

# The Arlington PDF Model

Peter Wyatt (CTO, PDF Association)



# Why a model of PDF?

- Highly complex file format
- ISO 32000 is a long specification
  - 1,000 pages
  - English prose
  - Errors, ambiguities, unstated assumptions, ...
- Differing understandings
- Differing interpretations

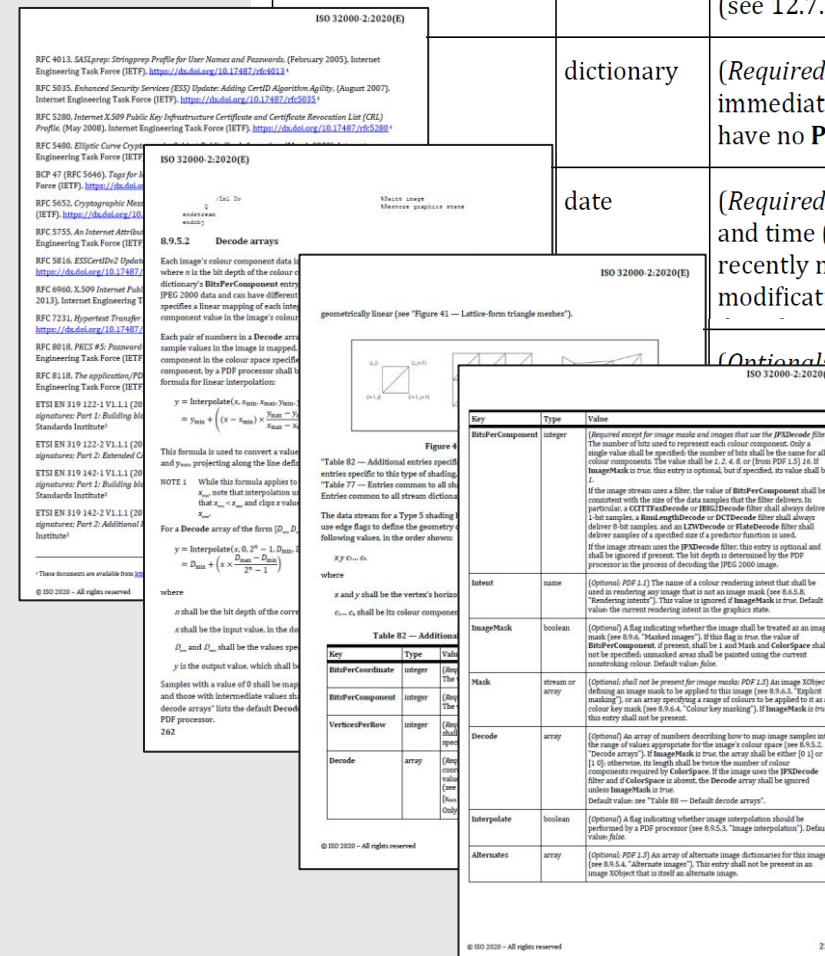
Table 31 — Entries in a page object

Key	Type	Value
Type	name	(Required) The type of PDF object that this dictionary describes; shall be <i>Page</i> for a page object or <i>Template</i> for an invisible Template page (see 12.7.7, "Named pages").
	dictionary	(Required; shall be an indirect reference) The page tree node that is the immediate parent of this page object. Objects of Type <i>Template</i> shall have no <b>Parent</b> key.
	date	(Required if <b>PieceInfo</b> is present; optional otherwise; PDF 1.3) The date and time (see 7.9.4, "Dates") when the page's contents were most recently modified. If a page-piece dictionary ( <b>PieceInfo</b> ) is present, the modification date shall be used to ascertain which of the application

(Optional; PDF 2.0) An array of one or more file specification (7.11.3, "File specification dictionaries") which denote the files for this page. See 14.13, "Associated files" and 14.13.8, "Files linked to DParts" for more details.

(Optional; PDF 2.0) An array of output intent dictionaries that shall describe the colour characteristics of output devices on which this page is rendered (see 14.11.5, "Output intents").

(Optional; PDF 2.0) An array of alternate image dictionaries for this page. If this page is within the range of a DPart, not permitted otherwise. (Optional; PDF 2.0) An indirect reference to the DPart dictionary whose key includes this page object (see 14.12.3, "Connecting the



Copyright © 2021, PDF Association

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0079.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Approved for public release.



# Applications for a model of PDF

- PDF reader (parser) generators
- PDF writer generators
- Checking the specification itself
- Documentation of PDF technology
- Test case generation
- Code coverage
- Code review
- OSS/TPS due diligence
- Extant data compliance (file validation)
- Digital forensics
  - File examination



# Extracting the data

Table 31 — Entries in a page object

Key	Type	Value
Type	name	<b>Required</b> The type of PDF object that this dictionary describes; shall be <b>Page</b> for a page object or <b>Template</b> for an invisible Template page (see 12.7.7, "Named pages").
Parent	dictionary	<b>Required</b> ; <b>shall be an indirect reference</b> The page tree node that is the immediate parent of this page object. <b>Objects of Type Template shall have no Parent key.</b>
LastModified	date	<b>(Required if PieceInfo is present; optional otherwise PDF 1.3)</b> The date and time (see 7.9.4, "Dates") when the page's contents were most recently modified. If a page-piece dictionary ( <b>PieceInfo</b> ) is present, the modification date shall be used to ascertain which of the application
AF	array of dictionaries	<b>(Optional PDF 2.0)</b> An array of one or more file specification dictionaries (7.11.3, "File specification dictionaries") which denote the associated files for this page. See 14.13, "Associated files" and 14.13.8, "Associated files linked to DParts" for more details.
OutputIntents	array	<b>(Optional PDF 2.0)</b> An array of output intent dictionaries that shall specify the colour characteristics of output devices on which this page might be rendered (see 14.11.5, "Output intents").
DPart	dictionary	<b>Required</b> if this page is within the range of a DPart, not permitted otherwise <b>PDF 2.0</b> An indirect reference to the DPart dictionary whose range of pages includes this page object (see 14.12.3, "Connecting the

Every key in every dictionary.  
For arrays, every array element

All basic PDF COS types, plus a few others for convenience

Version introduced (*unstated* = PDF 1.0).  
Version deprecated (*where appropriate*)

Indirect & Direct requirements

Required-ness / Optional, with conditions

Linkage to precise type(s) for each key value when dictionary or array



# The Arlington PDF data model

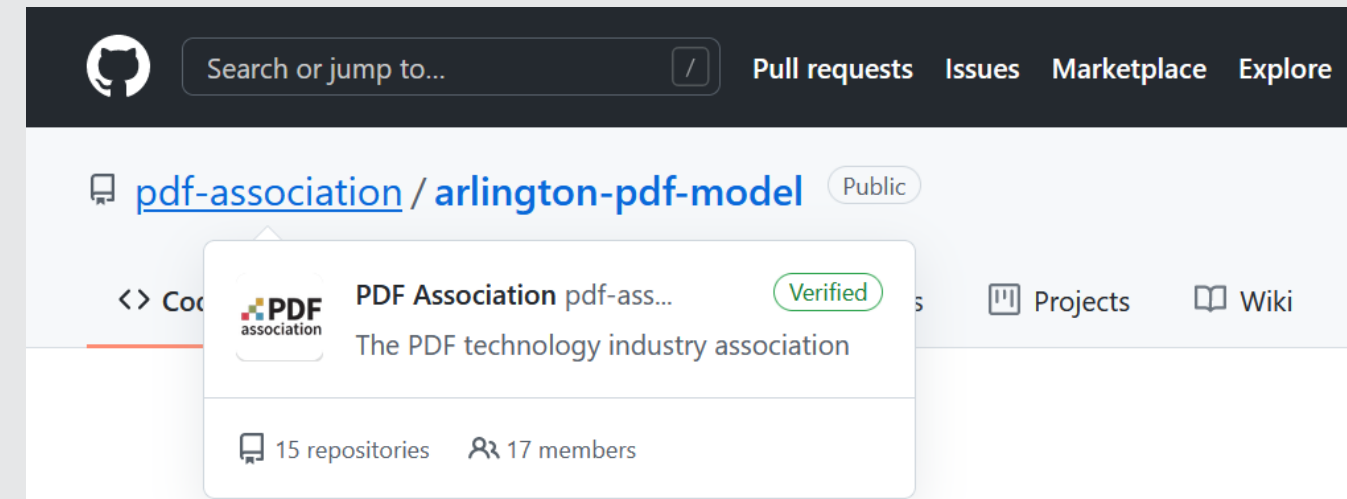
- Every data object is 12 fields:
  - Always identical structure
    - Dictionaries, arrays, streams, maps
  - Entire definition for a single PDF object
- Can include custom predicates:
  - `fn:Eval((@ca>=0.0) && (@ca<=1.0))`
  - `fn:IsMeaningful(@Subtype==Polygon)`
  - `fn:SinceVersion(1.6,Solidities)`

