



# Deriving HTML from PDF

How reusable content in pdf could be

Roman Toda, Normex



electronic  
document  
CONFERENCE

# 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
  - First time used different presentation model for assistive technology



# Problems 2019

- PDF is essential part of the web – but is it really?
- HTML users hate pdf
  - How to use pdfs on mobiles
  - In browsers
- HTML developers hate pdf
  - Can't control the user experience
  - Can't access/navigate pdf content



# Derivation

- Is it possible to deterministically interpret PDF content in the web based environment?
- Yes. **Tagged PDF** is the best way for capturing author's intent
  - provisions in ISO 32000-2 are rich enough to be unambiguously interpreted in html = **Derived**
- **Derivation** our new word for **rendering**



# Join us

- Download and read the publication
- PDF Association
- Next-Generation PDF working group
- Test our MVP
  - <https://github.com/Normex/PDF-Derivation>
  - Share your files
  - Discuss techniques
  - Give feedback



# What is Tagged PDF

- Structure tree
  - Marked content
  - Forms fields, links, annotations
- Attributes
- Classes
- Associated files
- Actions



# What is Tagged PDF

2018-07-16 15:27 3.0.7.0 PDF/UA


### PDF Document

**Title**  
Invoice for John Q. Doe - Kraxi Systems, Inc.

**Filename**  
PDFUA-Reference-09\_(English-invoice) - Copy.pdf

Language	Tags	Pages	Size
en	79	1	120 KB

### Result

 The PDF/UA requirements checked by PAC are fulfilled.

Checkpoint	Passed	Warned	Failed
<b>Basic Requirements</b>			
PDF Syntax	85	0	0
Fonts	8	0	0
Content	586	0	0
Embedded Files	0	0	0
Natural Language	226	0	0
<b>Logical Structure</b>			
Structure Elements	24	0	0
Structure Tree	158	0	0

Tags

- <Figure>
- <L>
- <H2>
- <Figure>
- <P>
- <Table>
- <TR>
- <TR>
- <TR>
- PDF Syntax
- <TD>
- <TD>
- <TD>
- <TR>



# What is Tagged PDF

2018-07-16 15:27 3.0.7.0 PDF/UA

### PDF Document

Title	
Invoice for John Q. Doe - Kraxi Systems, Inc.	
Filename	
PDFUA-Reference-09_(English-invoice) - Cop	
Language	Tags
en	79

### Result

The PDF/UA requirements checked by

Checkpoint	Passed
<b>Basic Requirements</b>	
PDF Syntax	85
Fonts	8
Content	586
Embedded Files	0
Natural Language	226
<b>Logical Structure</b>	
Structure Elements	24
Structure Tree	158

### Tag Element

- /A <<Dictionary>>
  - /O /Layout
  - /Placement /Box
  - /SpaceAfter 12.649994
- /AF [Array]
  - [0] <<Dictionary>>
    - /AFRelationship /Supplement
  - /EF <<Dictionary>>
    - /F (internal\_style\_3.css)
    - /Type /Filespec
    - /UF (internal\_style\_3.css)
- /K [Array]
- /P <<Dictionary>>
  - /S /Table
  - /Type /StructElem
- /C [Array]
  - [0] /Floating
  - [1] /Retro





# Standard structure element types

- Chapter 14.8.4 in ISO 32000-2
  - Annot, Artifact, Aside, Caption, Document, DocumentFragment, Div, Em, FEnote, Figure, Form, Formula, H1..Hn, L, Lbl, LBody, LI, Link, P, RB, RP, RT, Ruby, Span, Strong, Sub, Table, TBody, TD, TFoot, TH THead, Title, TR, Warichu, WT, WP
- Some sound like html tags: Div, Span, P, H1, LI
- Some are harder to imagine: Annot
- RoleMap may be involved



# Text

The screenshot shows a PDF editor interface. On the left, a 'Tags' panel displays a tree structure with elements like <Sect>, <P>, <H1>, and <P>. The main document area shows the title 'The Portable Document Fo' and a section titled 'Imaging model' which is highlighted with a blue selection box. A red arrow points from this selection to a code block on the right. A 'Tag Element' dialog box is open in the foreground, showing the following details:

- Path: Tag Element
- Type: Dictionary
- Tag Element (selected)
- /K 1
- /Lang (EN-US)
- /P <<Dictionary>>
- /Pg <<Dictionary>>
- /S /H1

Buttons at the bottom of the dialog include 'Change Item', 'Delete Item', 'New Item', 'Options', 'OK', and 'Cancel'.

```
<h1 lang="EN-US">  
    <span>Imaging model</span>  
</h1>
```



# Simple structure - Text

The image illustrates the process of generating HTML from a PDF document. It shows three main components:

- PDF Editor:** A 'Tag Element' dialog box is open, showing a tree structure of tags. The 'Tag Element' is selected, and its properties are listed: Path: Tag Element, Type: Dictionary, /K 1, /Lang (EN-US), /P <<Dictionary>>, /Pg <<Dictionary>>, /S /H1. Buttons for 'Change Item', 'Delete Item', 'New Item', 'Options', 'OK', and 'Cancel' are visible.
- PDF Document:** The document page is titled 'The Portable Document Fo...'. The text 'Imaging model' is highlighted with a blue selection box. A red arrow points from this box to the HTML code.
- HTML Output:** A browser window displays the rendered HTML. The code is: 

```
<h1 lang="EN-US">
  <span>Imaging model</span>
</h1>
```

 A red arrow points from this code to the browser window below.
- Browser Window:** The browser window shows the rendered HTML output. The title is 'Imaging model'. The content is: 

The basic design of how graphics are represented in PDF is very similar to that of PostScript, exc...  
The current transformation matrix (CTM), which determines the coordinate system



# Table

• The alpha constant, which is a key component of transparency

### Raster images

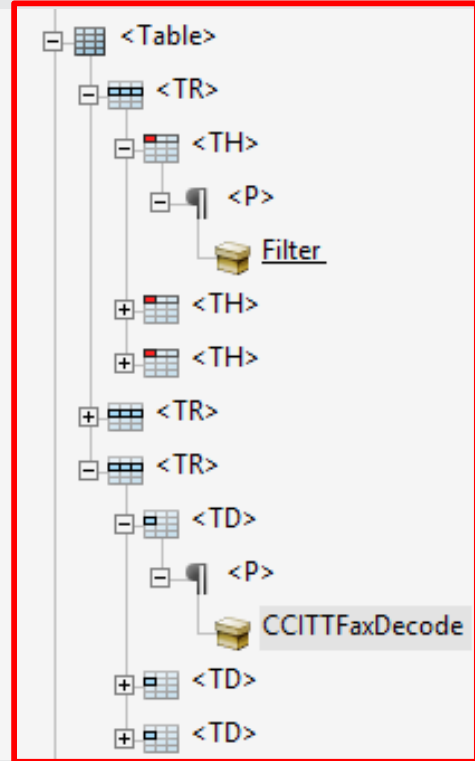
Raster images in PDF (called Image XObjects) are represented by dictionaries with an associated stream.

Filter	Extension	Description
DC1Decode	.jpg	a lossy filter based on the JPEG standard
CCITTFaxDecode	.tiff	a lossless bi-level (black/white) filter based on the Group 3 or Group 4 CCITT (ITU-T) fax compression standard defined in ITU-T T.4 and T.6

