transportation. Most definitions of *car* say they run primarily on roads, seat one to eight people, have four tires, and mainly transport people rather than goods.[2][3]

Cars came into global use during the 20th century, and developed economies depend on them. The year 1886 is regarded as the birth year of the modern car when German inventor Karl Benz patented his Benz Patent-Motorwagen. Cars became widely available in the early 20th century. One of the first cars accessible to the masses was the 1908 Model T, an American car manufactured by the Ford Motor Company. Cars were rapidly adopted in the US, where they replaced animal-drawn carriages and carts, but took much longer to be accepted in Western Europe and other parts of the world.

Car

Modern cars and trucks driving on an expressway in Ontario, Canada

Classification	Vehicle
Industry	Various
Application	Transportation
Fuel source	Gasoline, diesel, natural gas, electric, hydrogen, solar, vegetable oil
Powered	Yes
Self-propelled	Yes
Wheels	3–4
Axles	2
Inventor	Karl Benz[1]

Cars have controls for driving, parking, passenger comfort, and a variety of lights. Over the decades, additional features and controls have been added to vehicles, making them progressively more complex. These include rear reversing cameras, air conditioning, navigation systems, and in-car entertainment. Most cars in use in the 2010s are propelled by an internal combustion engine, fueled by the combustion of fossil fuels. Electric cars, which were invented early in the history of the car, began to become commercially available in 2008.

There are costs and benefits to car use. The costs to the individual include acquiring the vehicle, interest payments (if the car is financed), repairs and maintenance, fuel, depreciation, driving time, parking fees, taxes, and insurance.[4] The costs to society include maintaining roads, land use, road congestion, air pollution, public health, health care, and disposing of the vehicle at the end of its life. Road traffic accidents are the largest cause of injury-related deaths worldwide.[5]

The personal benefits include on-demand transportation, mobility, independence, and

HTML 2 PDF

- As many Layout attributes as possible
- 2.0 allows adding CSS, HTML attributes
- URL binding
- Associated files
- HTML Namespace
- Forms

Lessons learned - problems

PDF Days Online 2021

- HTML – small, PDF - big
- Invalid html
- Incompatible tagsets
- complex html, css, js frameworks, media queries etc..
- Interactivity (forms, fieldsets - import/export)
- Online vs. offline (in html world we reference everything by URI)
- Hard to include multiple targets (UA, WT PDF – or mobile vs. desktop)

Landscape change

- WT PDF technical working group
- Forms technical working group
- AnnexL update
- 1.7 and 2.0 Namespaces

- PDF/UA-2

Join us

- Implementation
  - https://github.com/Normex/PDF-Derivation
- Samples
  - https://normex.github.io/PDF-Derivation/
- PDF Association TWG
  - Derivation TWG
  - WT PDF TWG
  - PDF/UA TWG

?

Thanks !

Roman Toda, Normex

https://github.com/Normex/PDF-Derivation