HTML / Text (with & w/o annotations) Test

Text Extraction

```
$ time pdf2txt.py --all-texts ~/LINGEA/MASAPI/adobe_slides.pdf | pr -tw$COLUMNS -3
...
real    0m9.128s
user    0m9.042s
sys     0m0.096s
```

```
Role of PDF and Open Data                                                                 © 2013 Adobe Systems Incorporated.                                                      1
James C. King   |  Senior Principal Scientist


Outline                                                                                   □ Role of PDF                                                                                   □ PDF envelopes

□ Open Data Paradigm                                                                      □ PDF in the wild                                                                       © 2013 Adobe Systems Incorporated.

□ Who is here and why                                                                     □ PDF purpose-built                                                                     2

□ PDF                                                                                     □ Structured Data


Open Data Paradigm                                                                                                                                                                3
                                              © 2013 Adobe Systems Incorporated.

Providing Open Data

Open Data Paradigm                                                                                                                                                                4
                                              © 2013 Adobe Systems Incorporated.

3rd Party “Processors”

Open Data Paradigm                                                                                                                                                                5
                                              © 2013 Adobe Systems Incorporated.

Other uses of Open Data

Open Data Paradigm                                                                                                                                                                6
                                              © 2013 Adobe Systems Incorporated.

All need tools

Open Data Roles
                                              □ Processors                                                                              © 2013 Adobe Systems Incorporated.
□ Which is your role(s)?
                                              □ Tool Providers                                                                          7
□ Providers
                                              □ Did I miss some roles?
□ Consumers

PDF                                                                                       PDF 1.7
                                              (2006)                                                                            PDF 1.2
PDF introduced by Adobe in June 1993                                                                                                                                              (1996)
                                              PDF 1.6
PDF 1.7 became an ISO Standard in July 2008                                               (2004)                                                                                  PDF 1.3
                                                                                                                                    (1999)
PDF 1.0                                                                                   PDF 1.5
(1993)                                                                                    (2003)                                                                                  © 2013 Adobe Systems Incorporated.
                                              8
PDF 1.1                                                                                                                                                                           ISO Work on PDF is ongoing
(1994)                                                                                    PDF 1.4
                                              (2001)

Role of PDF and Open Data                                                                                                                                                         9
                                              © 2013 Adobe Systems Incorporated.
     •  PDF in the wild
     •  PDF purpose-built

Pre-existing PDFs  (PDF in the wild)                                                      □ Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/           © 2013 Adobe Systems Incorporated.

□ PDFs abound containing useful content                                                           □ If pages are textual (including tables) – can extract that text/tables                  10

□ but, PDF is a document format not a data format                                         □ see, Wikipedia entry for “List of PDF Software”

□ If pages contain graphics – can extract those graphics                                          □ If pages are images – must turn to OCR technology

□ Vector graphics: use Adobe Illustrator                                                          □ see, Wikipedia entry for “Comparison of optical character recognition software”


Purpose Built PDFs – Structured PDFs                                                      □ Content extraction tools can make use of this structure while extracting content        11

□ ISO Standard allows for optional structural information to be added to PDFs for        □ Structure best obtained from authoring tool (e.g., document processing tools)

□ reading order                                                                                   □ Can be added after-the-fact
□ tagging information (headings, footnotes, figures, math)                        © 2013 Adobe Systems Incorporated.

Purpose Built PDFs – PDF Attachments                                                □ Attach icon to page to select attachment
                                                                                                                          12
□ ISO Standard defines attachments to PDF files                                           □ Here is a sample of using attachments for datasets used in a presentation

□ attachments get compressed using same lossless technology as ZIP and PNG          © 2013 Adobe Systems Incorporated.


PDF Enveloping
                                             Schema                                                                           © 2013 Adobe Systems Incorporated.
□ Raw data needs defining information
                                             □ PDF can provide 1. and include data and schema as attachment                             13
1. documentation for source, ownership, semantics
                                             □ typical XML file gets reduced by an order of magnitude
2. schema for syntax
                                             □ PDF document features cover the attachments (authenticity, signatures, forms)
3. proof of authenticity
                                             □ Attachments easily extracted from mother PDF
Descriptive PDF
                                             □ see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf
XML or CSV file

References to more about PDF                                                        □ PDF package example:        http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_201 http://www.adobe.com/technology/people/san-jose/jim-king.htm

□ PDF attachment example:  http://www.w3.org/2013/04/odw/EducationalAttainment.pdf   □ My PDF blog:       http://blogs.adobe.com/insidepdf                                         © 2013 Adobe Systems Incorporated.

□ Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html □ My tutorial on what is inside of a PDF file:                                            14

□ Using the Acrobat web capture feature to convert HTML to PDF (14 pages)           http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_L

□ All of the 8 dataset files were downloaded and added to this PDF as attachments   □ Other presentations and papers by me:


© 2013 Adobe Systems Incorporated.
```

```
$ time pdftotext -layout ~/LINGEA/MASAPI/adobe_slides.pdf - | pr -tw$COLUMNS -2
...
real    0m0.198s
user    0m0.178s
sys     0m0.029s
```

```
       Role of PDF and Open Data
       James C. King | Senior Principal Scientist
                                                                   © 2013 Adobe Systems Incorporated.            1


  Outline                                                                                                                          □      PDF in the wild
                                                                       □      PDF purpose-built
  □     Open Data Paradigm                                                                                                                □ Structured Data
    □      Who is here and why                                                                                                            □ PDF envelopes



  □     PDF


                                                                   © 2013 Adobe Systems Incorporated.   2

  □     Role of PDF


  Open Data Paradigm


Providing Open Data
                                                                   © 2013 Adobe Systems Incorporated.   3


  Open Data Paradigm


3rd Party “Processors”
                                                                   © 2013 Adobe Systems Incorporated.   4


  Open Data Paradigm


Other uses of Open Data
                                                                   © 2013 Adobe Systems Incorporated.   5


  Open Data Paradigm


All need tools
                                                                   © 2013 Adobe Systems Incorporated.   6


  Open Data Roles
                                                                     □      Did I miss some roles?
  □     Which is your role(s)?
    □      Providers
    □      Consumers
    □      Processors
    □      Tool Providers                                                                                                © 2013 Adobe Systems Incorporated.   7


                                PDF                                                                                                                                                             (1996)
                    PDF introduced by Adobe in June 1993                                                                                                                                    PDF 1.3
                     PDF 1.7 became an ISO Standard in July 2008                                                                                                                             (1999)
```

```
                                                                                        PDF 1.7                       PDF 1.4
                                                                                         (2006)   PDF 1.6 PDF   1.5
                                                                                                   (2003)
                                                      PDF 1.0                                                                                                  (2001)
                                                    (1993)                                                                                 (2004)       8

                                                    PDF 1.1
                                                    (1994)
                                                                                                                  ISO Work on PDF is ongoing
                                              PDF 1.2                                 © 2013 Adobe Systems Incorporated.

               Role of PDF and Open Data                                                                                                                             • PDF purpose-built

                     • PDF in the wild                                                                                        © 2013 Adobe Systems Incorporated.            9

  Pre-existing PDFs (PDF in the wild)
                                                                      □     If pages are textual (including tables) – can extract that text/tables
  □     PDFs abound containing useful content                                                                                                       □      see, Wikipedia entry for “List of PDF Software”
    □      but, PDF is a document format not a data format

                                                                      □     If pages are images – must turn to OCR technology
  □     If pages contain graphics – can extract those graphics                                                                                              □      see, Wikipedia entry for “Comparison of optical character recognition software”
    □      Vector graphics: use Adobe Illustrator
    □      Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/
                                                                    © 2013 Adobe Systems Incorporated.                    10

  Purpose Built PDFs – Structured PDFs
                                                                      □     Structure best obtained from authoring tool (e.g., document processing tools)
  □     ISO Standard allows for optional structural information to be added to PDFs for
    □      reading order
    □      tagging information (headings, footnotes, figures, math)                                                                           □     Can be added after-the-fact


  □     Content extraction tools can make use of this structure while extracting content

                                                                    © 2013 Adobe Systems Incorporated.                    11

  Purpose Built PDFs – PDF Attachments

  □     ISO Standard defines attachments to PDF files                                                                                         □     Here is a sample of using attachments for datasets used in a presentation
    □    attachments get compressed using same lossless technology as ZIP and PNG



  □     Attach icon to page to select attachment                                                                                                    © 2013 Adobe Systems Incorporated.                    12

  PDF Enveloping                                                                                                                                    □       typical XML file gets reduced by an order of magnitude
                                                                        □       PDF document features cover the attachments (authenticity, signatures, forms)
                                             Descriptive PDF
  □     Raw data needs defining information
                                  XML or CSV file                                                         □      Attachments easily extracted from mother PDF
    1. documentation for source, ownership, semantics
    2. schema for syntax                                                 Schema
                                                                      □     see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf
    3. proof of authenticity

                                                                    © 2013 Adobe Systems Incorporated.                   13
  □     PDF can provide 1. and include data and schema as attachment

  References to more about PDF                                                                                                                □     My PDF blog: http://blogs.adobe.com/insidepdf
                                                                      □     My tutorial on what is inside of a PDF file:
  □     PDF attachment example: http://www.w3.org/2013/04/odw/EducationalAttainment.pdf                                                                     http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.p
    □      Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html                                           □     Other presentations and papers by me:
    □      Using the Acrobat web capture feature to convert HTML to PDF (14 pages)                                                                  http://www.adobe.com/technology/people/san-jose/jim-king.htm
    □      All of the 8 dataset files were downloaded and added to this PDF as attachments


  □     PDF package example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf
                                                                    © 2013 Adobe Systems Incorporated.                              14

© 2013 Adobe Systems Incorporated.
```