**PDF association MEMBER**



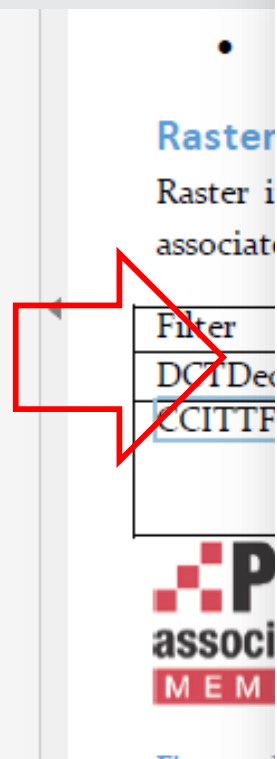
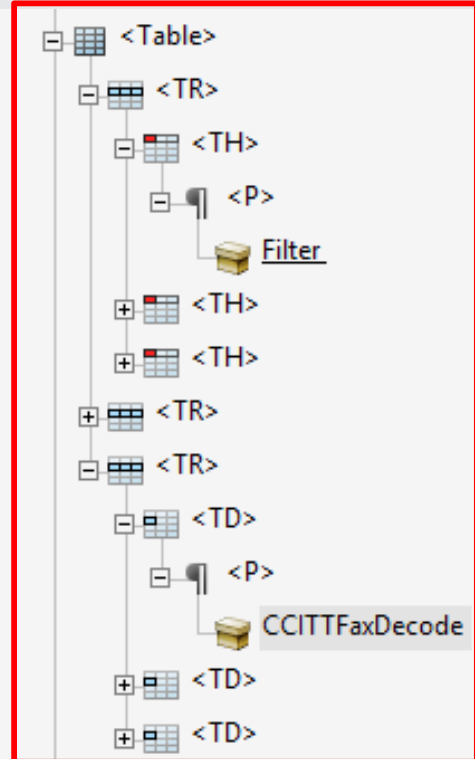
# Table - derivation



```
<table>
  <tr>
    <th>
      <p lang="EN-US">
        <span>Filter </span>
      </p>
    </th>
    <!--.....-->
  </tr>
  <!--.....-->
  <tr>
    <td>
      <p lang="EN-US">
        <span>CCITTFaxDecode </span>
      </p>
    </td>
    <!--.....-->
  </tr>
</table>
```



# Table - derivation



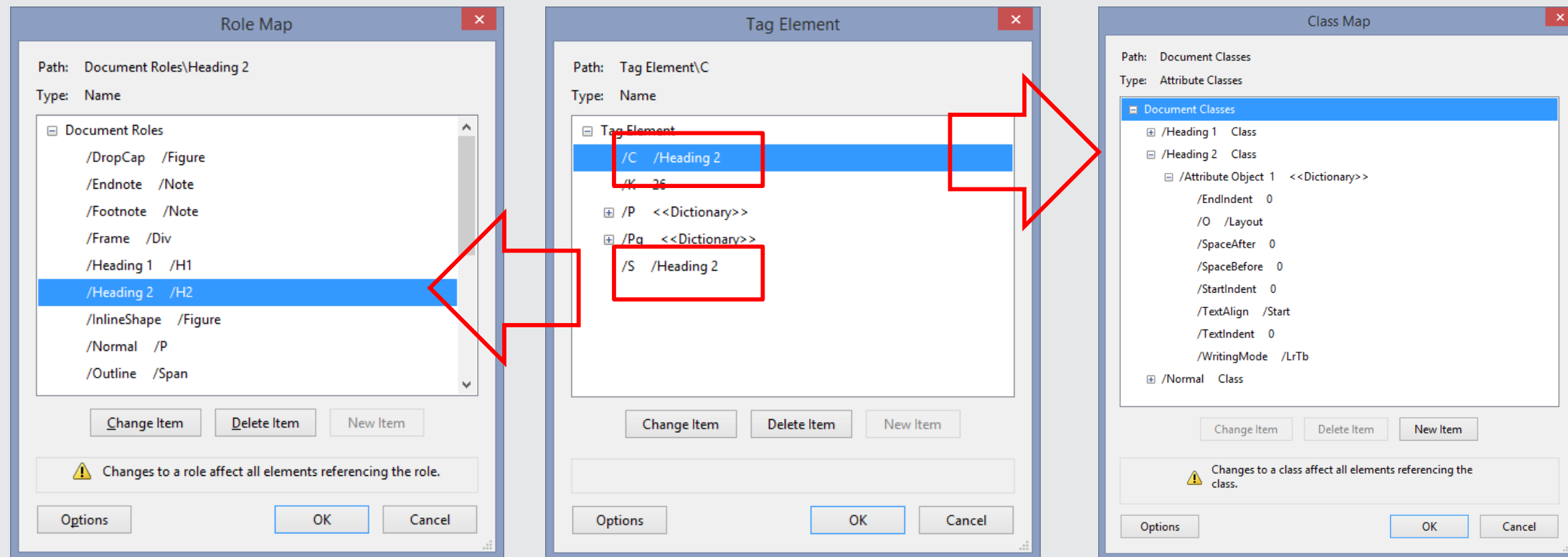
```
<table>
  <tr>
    <th>
      <p lang="EN-US">
        <span>Filter </span>
      </p>
    </th>
    <td>
      <p>
        Raster images in PDF (called Image XObjects) are represented by dictionaries with an associated stream.
      </p>
      <table>
        <thead>
          <tr>
            <th>Filter</th>
            <th>Extension</th>
            <th>Description</th>
          </tr>
        </thead>
        <tbody>
          <tr>
            <td>DCTDecode</td>
            <td>.jpg</td>
            <td>a lossy filter based on the JPEG standard</td>
          </tr>
          <tr>
            <td>CCITTFaxDecode</td>
            <td>.tiff</td>
            <td>a lossless bi-level (black/white) filter based on the Group 3 or Group 4 CCITT (ITU-T) fax compression standard defined in ITU-T T.4 and T.6</td>
          </tr>
        </tbody>
      </table>
    </td>
  </tr>
</table>
```