#	Column Name
1	Key name / array index
2	Type
3	Since Version
4	Deprecated In
5	Required?
6	Indirect Reference?
7	Inheritable?
8	Default Value
9	Possible Values
10	Special Case
11	Link
12	Notes ( <i>freeform text</i> )



# What is the Arlington PDF Model?

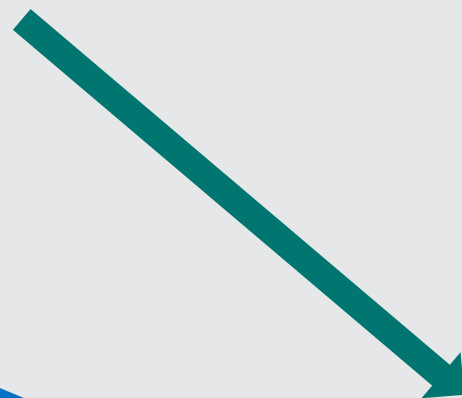
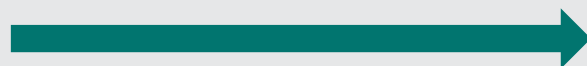
- Open source
- Extensible
- Specification derived
- Structured data
- Machine and human readable
- Vendor and implementation neutral
- Comprehensive definition of every PDF object
  - Every key in every dictionary and stream
  - Every array element in every array



## Arlington PDF Model



# Why TSV?



33 lines (33 sloc) 3.32 KB

Edit on HackMD Raw Blame

Search this file...

1	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleValues
2	Type	name	1.0		TRUE	FALSE	FALSE		[Catalog]
3	Version	name	1.4		FALSE	FALSE	FALSE		[1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,2.0]
4	Extensions	dictionary	1.7		FALSE	FALSE	FALSE		
5	Pages	dictionary	1.0		TRUE	TRUE	FALSE		
6	PageLabels	number-tree	1.3		FALSE	FALSE	FALSE		
7	Names	dictionary	1.2		FALSE	FALSE	FALSE		
8	Dests	dictionary	1.1		FALSE	TRUE	FALSE		
9	ViewerPreferences	dictionary	1.2		FALSE	FALSE	FALSE		
10	PageLayout	name	1.0		FALSE	FALSE	FALSE	SinglePage	[SinglePage, ...]

<xml/>  
{JSON}



Xtext



> diff/patch





# So what's possible?

## ■ No-code

- Spreadsheets
- Linux command line

## ■ Low-code

- Python
- Perl, ...

## ■ Code

- Java
- C++

## ■ Interactive

- Linux commands
- “Big data” TSV utilities
- Jupyter Notebook
- 3D/VR visualizations

## ■ Query-based

- JSON = JQ
- XML = XSD + XPath
- Proof of concept Java application

## ■ Non-interactive (batch)

- Proof of concept C++ application





```
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls
3dvisualize CHANGELOG.md CODE_OF_CONDUCT.md CONTRIBUTORS.txt INTERNAL_GRAMMAR.md LICENSE NOTICE.txt README.md TestGrammar gexml resources scripts tsv xml
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 2.0 latest
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.0
AnnotLink.tsv          ArrayOfNamesForProcSet.tsv      Dest0.tsv                FilterDCTDecode.tsv        FontTrueType.tsv          PageTreeNodeRoot.tsv
AnnotText.tsv          ArrayOfNumbersGeneral.tsv      Dest1.tsv                FilterLZWDecode.tsv       FontType1.tsv             Resource.tsv
ArrayOfAnnots.tsv      ArrayOfPageTreeNodeKids.tsv    Dest4.tsv                FontDescriptorTrueType.tsv FontType3.tsv             Stream.tsv
ArrayOfCompressionFilterNames.tsv ArrayOfStreamsGeneral.tsv      DestXYZ.tsv              FontDescriptorType1.tsv   IndexedColorSpace.tsv    Thumbnail.tsv
ArrayOfDecodeParams.tsv ArrayOf_4AnnotBorderCharacteristics.tsv DocInfo.tsv              FontDescriptorType3.tsv   Outline.tsv               XObjectFormType1.tsv
ArrayOfDifferences.tsv Catalog.tsv                    Encoding.tsv              FontFileType1.tsv         OutlineItem.tsv           XObjectImage.tsv
ArrayOfFilterNames.tsv CharProcMap.tsv                FileTrailer.tsv          FontMap.tsv               PageObject.tsv            XObjectMap.tsv
ArrayOfIntegersGeneral.tsv ColorSpaceMap.tsv              FilterCCITTFaxDecode.tsv FontMultipleMaster.tsv   PageTreeNode.tsv
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls -1 tsv/1.0 | wc -l
47
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.1
ActionGoTo.tsv          ArrayOfStreamsGeneral.tsv      CharProcMap.tsv          FontDescriptorTrueType.tsv OutlineItem.tsv
ActionGoToR.tsv         ArrayOfThreads.tsv            ColorSpaceMap.tsv        FontDescriptorType1.tsv   PageObject.tsv
ActionLaunch.tsv        ArrayOf_2DashNumbers.tsv      Dest0.tsv                FontDescriptorType3.tsv   PageTreeNode.tsv
ActionThread.tsv        ArrayOf_2Integers.tsv         Dest1.tsv                FontFileType1.tsv         PageTreeNodeRoot.tsv
ActionURI.tsv           ArrayOf_2StringsByte.tsv      Dest4.tsv                FontMap.tsv               Resource.tsv
AnnotLink.tsv          ArrayOf_4AnnotBorderCharacteristics.tsv DestDict.tsv             FontMultipleMaster.tsv   Stream.tsv
AnnotText.tsv          ArrayOf_4Integers.tsv         DestXYZ.tsv              FontTrueType.tsv         Thread.tsv
ArrayOfAnnots.tsv       ArrayOf_4NumbersColorAnnotation.tsv DestsMap.tsv             FontType1.tsv            Thumbnail.tsv
ArrayOfBeads.tsv        ArrayOf_9Numbers.tsv          DocInfo.tsv              FontType3.tsv            Transition.tsv
ArrayOfCompressionFilterNames.tsv Bead.tsv                      Encoding.tsv              Gamma.tsv                 URI.tsv
ArrayOfDecodeParams.tsv BeadFirst.tsv                 EncryptionPublicKey.tsv   IndexedColorSpace.tsv    Whitepoint.tsv
ArrayOfDifferences.tsv Blackpoint.tsv                EncryptionStandard.tsv   LabColorSpace.tsv        XObjectFormPS.tsv
ArrayOfFilterNames.tsv CalGrayColorSpace.tsv        FileSpecification.tsv    LabDict.tsv              XObjectFormPSpassthrough.tsv
ArrayOfIntegersGeneral.tsv CalGrayDict.tsv               FileTrailer.tsv          LabRange.tsv             XObjectFormType1.tsv
ArrayOfNamesForProcSet.tsv CalRGBColorSpace.tsv         FilterCCITTFaxDecode.tsv LinearizationParameterDict.tsv XObjectImage.tsv
ArrayOfNumbersGeneral.tsv CalRGBDict.tsv                FilterDCTDecode.tsv      MicrosoftWindowsLaunchParam.tsv XObjectMap.tsv
ArrayOfPageTreeNodeKids.tsv Catalog.tsv                   FilterLZWDecode.tsv      Outline.tsv
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls -1 tsv/1.1 | wc -l
84
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/latest/ | wc -l
515
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$
```



```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls
```

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/
```

```
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 2.0 latest
```

AnnotLink.tsv	ArrayOfNamesForProcSet.tsv	Dest0.tsv	FilterDCTDecode.tsv	FontTrueType.tsv	PageTreeNodeRoot.tsv
AnnotText.tsv	ArrayOfNumbersGeneral.tsv	Dest1.tsv	FilterLZWDecode.tsv	FontType1.tsv	Resource.tsv
ArrayOfAnnots.tsv	ArrayOfPageTreeNodeKids.tsv	Dest4.tsv	FontDescriptorTrueType.tsv	FontType3.tsv	Stream.tsv
ArrayOfCompressionFilterNames.tsv	ArrayOfStreamsGeneral.tsv	DestXYZ.tsv	FontDescriptorType1.tsv	IndexedColorSpace.tsv	Thumbnail.tsv
ArrayOfDecodeParams.tsv	ArrayOf_4AnnotBorderCharacteristics.tsv	DocInfo.tsv	FontDescriptorType3.tsv	Outline.tsv	XObjectFormType1.tsv
ArrayOfDifferences.tsv	Catalog.tsv	Encoding.tsv	FontFileType1.tsv	OutlineItem.tsv	XObjectImage.tsv
ArrayOfFilterNames.tsv	CharProcMap.tsv	FileTrailer.tsv	FontMap.tsv	PageObject.tsv	XObjectMap.tsv
ArrayOfIntegersGeneral.tsv	ColorSpaceMap.tsv	FilterCCITTFaxDecode.tsv	FontMultipleMaster.tsv	PageTreeNode.tsv	

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.0 | wc -l
```

```
47
```