HTML Extraction

```
$ time pdf2txt.py --layoutmode loose --output_type html -A ~/LINGEA/MASAPI/adobe_slides.pdf > html1.html

real    0m9.472s
user    0m9.396s
sys     0m0.076s
$ du -ch html1.html
208K    html1.html
208K    total
```

Page: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Page 1

Role of PDF and Open Data James C. King

| Senior Principal Scientist

Page 2

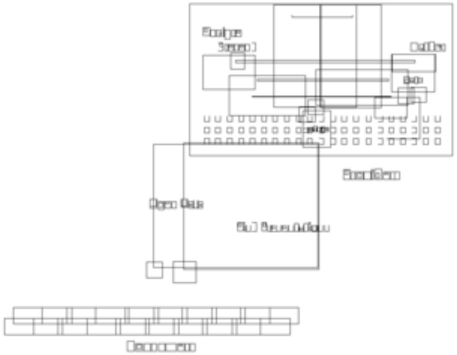
Outline

- Open Data Paradigm
  - Who is here and why
- PDF
- Role of PDF
  - PDF in the wild
  - PDF purpose-built
    - Structured Data
    - PDF envelopes

Page 3

Open Data Paradigm

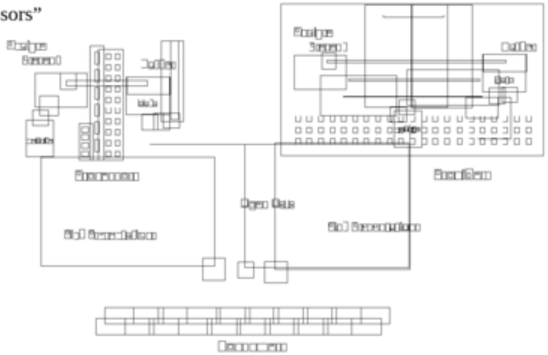
Providing Open Data



Page 4

Open Data Paradigm

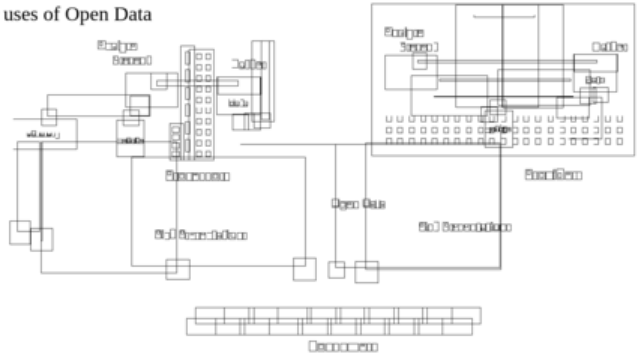
3rd Party “Processors”



Page 5

Open Data Paradigm

Other uses of Open Data



Page 6

Open Data Paradigm

All need tools

Page 7

Open Data Roles

- Which is your role(s)?
  - Providers
  - Consumers
  - Processors
  - Tool Providers

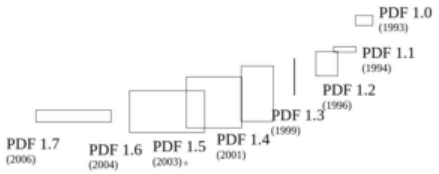
- Did I miss some roles?