Filter	Extension	Description
DCTDecode	.jpg	a lossy filter based on the JPEG standard
CCITTFaxDecode	.tiff	a lossless bi-level (black/white) filter based on the Group 3 or Group 4 CCITT (ITU-T) fax compression standard defined in ITU-T T.4 and T.6



# Attributes – general concept

- Each tag could have attributes
- Different practice is using ClassMap
- Role mapping takes place



# Attribute owners

Table 376: Standard structure attribute owners

Owner value for the attribute object's O entry	Description
Layout	Attributes governing the layout of content
List	Attributes governing the numbering of lists
PrintField	Attributes governing Form structure elements for non-interactive form fields
Table	Attributes governing the organization of cells in tables
Artifact	Attributes governing Artifact structure elements
XML-1.00	Additional attributes governing translation to XML, version 1.00
HTML-3.20	Additional attributes governing translation to HTML, version 3.20
HTML-4.01	Additional attributes governing translation to HTML, version 4.0
HTML-5.00	Additional attributes governing translation to HTML, version 5.0
OEB-1.00	Additional attributes governing translation to OEB (Open eBook), version 1.0
RTF-1.05	Additional attributes governing translation to Microsoft Rich Text Format, version 1.05
CSS-1.00	Additional attributes governing translation to a format using CSS, version 1.00
CSS-2.00	Additional attributes governing translation to a format using CSS, version 2.00
CSS-3.00	Additional attributes governing translation to a format using CSS, version 3.00
RDFa-1.10	Additional attributes governing translation to a format using RDFa version 1.1





# Attribute owners

Table 376: Standard structure attribute owners

Owner value for the attribute object's O entry	Description
Layout	Attributes governing the layout of content
List	Attributes governing the numbering of lists
PrintField	Attributes governing Form structure elements for non-interactive form fields
Table	Attributes governing the organization of cells in tables
Artifact	Attributes governing Artifact structure elements
XML-1.00	Additional attributes governing translation to XML, version 1.00
HTML-3.20	Additional attributes governing translation to HTML, version 3.20
HTML-4.01	Additional attributes governing translation to HTML, version 4.0
HTML-5.00	Additional attributes governing translation to HTML, version 5.0
OEB-1.00	Additional attributes governing translation to OEB (Open eBook), version 1.0
RTF-1.05	Additional attributes governing translation to Microsoft Rich Text Format, version 1.05
CSS-1.00	Additional attributes governing translation to a format using CSS, version 1.00
CSS-2.00	Additional attributes governing translation to a format using CSS, version 2.00
CSS-3.00	Additional attributes governing translation to a format using CSS, version 3.00
RDFa-1.10	Additional attributes governing translation to a format using RDFa version 1.1



# Attribute usage

The Portable

### Imaging model

The basic design of how gray except for the use of transparent

The current transformation

- The clipping path
- The color space
- The alpha constant,

### Raster images

Raster images in PDF (called associated stream.