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.1
```

ActionGoTo.tsv	ArrayOfStreamsGeneral.tsv	CharProcMap.tsv	FontDescriptorTrueType.tsv	OutlineItem.tsv
ActionGoToR.tsv	ArrayOfThreads.tsv	ColorSpaceMap.tsv	FontDescriptorType1.tsv	PageObject.tsv
ActionLaunch.tsv	ArrayOf_2DashNumbers.tsv	Dest0.tsv	FontDescriptorType3.tsv	PageTreeNode.tsv
ActionThread.tsv	ArrayOf_2Integers.tsv	Dest1.tsv	FontFileType1.tsv	PageTreeNodeRoot.tsv
ActionURI.tsv	ArrayOf_2StringsByte.tsv	Dest4.tsv	FontMap.tsv	Resource.tsv
AnnotLink.tsv	ArrayOf_4AnnotBorderCharacteristics.tsv	DestDict.tsv	FontMultipleMaster.tsv	Stream.tsv
AnnotText.tsv	ArrayOf_4Integers.tsv	DestXYZ.tsv	FontTrueType.tsv	Thread.tsv
ArrayOfAnnots.tsv	ArrayOf_4NumbersColorAnnotation.tsv	DestsMap.tsv	FontType1.tsv	Thumbnail.tsv
ArrayOfBeads.tsv	ArrayOf_9Numbers.tsv	DocInfo.tsv	FontType3.tsv	Transition.tsv
ArrayOfCompressionFilterNames.tsv	Bead.tsv	Encoding.tsv	Gamma.tsv	URI.tsv
ArrayOfDecodeParams.tsv	BeadFirst.tsv	EncryptionPublicKey.tsv	IndexedColorSpace.tsv	Whitepoint.tsv
ArrayOfDifferences.tsv	Blackpoint.tsv	EncryptionStandard.tsv	LabColorSpace.tsv	XObjectFormPS.tsv
ArrayOfFilterNames.tsv	CalGrayColorSpace.tsv	FileSpecification.tsv	LabDict.tsv	XObjectFormPSpassthrough.tsv
ArrayOfIntegersGeneral.tsv	CalGrayDict.tsv	FileTrailer.tsv	LabRange.tsv	XObjectFormType1.tsv
ArrayOfNamesForProcSet.tsv	CalRGBColorSpace.tsv	FilterCCITTFaxDecode.tsv	LinearizationParameterDict.tsv	XObjectImage.tsv
ArrayOfNumbersGeneral.tsv	CalRGBDict.tsv	FilterDCTDecode.tsv	MicrosoftWindowsLaunchParam.tsv	XObjectMap.tsv
ArrayOfPageTreeNodeKids.tsv	Catalog.tsv	FilterLZWDecode.tsv	Outline.tsv	

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.1 | wc -l
```

```
84
```

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/latest/ | wc -l
```

```
515
```

```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$
```



```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls
3dvisualize CHANGELOG.md CODE_OF_CONDUCT.md CONTRIBUTORS.txt INTERNAL_GRAMMAR.md LICENSE NOTICE.txt README.md TestGrammar gexml resources scripts tsv xml
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.0
AnnotLink.tsv      ArrayOfNamesForProcSet.tsv      Dest0.tsv      FilterDCTDecode.tsv      FontTrueType.tsv      PageTreeNodeRoot.tsv
AnnotText.tsv      ArrayOfNumbersGeneral.tsv      Dest1.tsv      FilterLZWDecode.tsv      FontType1.tsv      Resource.tsv
ArrayOfAnnots.tsv  ArrayOfPageTreeNodeKids.tsv    Dest4.tsv      FontDescriptorTrueType.tsv FontType3.tsv      Stream.tsv
ArrayOfCompressionFilterNames.tsv ArrayOfStreamsGeneral.tsv      DestXYZ.tsv    FontDescriptorType1.tsv  IndexedColorSpace.tsv Thumbnail.tsv
ArrayOfDecodeParams.tsv ArrayOf_4AnnotBorderCharacteristics.tsv DocInfo.tsv    FontDescriptorType3.tsv  Outline.tsv      XObjectFormType1.tsv
ArrayOfDifferences.tsv Catalog.tsv                    Encoding.tsv    FontFileType1.tsv        OutlineItem.tsv    XObjectImage.tsv
ArrayOfFilterNames.tsv CharProcMap.tsv                FileTrailer.tsv FontMap.tsv             PageObject.tsv    XObjectMap.tsv
ArrayOfIntegersGeneral.tsv ColorSpaceMap.tsv              FilterCCITTFaxDecode.tsv FontMultipleMaster.tsv PageTreeNode.tsv
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.0 | wc -l
47
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.1
ActionGoTo.tsv      ArrayOfStreamsGeneral.tsv      CharProcMap.tsv      FontDescriptorTrueType.tsv      OutlineItem.tsv
ActionGoToR.tsv     ArrayOfThreads.tsv            ColorSpaceMap.tsv    FontDescriptorType1.tsv        PageObject.tsv
ActionLaunch.tsv    ArrayOf_2DashNumbers.tsv      Dest0.tsv            FontDescriptorType3.tsv        PageTreeNode.tsv
ActionThread.tsv    ArrayOf_2Integers.tsv         Dest1.tsv            FontFileType1.tsv             PageTreeNodeRoot.tsv
ActionURI.tsv       ArrayOf_2StringsByte.tsv      Dest4.tsv            FontMap.tsv                   Resource.tsv
AnnotLink.tsv       ArrayOf_4AnnotBorderCharacteristics.tsv DestDict.tsv         FontMultipleMaster.tsv        Stream.tsv
AnnotText.tsv       ArrayOf_4Integers.tsv         DestXYZ.tsv          FontTrueType.tsv              Thread.tsv
ArrayOfAnnots.tsv   ArrayOf_4NumbersColorAnnotation.tsv DestsMap.tsv        FontType1.tsv                 Thumbnail.tsv
ArrayOfBeads.tsv    ArrayOf_9Numbers.tsv          DocInfo.tsv          FontType3.tsv                 Transition.tsv
ArrayOfCompressionFilterNames.tsv Bead.tsv                    Encoding.tsv          Gamma.tsv                     URI.tsv
ArrayOfDecodeParams.tsv BeadFirst.tsv                EncryptionPublicKey.tsv IndexedColorSpace.tsv        Whitepoint.tsv
ArrayOfDifferences.tsv Blackpoint.tsv                EncryptionStandard.tsv LabColorSpace.tsv            XObjectFormPS.tsv
ArrayOfFilterNames.tsv CalGrayColorSpace.tsv        FileSpecification.tsv LabDict.tsv                  XObjectFormPSpassthrough.tsv
ArrayOfIntegersGeneral.tsv CalGrayDict.tsv              FileTrailer.tsv      LabRange.tsv                  XObjectFormType1.tsv
ArrayOfNamesForProcSet.tsv CalRGBColorSpace.tsv        FilterCCITTFaxDecode.tsv LinearizationParameterDict.tsv XObjectImage.tsv
ArrayOfNumbersGeneral.tsv CalRGBDict.tsv               FilterDCTDecode.tsv  MicrosoftWindowsLaunchParam.tsv XObjectMap.tsv
ArrayOfPageTreeNodeKids.tsv Catalog.tsv                  FilterLZWDecode.tsv  Outline.tsv
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.1 | wc -l
84
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/latest/ | wc -l
515
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$
```