Page 8

PDF

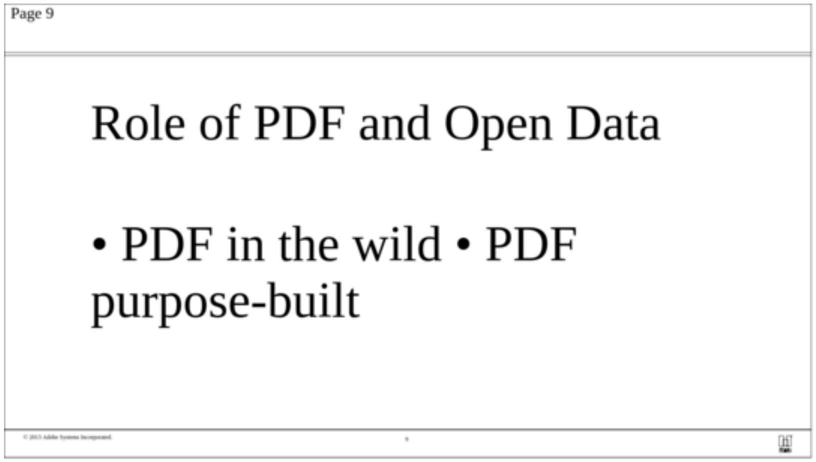
PDF introduced by Adobe in June 1993

PDF 1.7 became an ISO Standard in July 2008



ISO Work on PDF is ongoing

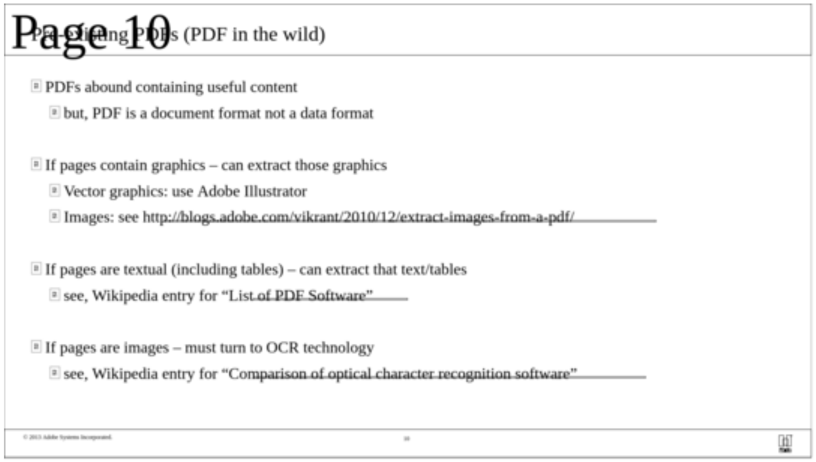
Page 9



Role of PDF and Open Data

• PDF in the wild • PDF purpose-built

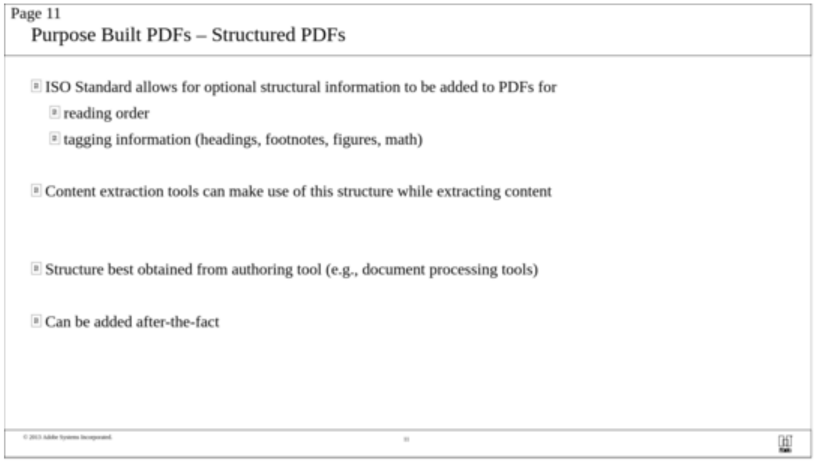
Page 10



Pre-existing PDFs (PDF in the wild)

- PDFs abound containing useful content
  - but, PDF is a document format not a data format
- If pages contain graphics – can extract those graphics
  - Vector graphics: use Adobe Illustrator
  - Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/
- If pages are textual (including tables) – can extract that text/tables
  - see, Wikipedia entry for "List of PDF Software"
- If pages are images – must turn to OCR technology
  - see, Wikipedia entry for "Comparison of optical character recognition software"

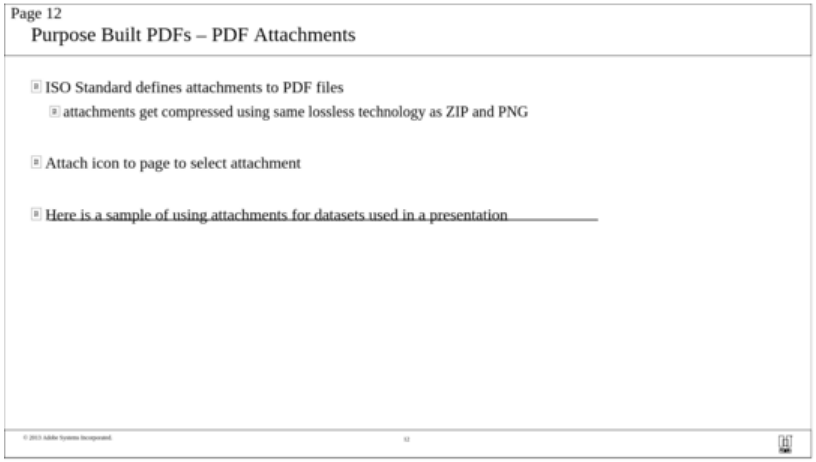
Page 11



Purpose Built PDFs – Structured PDFs

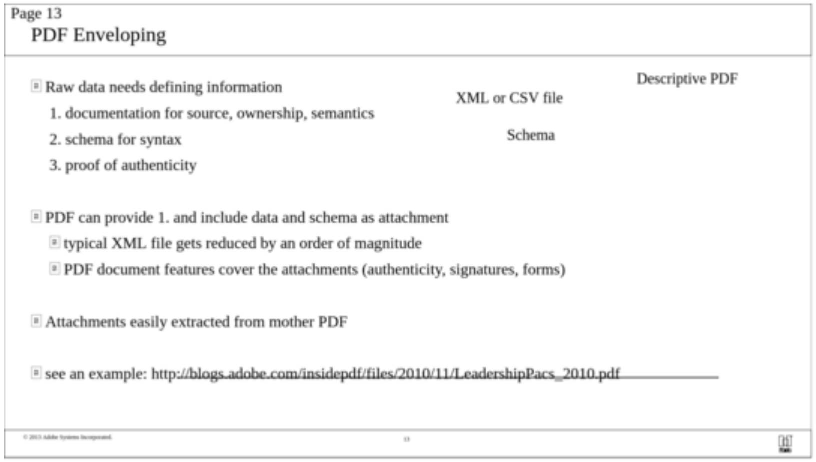
- ISO Standard allows for optional structural information to be added to PDFs for
  - reading order
  - tagging information (headings, footnotes, figures, math)
- Content extraction tools can make use of this structure while extracting content
- Structure best obtained from authoring tool (e.g., document processing tools)
- Can be added after-the-fact

Page 12



Purpose Built PDFs – PDF Attachments

- ISO Standard defines attachments to PDF files
  - attachments get compressed using same lossless technology as ZIP and PNG
- Attach icon to page to select attachment
- Here is a sample of using attachments for datasets used in a presentation



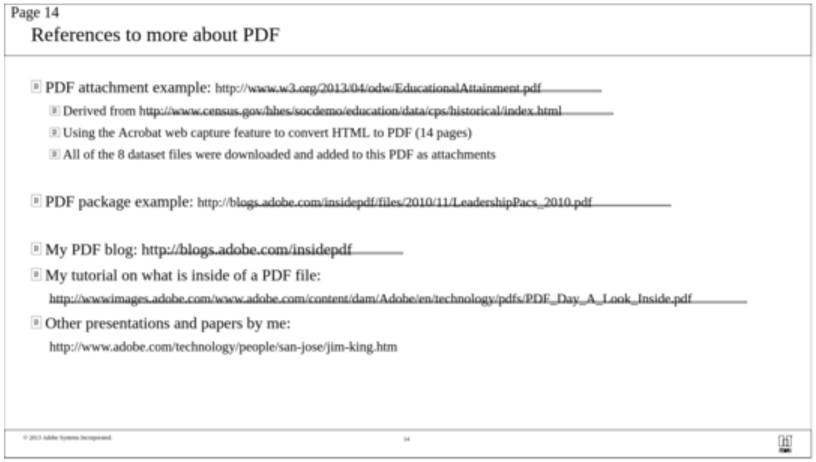
Page 13

PDF Enveloping

- Raw data needs defining information
  1. documentation for source, ownership, semantics
  2. schema for syntax
  3. proof of authenticity
- PDF can provide 1. and include data and schema as attachment
  - typical XML file gets reduced by an order of magnitude
  - PDF document features cover the attachments (authenticity, signatures, forms)
- Attachments easily extracted from mother PDF
- see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

© 2013 Adobe Systems Incorporated.

13



Page 14

References to more about PDF

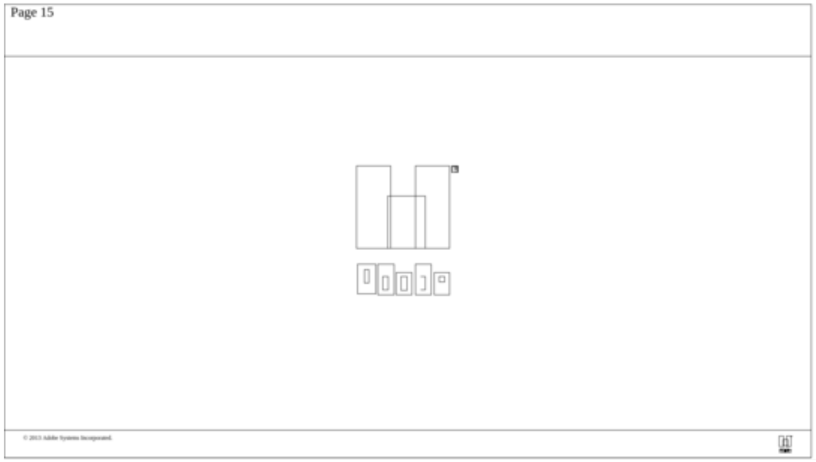
- PDF attachment example: http://www.w3.org/2013/04/odw/EducationalAttainment.pdf
  - Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html
  - Using the Acrobat web capture feature to convert HTML to PDF (14 pages)
  - All of the 8 dataset files were downloaded and added to this PDF as attachments
- PDF package example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf
- My PDF blog: http://blogs.adobe.com/insidepdf
- My tutorial on what is inside of a PDF file:

  http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.pdf
- Other presentations and papers by me:

  http://www.adobe.com/technology/people/san-jose/jim-king.htm

© 2013 Adobe Systems Incorporated.