Filter	Ext
DCTDecode	.jpg
CCITTFaxDecode	.tif

Attributes

Path: Attribute Objects\  
Type: Dictionary

- Attribute Objects
  - /Attribute Object 1 <<Dictionary>>
    - /O /CSS-3.00
    - /font-size (36px)
    - /text-align (center)
    - /color (blue)
    - /background-color (yellow)
    - /word-spacing (10px)

Change Item Delete Item New Item

- Layout
- Table
- List
- Can use the power of CSS



# Attribute usage

The Portable

## Imaging model

The basic design of how gray except for the use of transparent

The current transformation

- The clipping path
- The color space
- The alpha constant,

### Raster images

Raster images in PDF (called associated stream.

Filter	Ext
DCTDecode	.jpg
CCITTFaxDecode	.tif

Attributes

Path: Attribute Objects\  
Type: Dictionary

- Attribute Objects
  - /Attribute Object 1 <<Dictionary>>
    - /O /CSS-3.00
    - /font-size (36px)
    - /text-align (center)
    - /color (blue)
    - /background-color (yellow)
    - /word-spacing (10px)

Change Item Delete Item New Item

```
<h1 lang="EN-US" style="background-color: yellow; color: blue; font-size: 36px; text-align: center; word-spacing: 10px; ">
<span>Imaging model </span>
</h1>
```



# Attribute usage

The screenshot illustrates the relationship between PDF attributes and their rendered output. On the left, a PDF viewer shows the 'Attributes' panel for an 'Imaging model' section. The panel lists several attributes: /O (CSS-3.00), /font-size (36px), /text-align (center), /color (blue), /background-color (yellow), and /word-spacing (10px). A red arrow points from the 'Attributes' panel to a code editor window. The code editor displays the following HTML code:

```
<h1 lang="EN-US" style="background-color: yellow; color: blue; font-size: 36px; text-align: center; word-spacing: 10px; "><span>Imaging model </span></h1>
```

Another red arrow points from the code editor to a browser window. The browser window shows the rendered output of the HTML code, which is a yellow rectangular box containing the text 'Imaging model' in blue, centered, with a wide word-spacing. Below the browser window, the text from the PDF is visible: 'The basic design of how graphics are represented in PDF is very similar to that of PostScript, except for the use of transparency, which was added in PDF 1.4.'



# Associated files – css, js, svg, mathml

Tag Element

- /AF [Array]
  - [0] <<Dictionary>>
    - /AFRelationship /Supplement
    - /Desc ()
    - /EF <<Dictionary>>
      - /F (sortable.js)
      - /Type /Filespec
      - /UF (sortable.js)
  - [1] <<Dictionary>>
    - /AFRelationship /Supplement
    - /Desc ()
    - /EF <<Dictionary>>
      - /F (afstyle.css)
      - /Type /Filespec
      - /UF (afstyle.css)
- /K [Array]
- /P <<Dictionary>>
- /S /Document



Regulatory Capital Disclosures - Adobe Acrobat Pro DC

File Edit View Plug-Ins Window Help Colorado

Home Tools Regulatory Capital ... x

Sign In

Share

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: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
2	Regulatory capital balance sheet	Pages 23-24 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
3	Flow statement of the movements in regulatory capital	Page 25 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
4	Total capital risk-weighted assets	Page 26 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
5	Market Risk – Risk-weighted assets by approach (all-in basis)	Page 27 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
6	Total capital risk-weighted assets by business segments (all-in-basis)	Page 28 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
7	Movement of total capital risk weighted assets by risk type (all-in basis)	Page 28 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
8	Attributed capital	Page 28 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
9	Main Features Report: Main Features of Outstanding Regulatory Capital Instruments	<a href="http://www.rbc.com/investorrelations/pdf/capitalq414.pdf">http://www.rbc.com/investorrelations/pdf/capitalq414.pdf</a>

8,50 x 11,00 in



# Associated files – css, js, svg, mathml

The image shows a screenshot of Adobe Acrobat Pro DC displaying a PDF document titled "Regulatory Capital Disclosures - Adobe Acrobat Pro DC". The PDF content includes the Royal Bank of Canada logo and a table titled "Regulatory Capital Disclosures Index". The table has three columns: "Item #", "Description", and "Location of Disclosure".