```
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls
3dvisualize CHANGELOG.md CODE_OF_CONDUCT.md CONTRIBUTORS.txt INTERNAL_GRAMMAR.md LICENSE NOTICE.txt README.md TestGrammar gexml resources scripts tsv xml
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 2.0 latest
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.0
AnnotLink.tsv          ArrayOfNamesForProcSet.tsv      Dest0.tsv                FilterDCTDecode.tsv        FontTrueType.tsv          PageTreeNodeRoot.tsv
AnnotText.tsv          ArrayOfNumbersGeneral.tsv       Dest1.tsv                FilterLZWDecode.tsv        FontType1.tsv             Resource.tsv
ArrayOfAnnots.tsv      ArrayOfPageTreeNodeKids.tsv     Dest4.tsv                FontDescriptorTrueType.tsv FontType3.tsv             Stream.tsv
ArrayOfCompressionFilterNames.tsv ArrayOfStreamsGeneral.tsv      DestXYZ.tsv              FontDescriptorType1.tsv   IndexedColorSpace.tsv    Thumbnail.tsv
ArrayOfDecodeParams.tsv ArrayOf_4AnnotBorderCharacteristics.tsv DocInfo.tsv              FontDescriptorType3.tsv  Outline.tsv              XObjectFormType1.tsv
ArrayOfDifferences.tsv Catalog.tsv                     Encoding.tsv              FontFileType1.tsv         OutlineItem.tsv           XObjectImage.tsv
ArrayOfFilterNames.tsv CharProcMap.tsv                 FileTrailer.tsv          FontMap.tsv               PageObject.tsv            XObjectMap.tsv
ArrayOfIntegersGeneral.tsv ColorSpaceMap.tsv              FilterCCITTFaxDecode.tsv FontMultipleMaster.tsv   PageTreeNode.tsv
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.0 | wc -l
117
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.1
ActionGoTo.tsv          ArrayOfStreamsGeneral.tsv       CharProcMap.tsv          FontDescriptorTrueType.tsv  OutlineItem.tsv
ActionGoToR.tsv         ArrayOfThreads.tsv             ColorSpaceMap.tsv        FontDescriptorType1.tsv    PageObject.tsv
ActionLaunch.tsv        ArrayOf_2DashNumbers.tsv       Dest0.tsv                FontDescriptorType3.tsv    PageTreeNode.tsv
ActionThread.tsv        ArrayOf_2Integers.tsv          Dest1.tsv                FontFileType1.tsv          PageTreeNodeRoot.tsv
ActionURI.tsv           ArrayOf_2StringsByte.tsv       Dest4.tsv                FontMap.tsv                Resource.tsv
AnnotLink.tsv           ArrayOf_4AnnotBorderCharacteristics.tsv DestDict.tsv             FontMultipleMaster.tsv    Stream.tsv
AnnotText.tsv           ArrayOf_4Integers.tsv          DestXYZ.tsv              FontTrueType.tsv          Thread.tsv
ArrayOfAnnots.tsv       ArrayOf_4NumbersColorAnnotation.tsv DestsMap.tsv             FontType1.tsv             Thumbnail.tsv
ArrayOfBeads.tsv        ArrayOf_9Numbers.tsv           DocInfo.tsv              FontType3.tsv             Transition.tsv
ArrayOfCompressionFilterNames.tsv Bead.tsv                      Encoding.tsv              Gamma.tsv                 URI.tsv
ArrayOfDecodeParams.tsv BeadFirst.tsv                 EncryptionPublicKey.tsv   IndexedColorSpace.tsv     Whitepoint.tsv
ArrayOfDifferences.tsv Blackpoint.tsv                EncryptionStandard.tsv   LabColorSpace.tsv         XObjectFormPS.tsv
ArrayOfFilterNames.tsv CalGrayColorSpace.tsv         FileSpecification.tsv    LabDict.tsv               XObjectFormPSpassthrough.tsv
ArrayOfIntegersGeneral.tsv CalGrayDict.tsv               FileTrailer.tsv          LabRange.tsv              XObjectFormType1.tsv
ArrayOfNamesForProcSet.tsv CalRGBColorSpace.tsv          FilterCCITTFaxDecode.tsv LinearizationParameterDict.tsv XObjectImage.tsv
ArrayOfNumbersGeneral.tsv CalRGBDict.tsv                FilterDCTDecode.tsv      MicrosoftWindowsLaunchParam.tsv XObjectMap.tsv
ArrayOfPageTreeNodeKids.tsv Catalog.tsv                   FilterLZWDecode.tsv      Outline.tsv
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.1 | wc -l
84
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$ ls tsv/latest/ | wc -l
515
pwyatt@PETER-SURFACEBOOK:/mnt/c/Temp/share/arlington-pdf-model$
```





```
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls
3dvisualize CHANGELOG.md CODE_OF_CONDUCT.md CONTRIBUTORS.txt INTERNAL_GRAMMAR.md LICENSE NOTICE.txt README.md TestGrammar gexml resources scripts tsv xml
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 2.0 latest
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.0
AnnotLink.tsv          ArrayOfNamesForProcSet.tsv      Dest0.tsv                FilterDCTDecode.tsv        FontTrueType.tsv          PageTreeNodeRoot.tsv
AnnotText.tsv          ArrayOfNumbersGeneral.tsv       Dest1.tsv                FilterLZWDecode.tsv        FontType1.tsv             Resource.tsv
ArrayOfAnnots.tsv      ArrayOfPageTreeNodeKids.tsv     Dest4.tsv                FontDescriptorTrueType.tsv FontType3.tsv             Stream.tsv
ArrayOfCompressionFilterNames.tsv ArrayOfStreamsGeneral.tsv       DestXYZ.tsv              FontDescriptorType1.tsv   IndexedColorSpace.tsv    Thumbnail.tsv
ArrayOfDecodeParams.tsv ArrayOf_4AnnotBorderCharacteristics.tsv DocInfo.tsv              FontDescriptorType3.tsv   Outline.tsv               XObjectFormType1.tsv
ArrayOfDifferences.tsv Catalog.tsv                     Encoding.tsv              FontFileType1.tsv         OutlineItem.tsv           XObjectImage.tsv
ArrayOfFilterNames.tsv CharProcMap.tsv                 FileTrailer.tsv          FontMap.tsv               PageObject.tsv           XObjectMap.tsv
ArrayOfIntegersGeneral.tsv ColorSpaceMap.tsv              FilterCCITTFaxDecode.tsv FontMultipleMaster.tsv   PageTreeNode.tsv
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.0 | wc -l
47
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/1.1
ActionGoTo.tsv          ArrayOfStreamsGeneral.tsv       CharProcMap.tsv          FontDescriptorTrueType.tsv OutlineItem.tsv
ActionGoToR.tsv         ArrayOfThreads.tsv             ColorSpaceMap.tsv        FontDescriptorType1.tsv   PageObject.tsv
ActionLaunch.tsv        ArrayOf_2DashNumbers.tsv       Dest0.tsv                FontDescriptorType3.tsv   PageTreeNode.tsv
ActionThread.tsv        ArrayOf_2Integers.tsv          Dest1.tsv                FontFileType1.tsv         PageTreeNodeRoot.tsv
ActionURI.tsv           ArrayOf_2StringsByte.tsv       Dest4.tsv                FontMap.tsv               Resource.tsv
AnnotLink.tsv           ArrayOf_4AnnotBorderCharacteristics.tsv DestDict.tsv             FontMultipleMaster.tsv   Stream.tsv
AnnotText.tsv           ArrayOf_4Integers.tsv          DestXYZ.tsv              FontTrueType.tsv         Thread.tsv
ArrayOfAnnots.tsv       ArrayOf_4NumbersColorAnnotation.tsv DestsMap.tsv             FontType1.tsv            Thumbnail.tsv
ArrayOfBeads.tsv        ArrayOf_9Numbers.tsv           DocInfo.tsv              FontType3.tsv            Transition.tsv
ArrayOfCompressionFilterNames.tsv Bead.tsv                       Encoding.tsv              Gamma.tsv                 URI.tsv
ArrayOfDecodeParams.tsv BeadFirst.tsv                  EncryptionPublicKey.tsv   IndexedColorSpace.tsv    Whitepoint.tsv
ArrayOfDifferences.tsv Blackpoint.tsv                 EncryptionStandard.tsv   LabColorSpace.tsv        XObjectFormPS.tsv
ArrayOfFilterNames.tsv CalGrayColorSpace.tsv          FileSpecification.tsv    LabDict.tsv              XObjectFormPSpassthrough.tsv
ArrayOfIntegersGeneral.tsv CalGrayDict.tsv                FileTrailer.tsv          LabRange.tsv             XObjectFormType1.tsv
ArrayOfNamesForProcSet.tsv CalRGBColorSpace.tsv           FilterCCITTFaxDecode.tsv LinearizationParameterDict.tsv XObjectImage.tsv
ArrayOfNumbersGeneral.tsv CalRGBDict.tsv                 FilterDCTDecode.tsv      MicrosoftWindowsLaunchParam.tsv XObjectMap.tsv
ArrayOfPageTreeNodeKids.tsv Catalog.tsv                    FilterLZWDecode.tsv      Outline.tsv
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls -l tsv/1.1 | wc -l
84
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ ls tsv/latest/ | wc -l
515
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ cat tsv/latest/* | wc -l
4059
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model$ calc 4059 - 515
3544
```