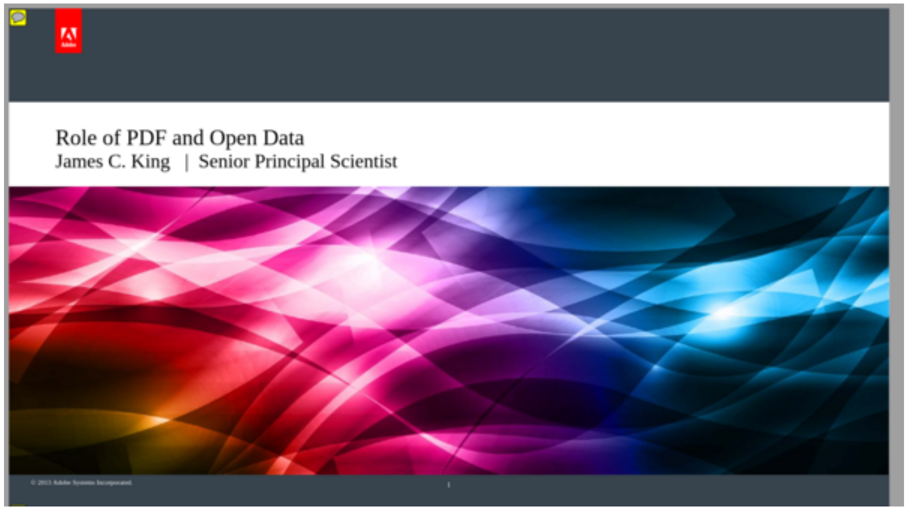
14



Page 15

© 2013 Adobe Systems Incorporated.

```
$ time pdftohtml  -c -s -noframes ~/LINGEA/MASAPI/adobe_slides.pdf > html2.html

real    0m17.805s
user    0m17.754s
sys     0m0.040s
$ du -ch html2*
36K     html2.html
620K    html2001.png
...
16K     html2015.png
1.6M    total
```

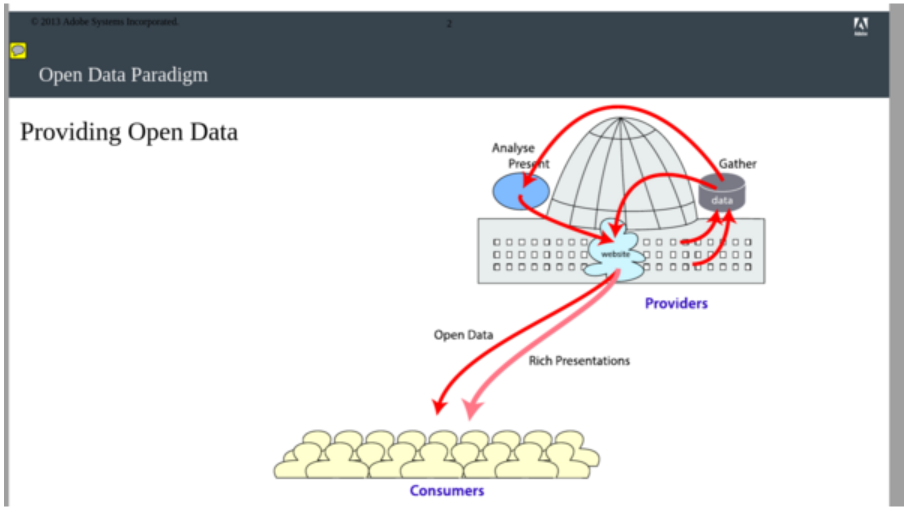


Role of PDF and Open Data

James C. King | Senior Principal Scientist

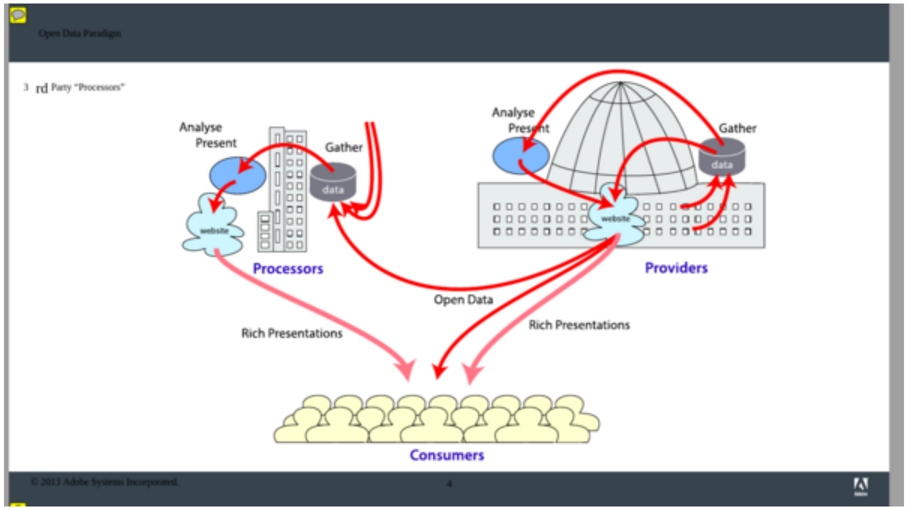
Outline

- Open Data Paradigm
  - Who is here and why
- PDF
- Role of PDF
  - PDF in the wild
  - PDF purpose-built
    - Structured Data
    - PDF envelopes



Open Data Paradigm

Providing Open Data

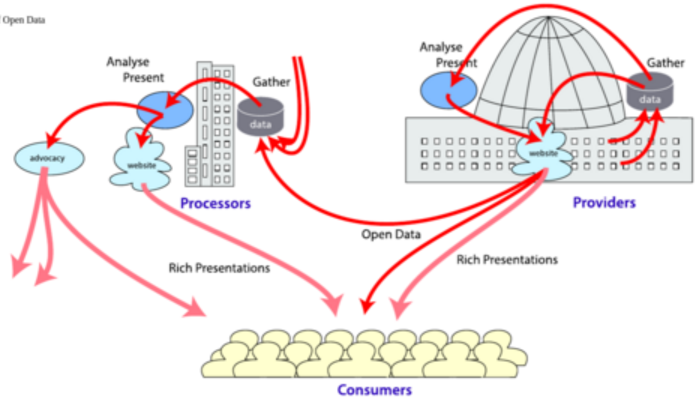


Open Data Paradigm

3 rd Party "Processors"

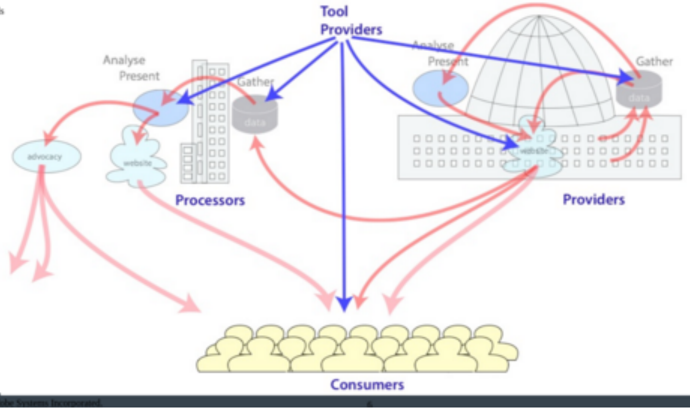
Open Data Paradigm

Other uses of Open Data



Open Data Paradigm

All need tools



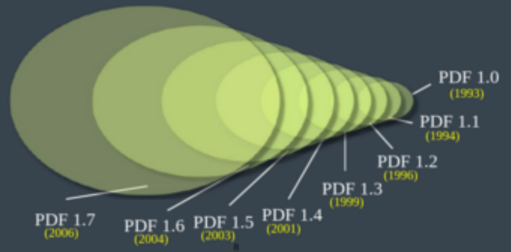
Open Data Roles

- Which is your role(s)?
  - Providers
  - Consumers
  - Processors
  - Tool Providers
- Did I miss some roles?