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: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
8	Attributed capital	Page 28 of the Supplementary Financial Information package, which can be found here: <a href="http://www.rbc.com/investorrelations/pdf/q414supp.pdf">http://www.rbc.com/investorrelations/pdf/q414supp.pdf</a>
2	Flow statement of the movements in	Page 25 of the Supplementary Financial Information package, which can be found here:

The "Tag Element" window shows the following XML structure:

```
<pre>/AF [Array]
  [0] <<Dictionary>>
    /AFRelationship /Supplement
    /Desc ()
    + /EF <<Dictionary>>
      /F (sortable.js)
      /Type /Filespec
      /UF (sortable.js)
    [1] <<Dictionary>>
      /AFRelationship /Supplement
      /Desc ()
      + /EF <<Dictionary>>
        /F (afstyle.css)
        /Type /Filespec
        /UF (afstyle.css)
    + /K [Array]
    + /P <<Dictionary>>
      /S /Document</pre>
```

# Namespaces – sample with html

The image shows a screenshot of a Wikipedia article titled "Car". On the left side, there is a "Tags" sidebar with a list of HTML tags. The tags listed are: <body>, <div>, <div>, <div>, <div>, <div>, <div>, <div>, <div>, <div>, <h3>, <div>, <ul>, <li>, <li>, <li>, <li>, <li>, <li>. The main article content includes the Wikipedia logo, navigation links (Main page, Contents, etc.), and the start of the article text: "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]". There is also a small image of a modern car and a classification table.

Tags

- <body>
- <div>
- <div>
- <div>
- <div>
- <div>
- <div>
- <div>
- <div>
- <div>
- <h3>
- <div>
- <ul>
- <li>
- <li>
- <li>
- <li>
- <li>

WIKIPEDIA  
The Free Encyclopedia

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

Interaction

- Help
- About Wikipedia
- Community porta
- Recent changes
- Contact page

Tools

- What links here
- Related changes
- Upload file

Not logged in

Article Talk Read View source


## Car

From Wikipedia, the free encyclopedia

*For other types of motorized vehicles, see [Motor vehicles](#). For other uses, see [Car \(disambiguation\)](#).*

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



Modern cars

Classification  
Industry



# Namespaces – sample with html

The image shows a screenshot of a Wikipedia page with a 'Tag Element' tooltip. The tooltip displays the following HTML namespace information:

```
Tag Element
  /A <<Dictionary>>
    /O /HTML-5.00
    /class (hatnote navigation-not-searchable)
    /role (note)
  /K [Array]
  /NS <<Dictionary>>
    /NS (http://www.w3.org/1999/xhtml)
    /Type /Namespace
  /P <<Dictionary>>
    /S /div
    /Type /StructElem
```

The background shows the Wikipedia main page with a sidebar containing navigation links like 'Main page', 'Contents', and 'Featured content'. A 'Tags' sidebar on the left lists various HTML tags like <body>, <div>, <h3>, and <ul>. A 'Tag Element' tooltip is overlaid on the page, showing the namespace information for a specific element. The tooltip is titled 'Tag Element' and contains a tree view of the element's namespace hierarchy. The main content area shows the beginning of an article about cars, with the text 'transport people rather than goods.[2][3]'. Below this, there is a section titled 'Modern cars' with a small image of a car and a table with columns for 'Classification' and 'Industry'.





# Namespaces – sample with html

The image shows a screenshot of a Wikipedia page with a 'Tag Element' tool overlay. The tool displays the HTML structure of a selected element, which is a hatnote. The HTML code is as follows:

```
<div data-pdf-se-type="div" data-pdf-se-type-original="div" id="2181" class="hatnote navigation-not-searchable" role="note" >  
<a href="https://en.wikipedia.org/wiki/Motor_vehicle" data-pdf-se-type="a" data-pdf-se-type-original="a" id="2182" title="Motor vehicle" >  
<span data-pdf-se-type="span" data-pdf-se-type-original="span" id="2184" >  
Motor vehicle</span>  
</a>
```