```

AnnotPopUp.tsv
AnnotPrinterMark.tsv
AnnotProjection.tsv
AnnotRedact.tsv
AnnotRichMedia.tsv
AnnotScreen.tsv
AnnotSound.tsv
AnnotSquare.tsv
AnnotSquiggly.tsv
AnnotStamp.tsv
AnnotStrikeOut.tsv
AnnotText.tsv
AnnotTrapNetwork.tsv
AnnotUnderline.tsv
AnnotWatermark.tsv
AnnotWidget.tsv
Appearance.tsv
AppearanceCharacteristics.tsv
AppearancePrinterMark.tsv
ArrayOfOCConfig.tsv
ArrayOfOCG.tsv
ArrayOfOCGState.tsv
ArrayOfOCUsage.tsv
ArrayOfOPI13Color.tsv
ArrayOfOptContentGroups.tsv
ArrayOfOptContentOrders1.tsv
ArrayOfOptContentOrders2.tsv
ArrayOfOutputIntents.tsv
ArrayOfPageTreeNodeKids.tsv
ArrayOfPages.tsv
ArrayOfPaths.tsv
ArrayOfQuadPoints.tsv
ArrayOfRequirements.tsv
ArrayOfRequirementsHandlers.tsv
ArrayOfRichMediaConfiguration.tsv
ArrayOfSignatureReferences.tsv
ArrayOfSoftwareIdentifiers.tsv
ArrayOfSourceInformations.tsv
CalGrayColorSpace.tsv
CalGrayDict.tsv
CalRGBColorSpace.tsv
CalRGBDict.tsv
Catalog.tsv
CertSeedValue.tsv
CharProcMap.tsv
ClassMap.tsv
Collection.tsv
CollectionColors.tsv
CollectionField.tsv
CollectionItem.tsv
CollectionSchema.tsv
CollectionSort.tsv
CollectionSplit.tsv
CollectionSubitem.tsv
ColorSpaceMap.tsv
ColorantsDict.tsv
CryptFilter.tsv
FontCIDType0.tsv
FontCIDType2.tsv
FontDescriptorCIDType0.tsv
FontDescriptorCIDType2.tsv
FontDescriptorTrueType.tsv
FontDescriptorType1.tsv
FontDescriptorType3.tsv
FontFile.tsv
FontFile2CIDType2.tsv
FontFile2TrueType.tsv
FontFile3CIDType0.tsv
FontFile3Type1.tsv
FontFileType1.tsv
FontMap.tsv
FontMultipleMaster.tsv
FontTrueType.tsv
FontType0.tsv
FontType1.tsv
FontType3.tsv
Namespace.tsv
NavNode.tsv
Navigator.tsv
NumberFormat.tsv
OPIVersion13.tsv
OPIVersion13Dict.tsv
OPIVersion20.tsv
OPIVersion20Dict.tsv
ObjectReference.tsv
ObjectStream.tsv
OptContentConfig.tsv
OptContentCreatorInfo.tsv
OptContentExport.tsv
OptContentGroup.tsv
OptContentLanguage.tsv
OptContentMembership.tsv
OptContentPageElement.tsv
OptContentPrint.tsv
OptContentProperties.tsv

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ more CalRGBDict.tsv
Key      Type      SinceVersion  DeprecatedIn  Required  IndirectReference  Inheritable  DefaultValue  PossibleValues  SpecialCase  Link  Note
WhitePoint  array    1.1           TRUE         FALSE    FALSE             [fn:SinceVersion(1.1,Whitepoint)]
BlackPoint  array    1.1           FALSE        FALSE    FALSE             [fn:SinceVersion(1.1,Blackpoint)]
Gamma       array    1.1           FALSE        FALSE    FALSE             [fn:SinceVersion(1.1,Gamma)]
Matrix      array    1.1           FALSE        FALSE    FALSE             [ArrayOf_9Numbers]

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ tabs 1,20,37,50,64,73,91,103,123,148,175,190,210,230

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ more CalRGBDict.tsv
Key      Type      SinceVersion  DeprecatedIn  Required  IndirectReference  Inheritable  DefaultValue  PossibleValues  SpecialCase
WhitePoint  array    1.1           TRUE         FALSE    FALSE             FALSE
BlackPoint  array    1.1           FALSE        FALSE    FALSE             FALSE          [0.0 0.0 0.0]
Gamma       array    1.1           FALSE        FALSE    FALSE             FALSE          [1.0 1.0 1.0]
Matrix      array    1.1           FALSE        FALSE    FALSE             FALSE          [1 0 0 0 1 0 0 0 1]

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ tsv-pretty CalRGBDict.tsv
Key      Type      SinceVersion  DeprecatedIn  Required  IndirectReference  Inheritable  DefaultValue  PossibleValues  SpecialCase  Link
WhitePoint  array    1.1           TRUE         FALSE    FALSE             FALSE             [fn:SinceVersion(1.1,Whit
BlackPoint  array    1.1           FALSE        FALSE    FALSE             FALSE             [fn:SinceVersion(1.1,Blac
Gamma       array    1.1           FALSE        FALSE    FALSE             FALSE             [fn:SinceVersion(1.1,Gamm
Matrix      array    1.1           FALSE        FALSE    FALSE             FALSE             [ArrayOf_9Numbers]

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ tsv-pretty Whitepoint.tsv
Key  Type      SinceVersion  DeprecatedIn  Required  IndirectReference  Inheritable  DefaultValue  PossibleValues  SpecialCase  Link  Note
0    number    1.1           TRUE         FALSE    FALSE             FALSE             [fn:Eval(@0>0)]
1    number    1.1           TRUE         FALSE    FALSE             FALSE             [1]
2    number    1.1           TRUE         FALSE    FALSE             FALSE             [fn:Eval(@2>0)]

```

```

/mnt/c/Temp/share/arlington-pdf-model/tsv/latest$

```





pwyatt@PETER-SURFACEBOOK: X



Command Prompt



```
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ cut -f 2 * | sort | uniq
```

```
Type
array
array;boolean
array;boolean;dictionary;integer;name;number;stream;string
array;boolean;dictionary;integer;name;stream;string
array;boolean;dictionary;name;null;number;stream;string
array;boolean;integer;number;string-text
array;dictionary
array;dictionary;integer
array;dictionary;name;stream
array;dictionary;null
array;dictionary;stream
array;dictionary;string
array;dictionary;string-text
array;fn:SinceVersion(1.3,dictionary);name;stream
array;integer
array;integer;string-byte
array;name
array;name;string-byte
array;null
array;number
array;stream
array;string
array;string-byte
```





```
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model/tsv/latest$ \  
> cut -f 2 * | sed -e 's/;/\n/g' | sort | uniq  
Type  
array  
bitmask  
boolean  
date  
dictionary  
fn:Deprecated(2.0,array)  
fn:SinceVersion(1.1,array)  
fn:SinceVersion(1.1,dictionary)  
fn:SinceVersion(1.1,name)  
fn:SinceVersion(1.2,string-byte)  
fn:SinceVersion(1.3,dictionary)  
fn:SinceVersion(1.6,array)  
fn:SinceVersion(2.0,array)  
integer  
matrix  
name  
name-tree  
null  
number  
number-tree  
rectangle  
stream  
string  
string-ascii  
string-byte  
string-text  
pwyatt@PETER-SURFACEBOOK: /mnt/c/Temp/share/arlington-pdf-model/tsv/latest$
```



## Arlington PDF Model Jupyter Notebook

This notebook demonstrates how the Arlington PDF Model can be combined into a monolithic TSV file suitable for use by pandas in a Jupyter Notebook. The Python script `arlington-to-pandas.py` will convert an Arlington file set into a single TSV file by adding the object name as the first (left-most) field called "Object".

```
In [1]: 1 %matplotlib inline
        2 import pandas as pd
        3 import numpy as np
        4 import matplotlib.pyplot as plt
        5 import seaborn as sns
        6 sns.set(style="darkgrid")
```

```
In [2]: 1 df = pd.read_csv('pandas.tsv', delimiter='\t', na_filter=False,
        2                   dtype={'Object': 'string', 'Key': 'string', 'Type': 'string',
        3                             'SinceVersion': 'string', 'DeprecatedIn': 'string', 'Required': 'string',
        4                             'IndirectReference': 'string', 'Inheritable': 'string',
        5                             'DefaultValue': 'string', 'PossibleValues': 'string',
        6                             'SpecialCase': 'string', 'Link': 'string', 'Note': 'string'})
```

The above two cells will have loaded the monolithic TSV into pandas as a Dataframe object. All fields are treated as strings to avoid NaN and other issues. This also matches how other Python scripts and the C++ PoC TestGrammar operate. Here is what the data looks like:

The above two cells will have loaded the monolithic TSV into pandas as a Dataframe object. All fields are treated as strings to avoid NaN and other issues. This also matches how other Python scripts and the C++ PoC TestGrammar operate. Here is what the data looks like:

In [3]:

```
1 df.head()
```

Out[3]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleValue
0	3DActivation	A	name	1.6		FALSE	FALSE	FALSE	XA	[PO,PV,X
1	3DActivation	AIS	name	1.6		FALSE	FALSE	FALSE	L	[I
2	3DActivation	D	name	1.6		FALSE	FALSE	FALSE	PI	[PC,PI,X
3	3DActivation	DIS	name	1.6		FALSE	FALSE	FALSE	U	[U,I
4	3DActivation	TB	boolean	1.7		FALSE	FALSE	FALSE	true	

In [4]:

```
1 df.tail()
```

Out[4]:

	Object	Key	Type	SinceVersion	DeprecatedIn
3539	XRefStream	Info	dictionary	1.5	
3540	XRefStream	ID	array	1.5	fn:IsRequired(fn:Sin
3541	XRefStream	Encrypt	dictionary	1.5	
3542	_UniversalArray	*	array;boolean;dictionary;name>null;number;stre...	1.0	
3543	_UniversalDictionary	*	array;boolean;dictionary;name>null;number;stre...	1.0	

The above two cells will have loaded the monolithic TSV into pandas as a Dataframe object. All fields are treated as strings to avoid NaN and other issues. This also matches how other Python scripts and the C++ PoC TestGrammar operate. Here is what the data looks like:

In [3]:

```
1 df.head()
```

Out [3]:

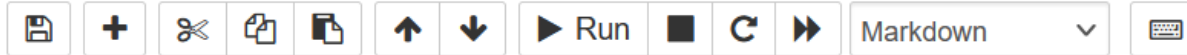
	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleValue
0	3DActivation	A	name	1.6		FALSE	FALSE	FALSE	XA	[PO,PV,X
1	3DActivation	AIS	name	1.6		FALSE	FALSE	FALSE	L	[I
2	3DActivation	D	name	1.6		FALSE	FALSE	FALSE	PI	[PC,PI,X
3	3DActivation	DIS	name	1.6		FALSE	FALSE	FALSE	U	[U,I
4	3DActivation	TB	boolean	1.7		FALSE	FALSE	FALSE	true	

In [4]:

```
1 df.tail()
```

Out [4]:

	Object	Key	Type	SinceVersion	DeprecatedIn
3539	XRefStream	Info	dictionary	1.5	
3540	XRefStream	ID	array	1.5	fn:IsRequired(fn:Sin
3541	XRefStream	Encrypt	dictionary	1.5	
3542	_UniversalArray	*	array;boolean;dictionary;name>null;number;stre...	1.0	
3543	_UniversalDictionary	*	array;boolean;dictionary;name>null;number;stre...	1.0	



The above two cells will have loaded the monolithic TSV into pandas as a Dataframe object. All fields are treated as strings to avoid NaN and other issues. This also matches how other Python scripts and the C++ PoC TestGrammar operate. Here is what the data looks like:

```
In [3]: 1 df.head()
```

Out[3]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleValue
0	3DActivation	A	name	1.6		FALSE	FALSE	FALSE	XA	[PO,PV,XA]
1	3DActivation	AIS	name	1.6		FALSE	FALSE	FALSE	L	[L]
2	3DActivation	D	name	1.6		FALSE	FALSE	FALSE	PI	[PC,PI,XA]
3	3DActivation	DIS	name	1.6		FALSE	FALSE	FALSE	U	[U,I]
4	3DActivation	TB	boolean	1.7		FALSE	FALSE	FALSE	true	
<div>&lt; &gt;</div>										

```
In [4]: 1 df.tail()
```

Out[4]:

	Object	Key	Type	SinceVersion	DeprecatedIn
3539	XRefStream	Info	dictionary	1.5	
3540	XRefStream	ID	array	1.5	fn:IsRequired(fn:SinceVersion)
3541	XRefStream	Encrypt	dictionary	1.5	
3542	_UniversalArray	*	array;boolean;dictionary;name>null;number;string	1.0	
3543	_UniversalDictionary	*	array;boolean;dictionary;name>null;number;string	1.0	

The above two cells will have loaded the monolithic TSV into pandas as a Dataframe object. All fields are treated as strings to avoid NaN and other issues. This also matches how other Python scripts and the C++ PoC TestGrammar operate. Here is what the data looks like:

In [3]:

```
1 df.head()
```

Out [3]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleValue
0	3DActivation	A	name	1.6		FALSE	FALSE	FALSE	XA	[PO,PV,X
1	3DActivation	AIS	name	1.6		FALSE	FALSE	FALSE	L	[I
2	3DActivation	D	name	1.6		FALSE	FALSE	FALSE	PI	[PC,PI,X
3	3DActivation	DIS	name	1.6		FALSE	FALSE	FALSE	U	[U,I
4	3DActivation	TB	boolean	1.7		FALSE	FALSE	FALSE	true	

In [4]:

```
1 df.tail()
```

Out [4]:

	Object	Key	Type	SinceVersion	DeprecatedIn
3539	XRefStream	Info	dictionary	1.5	
3540	XRefStream	ID	array	1.5	fn:IsRequired(fn:Sin
3541	XRefStream	Encrypt	dictionary	1.5	
3542	_UniversalArray		array;boolean;dictionary;name>null;number;stre...	1.0	
3543	_UniversalDictionary		array;boolean;dictionary;name>null;number;stre...	1.0	





So how many keys or array elements does Arlington define? This will be the number of rows in the Dataframe!

```
In [ ]: 1 len(df)
```

```
In [ ]: 1 df.dtypes
```

```
In [ ]: 1 df.columns
```

The fields (columns) are easy to access:

```
In [ ]: 1 df['Object']
```

```
In [ ]: 1 len(pd.unique(df['Object']))
```

```
In [ ]: 1 df['Type'].str.split(';')
```

Using Datafield query method we can create some simple filters over the data:

```
In [ ]: 1 df.query('Key == "Subtype"')
```

```
In [ ]: 1 df.query('("string-ascii" in Type) and (Required == "TRUE")')
```

Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

```
In [ ]: 1 bin sizes, , = plt.hist(df['SinceVersion'].sort values(ascending=True))
```

Let's try and identify this mysterious PDF object:





So how many keys or array elements does Arlington define? This will be the number of rows in the Dataframe!

In [5]: 1 len(df)

Out[5]: 3544

In [6]: 1 df.dtypes

Out[6]:

Object	string
Key	string
Type	string
SinceVersion	string
DeprecatedIn	string
Required	string
IndirectReference	string
Inheritable	string
DefaultValue	string
PossibleValues	string
SpecialCase	string
Link	string
Note	string
dtype:	object

In [7]: 1 df.columns

Out[7]: Index(['Object', 'Key', 'Type', 'SinceVersion', 'DeprecatedIn', 'Required',  
'IndirectReference', 'Inheritable', 'DefaultValue', 'PossibleValues',  
'SpecialCase', 'Link', 'Note'],  
dtype='object')

The fields (columns) are easy to access:

In [ ]: 1 df['Object']



The fields (columns) are easy to access:

```
In [ ]: 1 df['Object']
```

```
In [ ]: 1 len(pd.unique(df['Object']))
```

```
In [ ]: 1 df['Type'].str.split(';')
```

Using Datafield query method we can create some simple filters over the data:

```
In [ ]: 1 df.query('Key == "Subtype"')
```

```
In [ ]: 1 df.query('("string-ascii" in Type) and (Required == "TRUE")')
```

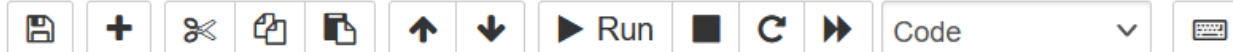
Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

```
In [ ]: 1 bin_sizes, _, _ = plt.hist(df['SinceVersion'].sort_values(ascending=True))
```

Let's try and identify this mysterious PDF object:

```
52 0 obj
<</R[146 522 153 542]/N 53 0 R/P 51 0 R/T 570 0
R/V 53 0 R>>
endobj
53 0 obj
<</R[352 570 359 590]/N 52 0 R/P 51 0 R/T 570 0
R/V 52 0 R>>
endobj
```





The fields (columns) are easy to access:

```
In [8]: 1 df['Object']
```

```
Out[8]: 0      3DActivation
        1      3DActivation
        2      3DActivation
        3      3DActivation
        4      3DActivation
        ...
        3539      XRefStream
        3540      XRefStream
        3541      XRefStream
        3542      _UniversalArray
        3543      _UniversalDictionary
        Name: Object, Length: 3544, dtype: string
```

```
In [9]: 1 len(pd.unique(df['Object']))
```

```
Out[9]: 515
```

```
In [ ]: 1 df['Type'].str.split(';')
```

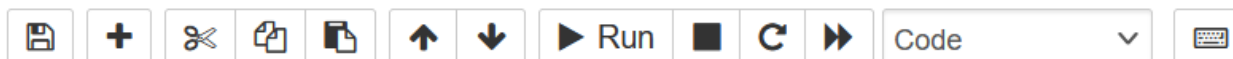
Using Datafield query method we can create some simple filters over the data:

```
In [ ]: 1 df.query('Key == "Subtype"')
```

```
In [ ]: 1 df.query('("string-ascii" in Type) and (Required == "TRUE")')
```

Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:





```
In [9]: 1 len(pd.unique(df['Object']))
```

```
Out[9]: 515
```

```
In [10]: 1 df['Type'].str.split(';')
```

```
Out[10]: 0 [name]
1 [name]
2 [name]
3 [name]
4 [boolean]
...
3539 [dictionary]
3540 [array]
3541 [dictionary]
3542 [array, boolean, dictionary, name, null, numbe...]
3543 [array, boolean, dictionary, name, null, numbe...]
Name: Type, Length: 3544, dtype: object
```

Using Datafield query method we can create some simple filters over the data:

```
In [ ]: 1 df.query('Key == "Subtype"')
```

```
In [ ]: 1 df.query('("string-ascii" in Type) and (Required == "TRUE")')
```

Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