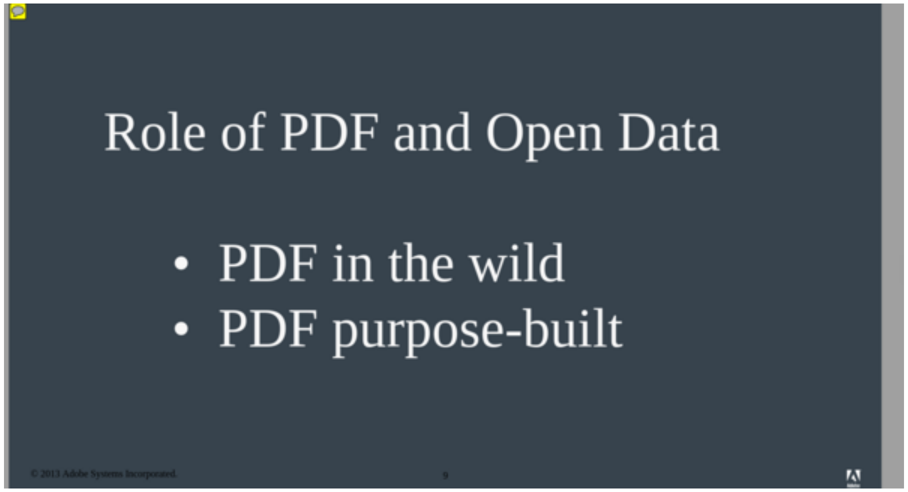
PDF

PDF introduced by Adobe in June 1993

PDF 1.7 became an ISO Standard in July 2008

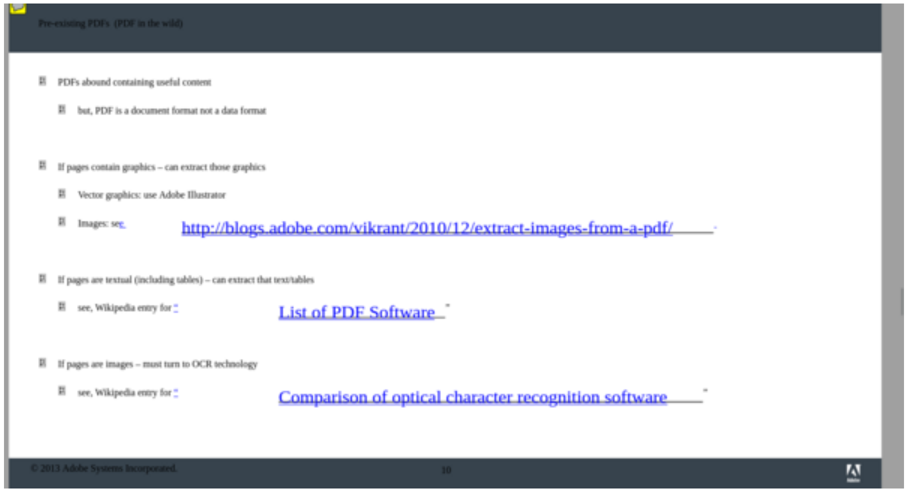


ISO Work on PDF is ongoing



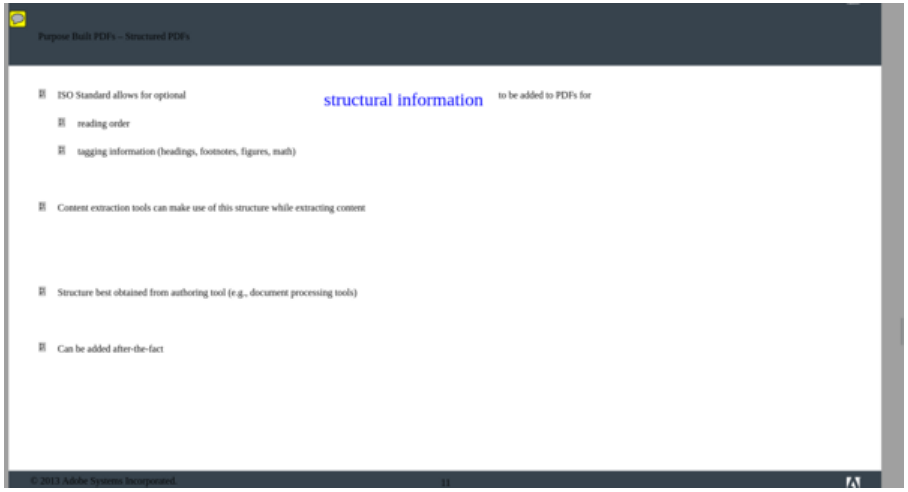
Role of PDF and Open Data

- PDF in the wild
- PDF purpose-built



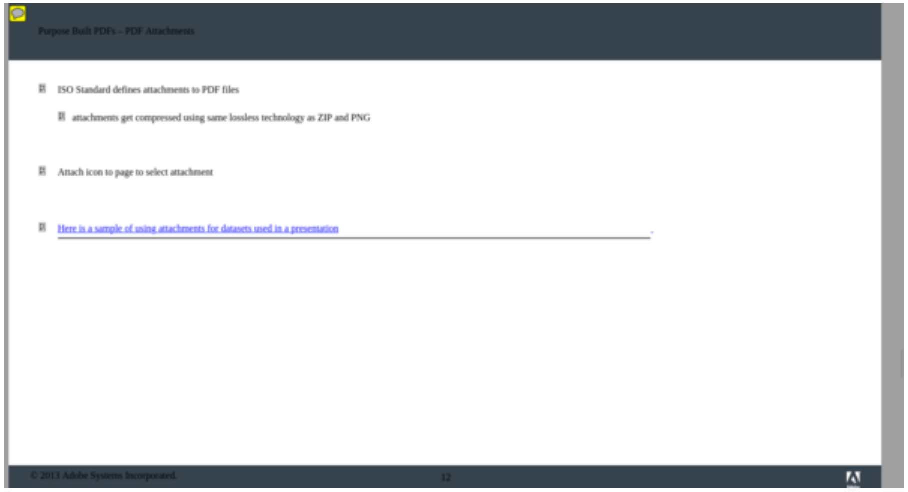
Pre-existing PDFs (PDF in the wild)

- PDFs abound containing useful content
  - but, PDF is a document format not a data format
- If pages contain graphics – can extract those graphics
  - Vector graphics: use Adobe Illustrator
  - Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/.
- If pages are textual (including tables) – can extract that text/tables
  - see, Wikipedia entry for "List of PDF Software"
- If pages are images – must turn to OCR technology
  - see, Wikipedia entry for "Comparison of optical character recognition software"



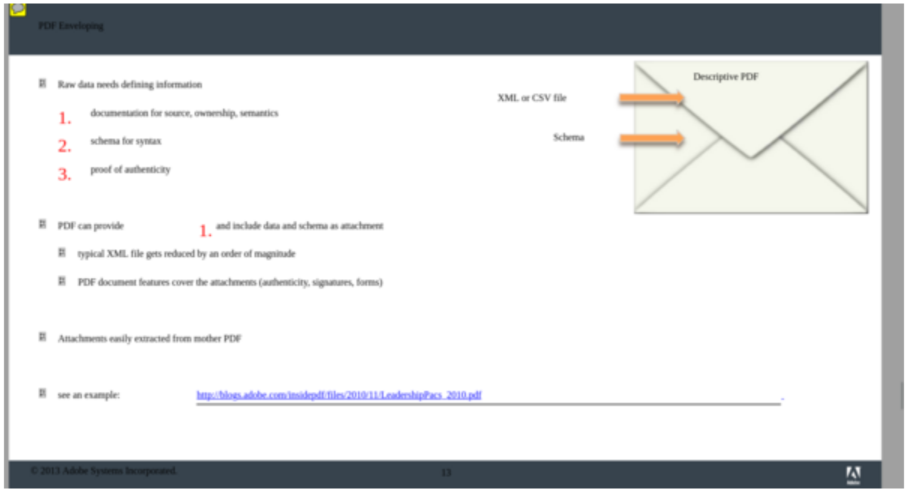
Purpose Built PDFs – Structured PDFs

- ISO Standard allows for optional structural information to be added to PDFs for
  - reading order
  - tagging information (headings, footnotes, figures, math)
- Content extraction tools can make use of this structure while extracting content
- Structure best obtained from authoring tool (e.g., document processing tools)
- Can be added after-the-fact