The 'Tag Element' tool also shows the following structure:

```
/A <<Dictionary>>  
/O /HTML-5.00  
/class (hatnote navigation-not-searchable)  
/role (note)  
/K [Array]  
/NS <<Dictionary>>  
/NS (http://www.w3.org/1999/xhtml)  
/Type /Namespace  
/P <<Dictionary>>  
/S /div  
/Type /StructElement
```

The Wikipedia page in the background shows the 'WIKIPEDIA The Free Encyclopedia' logo, navigation links (Main page, Contents, etc.), and a list of links under 'Interaction' (Help, About Wikipedia, etc.). The main content area shows the beginning of an article about cars, mentioning 'transport people rather than horses' and 'Cars came into global use in the late 19th century, and development continued through the 20th century, and development continues to this day. The design of cars has evolved to meet the needs of the modern world, and they are now an essential part of our lives. The first motor vehicles were simple, open-top cars that were used for recreation and transport. They were often referred to as 'horseless carriages' or 'motor cars'. The first mass-produced motor car was the Ford Model T, which was introduced in 1908. It was a simple, affordable car that made it possible for many people to own a car. The Ford Model T was a revolution in transportation, and it paved the way for the modern car. Today, cars are used for a wide variety of purposes, from commuting to work to recreation. They are an essential part of our lives, and they continue to evolve and improve. The future of cars is bright, and we can expect to see many exciting developments in the years to come.



# Forms sample

The image shows a web form with a 'Tags' panel on the left and a form body on the right. The 'Tags' panel lists various HTML elements, with '<Form>' highlighted. The form body contains fields for personal information and health metrics.

**Tags Panel:**

- Tags
- <Document>
- > <H1>
- <Div>
- <Form>
- > <LbI>
- text\_firstname - OBJR
- <Div>
- > <Form>
- <Div>
- > <Form>
- > <Form>
- <Div>
- > <Form>
- > <Form>
- <Div>
- > <Form>

**Form Body:**

First name:	<input type="text"/>
Last name:	<input type="text"/>
Height:	<input type="text"/> <input type="text" value="cm"/>
Weight:	<input type="text"/> <input type="text" value="kg"/>
Gender:	<input checked="" type="radio"/> male <input type="radio"/> female
Ethnic group:	<input type="text" value="Asians"/> <input type="text" value="Europeans"/> <input type="text" value="Indians"/>
Age:	<input type="text"/> years
Adult:	<input type="checkbox"/>
Hip measure:	<input type="text"/> <input type="text" value="cm"/>

BMI:

BAI:

[Click here to check results](#)



# Forms sample

Tags

- <Document>
  - <H1>
  - <Div>
    - <Form>
    - <Lbl>
      - text\_firstname - OBJR
  - <Div>
    - <Form>
  - <Div>
    - <Form>
  - <Div>
    - <Form>
  - <Div>
    - <Form>
  - <Div>
    - <Form>
  - <Div>
    - <Form>

First name:  
Last name:  
  
Height:  
Weight:  
Gender:  
Ethnic group:  
  
Age:  
Adult:  
Hip measure:  
  
BMI:  
BAI:  
  
Clear

### Calculator

First name:

Last name:

Height:  cm

Weight:  kg

Gender:  male  female

Ethnic group:  
Africans  
Asians  
Europeans  
Indians

Age:  years

Adult:

Hip measure:  cm

BMI:

BAI:

[Click here to check results](#)

```
script.js
35 this.getField("text_weight").value = weight*2.205;
36 }
37 }
38 }
39 all_fields.push({name:"group_gender", id:"28", annots:[]})
40 all_fields.push({name:"group_gender", id:"26", annots:[]})
41 all_fields.push({name:"listbox_ethnic", id:"27", annots:[]})
42 all_fields.push({name:"text_age", id:"40", annots:[]});
43 function K40() {
44 AFNumber_Keystroke(0, 0, 0, 0, "", true);
45 }
46 function F40() {
47 AFNumber_Format(0, 0, 0, 0, "", true);
48 }
49 function B140() {
50 }
51 if(this.getField("text_age").value >= 18){
52 this.getField("checkbox_adult").checkThisBox(0,true);
53 }
54 else{
55 this.getField("checkbox_adult").checkThisBox(0,false);
56 }
57 }
58 }
59 all_fields.push({name:"checkbox_adult", id:"29", annots:[]})
60 all_fields.push({name:"text_hip", id:"30", annots:[]});
61 function K30() {
62 AFNumber_Keystroke(1, 0, 0, 0, "", true);
63 }
64 function F30() {
65 AFNumber_Format(1, 0, 0, 0, "", true);
66 }
67 all_fields.push({name:"dropdown_hip", id:"31", annots:[]})
68 function V31() {
69 var hip = this.getField("text_hip").value;
70 if(hip > 0){
71 if(event.value=="cm")
72 this.getField("text_hip").value = hip*2.54;
73 else
74 this.getField("text_hip").value = hip/2.54;
75 }
76 }
77 }
78 all_fields.push({name:"text_bmi", id:"32", annots:[]});
79 function K32() {
80 AFNumber_Keystroke(2, 0, 0, 0, "", true);
81 }
82 function F32() {
83 AFNumber_Format(2, 0, 0, 0, "", true);
84 }
```