```
In [ ]: 1 bin_sizes, _, _ = plt.hist(df['SinceVersion'].sort_values(ascending=True))
```

Let's try and identify this mysterious PDF object:





```
Name: Type, Length: 3544, dtype: object
```

Using Datafield query method we can create some simple filters over the data:

```
In [12]: 1 df.query('Key == "Subtype"')
```


Out[12]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue
10	3DAnimationStyle	Subtype	name	1.7		FALSE	FALSE	FALSE	None
14	3DBackground	Subtype	name	1.6		FALSE	FALSE	FALSE	SC
28	3DLightingScheme	Subtype	name	1.7		TRUE	FALSE	FALSE	
30	3DMeasure3DC	Subtype	name	2.0		TRUE	FALSE	FALSE	
41	3DMeasureAD3	Subtype	name	2.0		TRUE	FALSE	FALSE	
...	...	...	...	...	...	...	...	...	...
3391	XObjectFormTrapNet	Subtype	name	1.3	2.0	TRUE	FALSE	FALSE	
3421	XObjectFormType1	Subtype	name	1.0		TRUE	FALSE	FALSE	
3448	XObjectImage	Subtype	name	1.0		TRUE	FALSE	FALSE	
3478	XObjectImageMask	Subtype	name	1.3		TRUE	FALSE	FALSE	
3506	XObjectImageSoftMask	Subtype	name	1.4		TRUE	FALSE	FALSE	

80 rows × 13 columns

```
In [ ]: 1 df.query('("string-ascii" in Type) and (Required == "TRUE")')
```


Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

In [13]:  1 df.query('("string-ascii" in Type) and (Required == "TRUE")')

Out[13]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue
299	ActionURI	URI	string-ascii	1.1		TRUE	FALSE	FALSE	
1493	CIDSystemInfo	Registry	string-ascii	1.2		TRUE	FALSE	FALSE	
1494	CIDSystemInfo	Ordering	string-ascii	1.2		TRUE	FALSE	FALSE	
3078	SoftwareIdentifier	U	string-ascii	1.5		TRUE	FALSE	FALSE	
3241	TimeStampDict	URL	string-ascii	1.6		TRUE	FALSE	FALSE	
3252	URLAlias	U	string-ascii	1.3		TRUE	FALSE	FALSE	
3308	WebCaptureCommand	URL	string-ascii	1.3		TRUE	FALSE	FALSE	

Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

In [ ]:  1 bin\_sizes, \_, \_ = plt.hist(df['SinceVersion'].sort\_values(ascending=True))

Let's try and identify this mysterious PDF object:

```
52 0 obj
<</R[146 522 153 542]/N 53 0 R/P 51 0 R/T 570 0
```

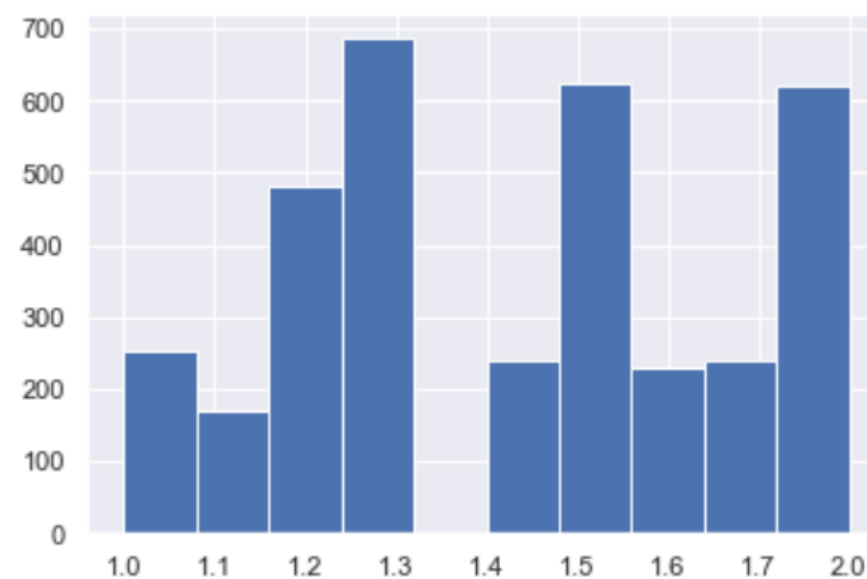




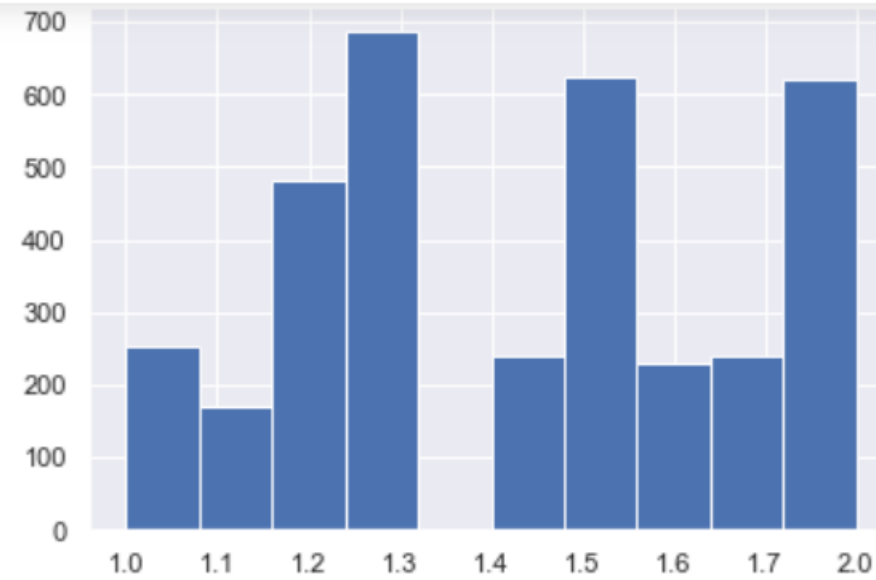
1494	CIDSystemInfo	Ordering	string-ascii	1.2	TRUE	FALSE	FALSE
3078	SoftwareIdentifier	U	string-ascii	1.5	TRUE	FALSE	FALSE
3241	TimeStampDict	URL	string-ascii	1.6	TRUE	FALSE	FALSE
3252	URLAlias	U	string-ascii	1.3	TRUE	FALSE	FALSE
3308	WebCaptureCommand	URL	string-ascii	1.3	TRUE	FALSE	FALSE

Not many fields are numeric, but we can do a simple histogram to get a feel of how many new keys were introduced in each PDF version:

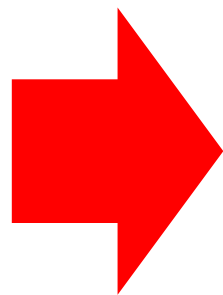
```
In [14]: 1 bin_sizes, _, _ = plt.hist(df['SinceVersion'].sort_values(ascending=True))
```







Let's try and identify this mysterious PDF object:



```
52 0 obj
<</R[146 522 153 542]/N 53 0 R/P 51 0 R/T 570 0
R/V 53 0 R>>
endobj
53 0 obj
<</R[352 570 359 590]/N 52 0 R/P 51 0 R/T 570 0
R/V 52 0 R>>
endobj
```

Lets make small Dataframes for each key we are interested in and limit the output to just the Object, Key and Type:

```
In [ ]: 1 n_key = df.loc[df['Key']=="N", ['Object', 'Key', 'Type']]
        2 p_key = df.loc[df['Key']=="P", ['Object', 'Key', 'Type']]
        3 t_key = df.loc[df['Key']=="T", ['Object', 'Key', 'Type']]
        4 r_key = df.loc[df['Key']=="R", ['Object', 'Key', 'Type']]
```





```
endobj
53 0 obj
<</R[352 570 359 590]/N 52 0 R/P 51 0 R/T 570 0
R/V 52 0 R>>
endobj
```

Lets make small Dataframes for each key we are interested in and limit the output to just the Object, Key and Type:

```
In [15]: 1 n_key = df.loc[df['Key']=="N", ['Object', 'Key', 'Type']]
2 p_key = df.loc[df['Key']=="P", ['Object', 'Key', 'Type']]
3 t_key = df.loc[df['Key']=="T", ['Object', 'Key', 'Type']]
4 r_key = df.loc[df['Key']=="R", ['Object', 'Key', 'Type']]
```

Now we can do inner-merge to intersect the sets with each other

```
In [17]: 1 df1 = pd.merge(n_key, p_key, how='inner', on=['Object'])
2 df2 = pd.merge(t_key, r_key, how='inner', on=['Object'])
3 pd.merge(df1, df2, how='inner', on=['Object'])
```

Out[17]:

	Object	Key_x_x	Type_x_x	Key_y_x	Type_y_x	Key_x_y	Type_x_y	Key_y_y	Type_y_y
0	Bead	N	dictionary	P	dictionary	T	dictionary	R	rectangle
1	BeadFirst	N	dictionary	P	dictionary	T	dictionary	R	rectangle
2	Target	N	string-byte	P	integer;string-byte	T	dictionary	R	name

From this result we can see that only Bead and BeadFirst objects have an R key which is an array (specifically, a rectangle in the Arlington predefined types). The Target R key needs to be a PDF name which the fragment is clearly not. So the answer is that it is a Bead/BeadFirst dictionary!

Let's now investigate inheritable-ness...



Let's now investigate inheritable-ness...