Purpose Built PDFs – PDF Attachments

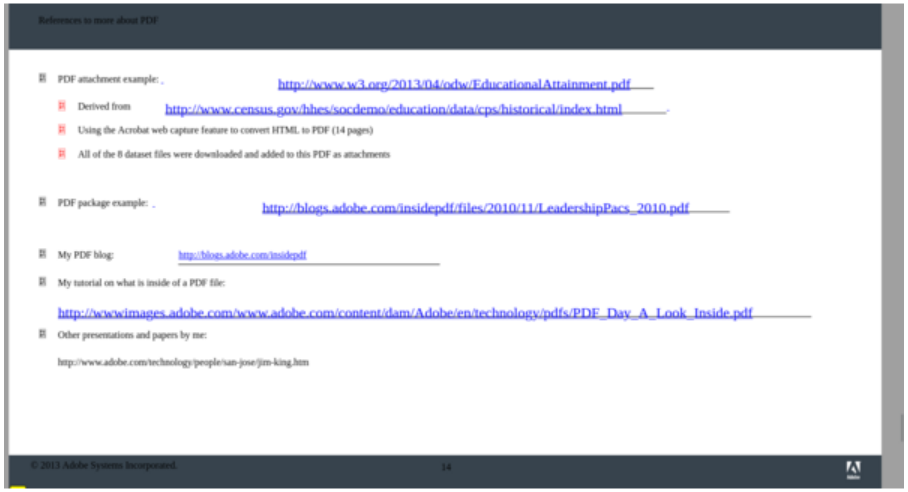
- ISO Standard defines attachments to PDF files
  - attachments get compressed using same lossless technology as ZIP and PNG
- Attach icon to page to select attachment
- Here is a sample of using attachments for datasets used in a presentation.



PDF Enveloping

- Raw data needs defining information
  1. documentation for source, ownership, semantics
  2. schema for syntax
  3. proof of authenticity
- PDF can provide 1. and include data and schema as attachment
  - typical XML file gets reduced by an order of magnitude
  - PDF document features cover the attachments (authenticity, signatures, forms)
- Attachments easily extracted from mother PDF
- see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

© 2013 Adobe Systems Incorporated. 13



References to more about PDF

- PDF attachment example: http://www.w3.org/2013/04/odw/EducationalAttainment.pdf
  - Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html
  - Using the Acrobat web capture feature to convert HTML to PDF (14 pages)
  - All of the 8 dataset files were downloaded and added to this PDF as attachments
- PDF package example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf
- My PDF blog: http://blogs.adobe.com/insidepdf
- My tutorial on what is inside of a PDF file:

  http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.pdf
- Other presentations and papers by me:

  http://www.adobe.com/technology/people/san-jose/jim-king.htm

© 2013 Adobe Systems Incorporated. 14

© 2013 Adobe Systems Incorporated.