# 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/figcaption	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				
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	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		
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		
H	h1-h6		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	0..n	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			
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			
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		
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		
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		
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		
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		
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		
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	
RB	rb							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
RT	rt							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
RP	rp							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
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	
WT	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
WP	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
FENote	p					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	0..n									0..n	0..n															0..n	0..n	0..n	
LI	li							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
LBody	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	
Table	table		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																																					0..n	0..n	0..n
TD	td							0..n																																					0..n	0..n	0..n
THead	thead							0..n																																					0..1	0..1	0..1
TBody	tbody							0..n																																					0..n	0..n	0..n
TFoot	tfoot							0..n																																					0..n	0..n	0..n
Caption	caption/figcaption					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	
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



# Summary

- World of single representation is over. Today we need to read out loud our pdfs, but what's going to be tomorrow?
- It's time to stop producing bad pdfs
- It's time to give authors full control over pdf interpretation
- Author decides instead of tool decides
- Instead of saying there is only one way to interpret PDF we will be saying that there is **always deterministic** way of interpreting pdf



# Resources

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



# Resources

The screenshot shows the PDF Derivation tool interface. The browser address bar is highlighted with a red box and contains the URL: `file:///W:/PDF-Derivation/pdf/PAC Report style classmap.pdf`. The main content area displays a "PAC Test Report" for an invoice. The report includes a "Check" table, a "PDF Document" summary, and a "Result" section with a green checkmark indicating that the PDF/UA requirements are fulfilled. A detailed "Checkpoint" table is also visible at the bottom.

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

Title
Invoice for John Q. Doe - Kraxi Systems, Inc.

Filename
PDFUA-Reference-09_(English-invoice) - Copy.pdf

Language	Tags	Pages	Size
en	79	1	120 KB

**Result**

The PDF/UA requirements checked by PAC are fulfilled.

Checkpoint	Passed	Warned	Failed
<b>Basic Requirements</b>			
PDF Syntax	85	0	0
Fonts	8	0	0
Content	586	0	0
Embedded Files	0	0	0
Natural Language	226	0	0
<b>Logical Structure</b>			
Structure Elements	24	0	0
Structure Tree	158	0	0

Version: 1.0.13 (PDFix 5.0.22)

The screenshot shows the PDF Derivation tool interface. The browser address bar is highlighted with a red box and contains the URL: `file:///P:/PDF-Derivation/output/PAC Report style classmap/index.html`. The main content area displays a "PAC Test Report" for an invoice, which is a derived HTML version of the PDF report. It includes the same "Check" table, "PDF Document" summary, and "Result" section with a green checkmark. The "Checkpoint" table is partially visible at the bottom.

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

Title
Invoice for John Q. Doe - Kraxi Systems, Inc.

Filename
PDFUA-Reference-09_(English-invoice) - Copy.pdf

Language	Tags	Pages	Size
en	79	1	120 KB

**Result**

The PDF/UA requirements checked by PAC are fulfilled.

Checkpoint	Passed	Warned	Failed
<b>Basic Requirements</b>			
PDF Syntax	85	0	0
Fonts	8	0	0

Version: 1.0.13 (PDFix 5.0.22)





Thanks !

Roman Toda, Normex

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