```
In [18]: 1 inheritable_keys = df.loc[df['Inheritable']=="TRUE", ['Object', 'Key', 'Type', 'Required']]
2 len(inheritable_keys)
```

Out[18]: 68

```
In [19]: 1 inheritable_keys
```

Out[19]:

	Object	Key	Type	Required	Inheritable	DefaultValue
284	ActionSubmitForm	Flags	bitmask	FALSE	TRUE	0
285	ActionSubmitForm	CharSet	string	FALSE	TRUE	
1632	DocTimeStamp	Filter	name	TRUE	TRUE	
1733	FieldBtn	FT	name	TRUE	TRUE	
1739	FieldBtn	Ff	bitmask	FALSE	TRUE	
...	...	...	...	...	...	...
3158	StandardLayoutAttributesILSE	TextDecorationType	name	FALSE	TRUE	None
3159	StandardLayoutAttributesILSE	RubyAlign	name	FALSE	TRUE	Distribute
3160	StandardLayoutAttributesILSE	RubyPosition	name	FALSE	TRUE	Before
3161	StandardLayoutAttributesILSE	GlyphOrientationVertical	name;number	FALSE	TRUE	[Auto];[]
3163	StandardListAttributes	ListNumbering	name	FALSE	TRUE	None

68 rows × 6 columns

So there are a total of 68 keys defined to have inheritance (of some kind) in PDF 2.0 (as defined by ISO 32000-2:2020).

68 rows × 6 columns

So there are a total of 68 keys defined to have inheritance (of some kind) in PDF 2.0 (as defined by ISO 32000-2:2020).

Let's narrow to required keys that also inheritable:

```
In [20]: 1 inheritable_and_required_keys = inheritable_keys.loc[inheritable_keys['Required']=="TRUE"]
        2 inheritable_and_required_keys
```

Out[20]:

	Object	Key	Type	Required	Inheritable	DefaultValue
1632	DocTimeStamp	Filter	name	TRUE	TRUE	
1733	FieldBtn	FT	name	TRUE	TRUE	
1741	FieldBtn	DA	string	TRUE	TRUE	
1748	FieldCh	FT	name	TRUE	TRUE	
1758	FieldCh	DA	string	TRUE	TRUE	
1769	FieldSig	FT	name	TRUE	TRUE	
1779	FieldSig	DA	string	TRUE	TRUE	
1785	FieldTx	FT	name	TRUE	TRUE	
1795	FieldTx	DA	string	TRUE	TRUE	
2608	PageObject	Resources	dictionary	TRUE	TRUE	
2609	PageObject	MediaBox	rectangle	TRUE	TRUE	
3028	Signature	Filter	name	TRUE	TRUE	

Everyone thinks that there are more required and inheritable keys in the PageObject...



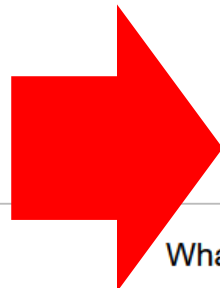
1748	FieldCh	FT	name	TRUE	TRUE
1758	FieldCh	DA	string	TRUE	TRUE
1769	FieldSig	FT	name	TRUE	TRUE
1779	FieldSig	DA	string	TRUE	TRUE
1785	FieldTx	FT	name	TRUE	TRUE
1795	FieldTx	DA	string	TRUE	TRUE
2608	PageObject	Resources	dictionary	TRUE	TRUE
2609	PageObject	MediaBox	rectangle	TRUE	TRUE
3028	Signature	Filter	name	TRUE	TRUE

Everyone thinks that there are more required and inheritable keys in the PageObject...

In [21]: 1 inheritable\_keys.query('Object=="PageObject"')

Out[21]:

	Object	Key	Type	Required	Inheritable	DefaultValue
2608	PageObject	Resources	dictionary	TRUE	TRUE	
2609	PageObject	MediaBox	rectangle	TRUE	TRUE	
2610	PageObject	CropBox	rectangle	FALSE	TRUE	@MediaBox
2616	PageObject	Rotate	integer	FALSE	TRUE	0
2638	PageObject	Hid	boolean	FALSE	TRUE	false




What is this "Hid" key?? That's not in ISO 32000-1 or ISO 32000-2! Let's get the full data for that key:

In [ ]: 1 df.query('Object=="PageObject" and Key=="Hid"')

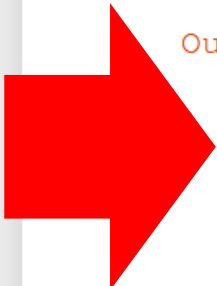




What is this "Hid" key?? That's not in ISO 32000-1 or ISO 32000-2! Let's get the full data for that key:


In [22]:  1 df.query('Object=="PageObject" and Key=="Hid"')

Out[22]:



	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleV
2638	PageObject	Hid	boolean	1.1	1.3	FALSE	FALSE	TRUE	false	

And what about all the other ...Box keys?

In [23]:  1 page\_obj = df.query('Object=="PageObject"')  
2 page\_obj[page\_obj.Key.str.contains("Box")]


Out[23]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	
2609	PageObject	MediaBox	rectangle	1.0		TRUE	FALSE	TRUE		
2610	PageObject	CropBox	rectangle	1.0		FALSE	FALSE	TRUE	@MediaBox	
2611	PageObject	BleedBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2612	PageObject	TrimBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2613	PageObject	ArtBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2614	PageObject	BoxColorInfo	dictionary	1.4		FALSE	FALSE	FALSE		






What is this "Hid" key?? That's not in ISO 32000-1 or ISO 32000-2! Let's get the full data for that key:

In [22]:  1 df.query('Object=="PageObject" and Key=="Hid"')

Out[22]:

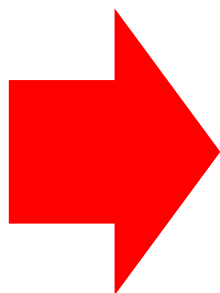
	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	PossibleV
2638	PageObject	Hid	boolean	1.1	1.3	FALSE	FALSE	TRUE	false	

And what about all the other ...Box keys?

In [23]:  1 page\_obj = df.query('Object=="PageObject"')  
2 page\_obj[page\_obj.Key.str.contains("Box")]

Out[23]:

	Object	Key	Type	SinceVersion	DeprecatedIn	Required	IndirectReference	Inheritable	DefaultValue	
2609	PageObject	MediaBox	rectangle	1.0		TRUE	FALSE	TRUE		
2610	PageObject	CropBox	rectangle	1.0		FALSE	FALSE	TRUE	@MediaBox	
2611	PageObject	BleedBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2612	PageObject	TrimBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2613	PageObject	ArtBox	rectangle	1.3		FALSE	FALSE	FALSE	@CropBox	
2614	PageObject	BoxColorInfo	dictionary	1.4		FALSE	FALSE	FALSE		





# C++ “TestGrammar” file validation

Warning: second/third class key 'AAPL:Keywords' is not defined in Arlington for DocInfo

Error: wrong value: Type (“XObjectImageSoftMask”) should be: name [XObject] and is name=**XObjcect**

Error: wrong value: Type ("FileSpecification") should be: name [Filespec] and is name=**FileSpec**

Error: wrong value: Type ("XObjectImage") should be: name [XObject] and is name=**Xobject**

Error: wrong type: Name (“RichMediaConfiguration”) should be: STRING-TEXT and is NAME

Error: not an indirect reference as required: FontDescriptor ("FontTrueType")

Error: non-inheritable required key doesn't exist: Subtype (“Thumbnail”)

Spelling

Capitalization

Subtype	name	(Required) The type of XObject that this dictionary describes; shall be
	Subtype	name
		(Optional when used only as a thumbnail image, required otherwise) The type of XObject that this dictionary describes; shall be <i>Image</i> for an image XObject.
NOTE The conditions for when the <b>Subtype</b> key is required are given in the document (		

Changed in ISO 32000-2:2020  
→ also changed in Arlington PDF model



# Conclusion

- **First open access, vendor neutral, specification-derived, machine and human readable, comprehensive definition of all PDF objects**
- Easy to understand and use
  - Text-based TSV file sets
  - 12 fields with custom predicates
  - Platform and language agnostic
  - Easily transformable
  - No code / low code / code



Arlington PDF Model

<https://github.com/pdf-association/arlington-pdf-model>



# Thank you Questions?

**[peter.wyatt@pdfa.org](mailto:peter.wyatt@pdfa.org)**