Document Outline

Annotations

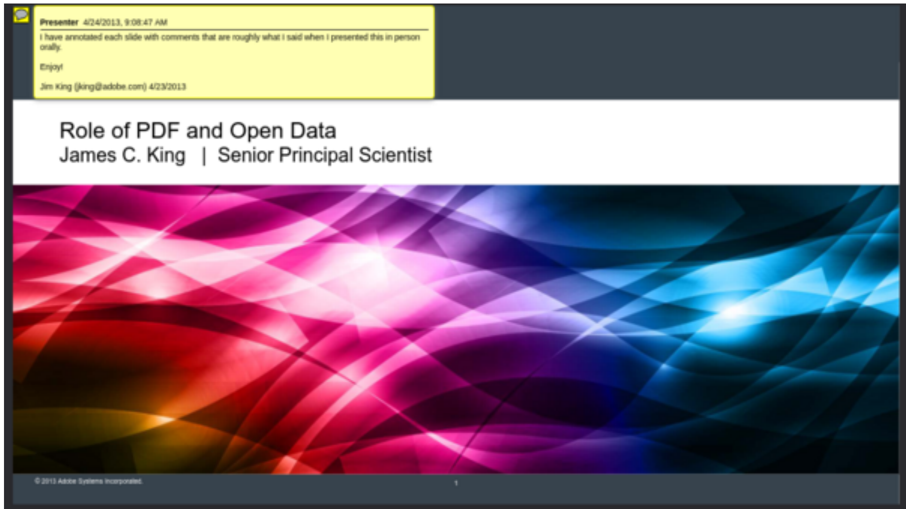
```
$ time leela annots adobe_slides.pdf /dev/stdout > leela.xml

real	0m0.059s
user	0m0.050s
sys	0m0.009s
```

```
<annot page="0" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>9d19d5d0-cc06-452b-a980-7a7497b07b9c</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
</text>g (jking@adobe.com) 4/23/2013with comments that are roughly what I said when I presented this in person orally.
</annot>
<annot page="1" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>3d32e856-dd6c-49bf-832f-4b06fa874e70</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>This is the outline of what I plan to say.  I am really curious as to who the workshop attendees are and what part of world of open data they are wor</text>talk in
general about PDF and it history (very briefly) and finally turn to the material reflected in the title of the presentation.
</annot>
<annot page="2" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>09a93474-6bbb-4514-a518-357a42528015</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>The push for supplying open data has been primarily directed at government agencies. They supply nicely formatted rich presentations as well as the oI like to think
of the standard definitions of “data” and “information” where the data is basically raw material collected and information is something more So this slide identifies two
roles in my open data paradigm: the provider of data and information and the consumer of it.  I think of the consumer as any cit</text>o is interested in the information
collected and disseminated by their governments.  This also applies to entities other than government agencies.
</annot>
<annot page="3" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>0cdf1ceb-1a64-48bf-bcf8-2792488bded5</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>This slide introduces what I call “processors”, those people or organizations to take the open data and process it in some manner to add value or turIn my mind the
ration of processors to consumers is 1 to 10000 or something like that.  Anyone who surf the web is a potential consumer but only a relatively small number of people want to
analyse and make presentation from the open data provided.  Ultimately, at least on this slide, we show that the processors also create rich presentations to supply to the
consumers.</text>
</annot>
<annot page="4" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>cb56508b-c1e8-4cbc-b72a-ce77474caac1</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>Of course, we still don’t know all the uses for which open data may be used. I have lumped these activities into advocacy since I suspect that that iSo I have
introduced three roles in my open data paradigm: providers, consumers and processors.</text>
</annot>
<annot page="5" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>0bc9484d-82af-47a8-addf-115849a7e8ed</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>Of course, all of these people need tools to accomplish their objectives. And I think that the tools for this paradigm are far from mature and ideal.At the workshop
people suggested that standards or to keep the terminology consistent, people who work on standards, represent another important role.  I thi</text> is true.
</annot>
<annot page="6" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>98ac1bd9-fde0-4a75-998d-c9d943b2e39e</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>I did a poll of the audience and it seemed that a lot of people raised their hands for each role.  I rated Providers at about 20% of the audience, Consumers at about
70%, Processor at maybe 60% and Tool providers (to my surprise) at about 70%.
Did anyone else have better estimates.  This happened really quickly and what I wrote about might be way off.</text>
</annot>
<annot page="7" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>aa6cb1d2-d806-44a2-9098-03e4ca2dd2de</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>OK.  Just some basic level setting about PDF.  In June 2013 PDF will be 20 years old.  Adobe introduced PDF in 1993 as the file format supported by its Acrobat
product line.  Based on some experience we had with PostScript being both a file format and a product, we decided with PDF to make a clear distinction between the file format
(PDF) and the Adobe products that support it (Acrobat).
We trademarked Acrobat but did not trademark PDF.  We published the full specification of PDF 1.0 in June 1993 in paperback form and wanted other to develop Through the years
as function was added to Adobe’s Acrobat and its use of PDF we revised the PDF specification and published it each time.  So the 8th version (PDF 1.7) was published by Adobe
in 2006.  In 2007 Adobe inked an agreement with AIIM and ISO to hand over control of PDF to ISO.  We had always gotten complaints that had total control over what new things
went into each new version.  With ISO owning the specification, then the world would have control over thThis picture is my attempt to depict that a file made to conform to
PDF 1.0 or 1.4 also conforms to PDF 1.7.  We did not want to ever obsolete any existing PISO published its first PDF standard ISO 32000-1 in July 2008.  (Yes, the public has
“owned” the standard for almost 5 years.  ISO is now working on PDF 2.0 which should come out in a year or so. </text>
</annot>
<annot page="8" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>f24c998d-361a-4978-ad79-e8d838a91f4d</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>OK.  Finally we get to the material suggest by the title of the talk.
We want to clearly distinguish between the two kinds of PDFs: existing, in the wild, PDFs and ones that are newly purpose built.</text>
</annot>
<annot page="9" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>1dd6e385-e7fa-4631-b2ad-3a3f9094fa94</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>There are billions of pre-existing PDFs that contain a tremendous amount of very valuable and interesting content.  Wouldn’t it be nice if a lot of that content could
become open data.
Well that is asking a lot since PDF was formulated 20 years ago as “Portable Document Format”.  It is not a data format.
However, it is possible to extract the content from PDF files, but a lot of the software available to do this has its limitations.  Further there is often the stated
requirement that such software must be open source software. I’m not sure I see where open data requires open software but I guess if you are an opeThe most compact PDFs and
the ones most amenable to harvesting data from are ones that have text as text, vector graphics as graphics and only those parts that must be images, be images.  If you
create a PDF file by scanning paper pages then the simplest systems will produce PDFs where each page is a full page imI give the best references I could find for processing
the different kinds of content found in PDFs.  I would take a lot of flake it I started telling you which software, besides Adobe’s, is the best for each job. ( Besides I
really don’t know.)</text>
</annot>
<annot page="9" index="1" type="link">
  <rect x1="190.183000" y1="279.035000" x2="599.863000" y2="301.355000"/>
</annot>
<annot page="9" index="2" type="link">
  <rect x1="599.863000" y1="279.035000" x2="606.523000" y2="301.355000"/>
</annot>
<annot page="9" index="3" type="link">
  <rect x1="606.523000" y1="279.035000" x2="670.903000" y2="301.355000"/>
</annot>
<annot page="9" index="4" type="link">
  <rect x1="670.903000" y1="279.035000" x2="677.563000" y2="301.355000"/>
</annot>
<annot page="9" index="5" type="link">
  <rect x1="677.563000" y1="279.035000" x2="717.463000" y2="301.355000"/>
</annot>
<annot page="9" index="6" type="link">
  <rect x1="717.463000" y1="279.035000" x2="724.123000" y2="301.355000"/>
</annot>
<annot page="9" index="7" type="link">
  <rect x1="724.123000" y1="279.035000" x2="735.223000" y2="301.355000"/>
```

```
</annot>
<annot page="9" index="8" type="link">
  <rect x1="735.223000" y1="279.035000" x2="741.883000" y2="301.355000"/>
</annot>
<annot page="9" index="9" type="link">
  <rect x1="741.883000" y1="279.035000" x2="775.123000" y2="301.355000"/>
</annot>
<annot page="9" index="10" type="link">
  <rect x1="296.863000" y1="186.035000" x2="333.463000" y2="208.355000"/>
</annot>
<annot page="9" index="11" type="link">
  <rect x1="333.463000" y1="186.035000" x2="401.203000" y2="208.355000"/>
</annot>
<annot page="9" index="12" type="link">
  <rect x1="401.203000" y1="186.035000" x2="480.043000" y2="208.355000"/>
</annot>
<annot page="9" index="13" type="link">
  <rect x1="296.863000" y1="93.035400" x2="763.363000" y2="115.355000"/>
</annot>
<annot page="10" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>bc7d9ec7-a6e8-438d-9352-bdf50671b69d</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>It seems that since day one, people have wanted to extract content from PDF files.  Given that it was designed to be a presentation format, any information as to
reading order of the text strings found in the file, what purpose some text supports like being a heading at some level or a footnote or …  was So, structured PDF was
invented quite a few years ago and a particular form of that called Tagged PDF was also invented.  These add tagging information, optionally, to the PDF file to assist
software that wants extract the content with more structural properties.  The structure is also very essential for producing PDFs that are accessible (like for blind people
to have them read aloud).
There are a lot of structured PDF file in existence because of the accessibility requirements of our governments, and because Adobe’s software includes it whenever possible.
</text>
</annot>
<annot page="11" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>a40723bb-fd0d-4be7-9658-7185745044e7</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
I think it is cool to attach a XLS or CSV file for each chart or other rendering that uses data.  One can add an annotation that can be clicked upon to retriI have
hyperlinked to a great example I made by copying some web pages produced by the US Bureau of Labor Statistics.  Take a look!</text>
</annot>
<annot page="11" index="1" type="link">
  <rect x1="54.283000" y1="281.435000" x2="705.883000" y2="303.755000"/>
</annot>
<annot page="12" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>a2cfa9e5-ca2c-48cc-a842-ffc34efb4163</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
  <text>Given that PDFs can become an envelope for attachments, and given that data sets don’t really stand on their own, I think a great idea is to use the I have also
created an example of this use which you should look at. </text>ses significantly.undle.  order to properly process it.
</annot>
<annot page="12" index="1" type="link">
  <rect x1="206.383000" y1="93.035400" x2="239.743000" y2="115.355000"/>
</annot>
<annot page="12" index="2" type="link">
  <rect x1="239.743000" y1="93.035400" x2="849.223000" y2="115.355000"/>
</annot>
<annot page="13" index="0" type="link">
  <rect x1="296.263000" y1="434.855000" x2="709.483000" y2="452.795000"/>
</annot>
<annot page="13" index="1" type="link">
  <rect x1="172.063000" y1="407.735000" x2="723.283000" y2="425.675000"/>
```

```
</annot>
<annot page="13" index="2" type="link">
  <rect x1="278.683000" y1="299.075000" x2="792.943000" y2="317.015000"/>
</annot>
<annot page="13" index="3" type="link">
  <rect x1="186.463000" y1="240.995000" x2="474.043000" y2="263.315000"/>
</annot>
<annot page="13" index="4" type="link">
  <rect x1="54.283000" y1="183.755000" x2="80.923000" y2="201.695000"/>
</annot>
<annot page="13" index="5" type="link">
  <rect x1="80.923000" y1="183.755000" x2="672.103000" y2="201.695000"/>
</annot>
<annot page="13" index="6" type="link">
  <rect x1="672.103000" y1="183.755000" x2="882.823000" y2="201.695000"/>
</annot>
<annot page="14" index="0" type="text">
  <rect x1="1.000020" y1="521.000000" x2="19.000000" y2="539.000000"/>
  <name>c1843d36-ebec-45a5-8d29-7588ae3b70ce</name>
  <color r="65535" g="65535" b="0"/>
  <label>Presenter</label>
  <subject>Presentation Notes</subject>
  </markup>
Jim King</text>d something valuable from this.you interesting things about PDF (and I have done that in the past).
</annot>
```

Original Slides



Presenter 4/24/2013, 9:08:47 AM

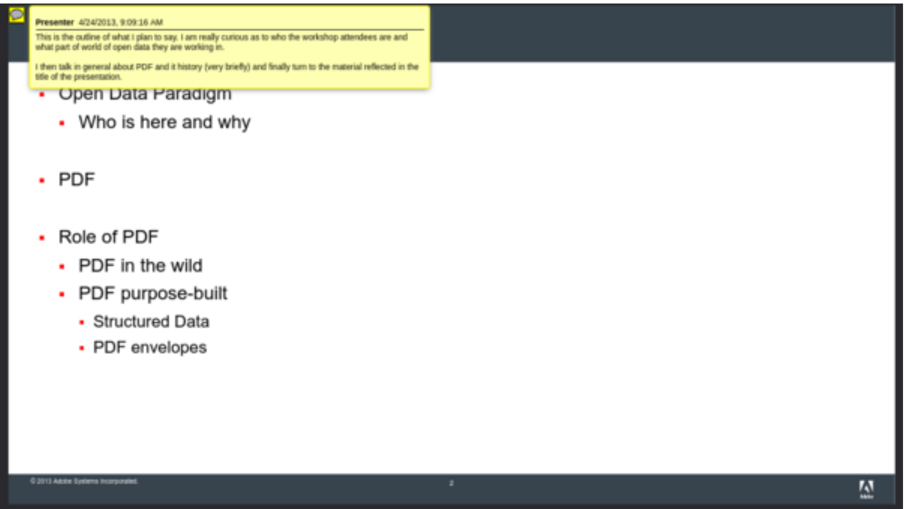
I have annotated each slide with comments that are roughly what I said when I presented this in person orally.

Enjoy!

Jim King (jking@adobe.com) 4/23/2013

Role of PDF and Open Data

James C. King | Senior Principal Scientist

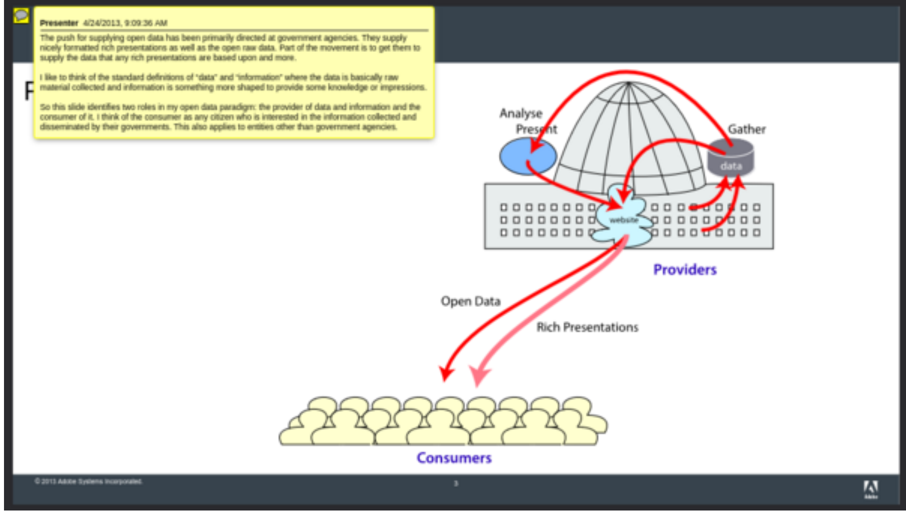


Presenter 4/24/2013, 9:09:16 AM

This is the outline of what I plan to say. I am really curious as to who the workshop attendees are and what part of world of open data they are working in.

I then talk in general about PDF and it history (very briefly) and finally turn to the material reflected in the title of the presentation.

- Open Data Paradigm
  - Who is here and why
- PDF
- Role of PDF
  - PDF in the wild
  - PDF purpose-built
    - Structured Data
    - PDF envelopes

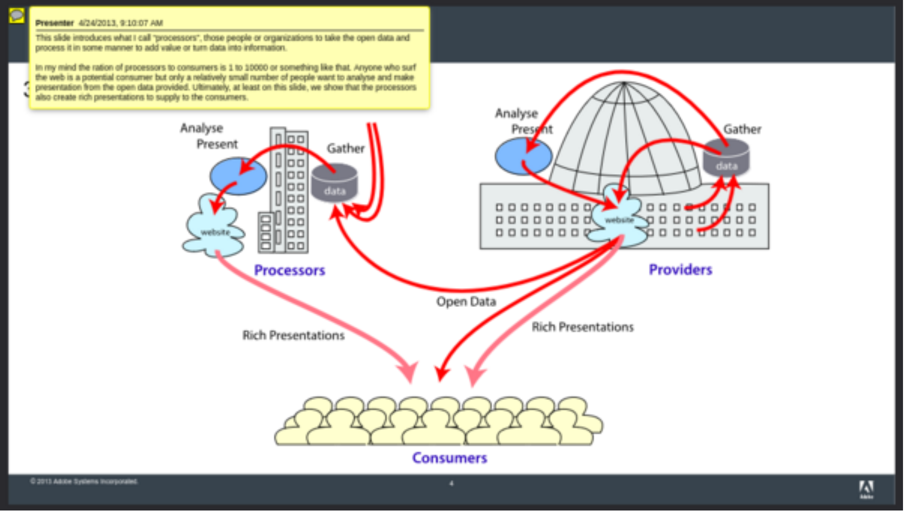


Presenter 4/24/2013, 9:09:36 AM

The push for supplying open data has been primarily directed at government agencies. They supply nicely formatted rich presentations as well as the open raw data. Part of the movement is to get them to supply the data that any rich presentations are based upon and more.

I like to think of the standard definitions of "data" and "information" where the data is basically raw material collected and information is something more shaped to provide some knowledge or impressions.

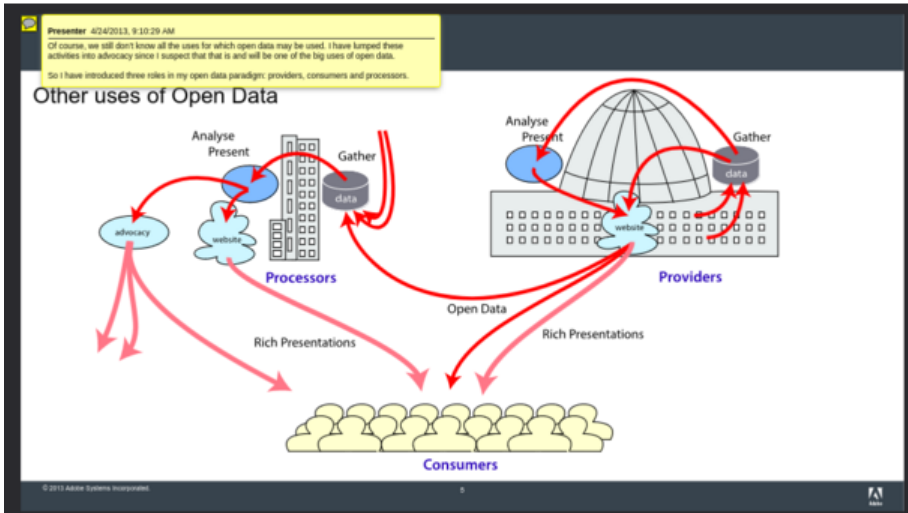
So this slide identifies two roles in my open data paradigm: the provider of data and information and the consumer of it. I think of the consumer as any citizen who is interested in the information collected and disseminated by their governments. This also applies to entities other than government agencies.



Presenter 4/24/2013, 9:10:07 AM

This slide introduces what I call "processors", those people or organizations to take the open data and process it in some manner to add value or turn data into information.

In my mind the ration of processors to consumers is 1 to 10000 or something like that. Anyone who surf the web is a potential consumer but only a relatively small number of people want to analyse and make presentation from the open data provided. Ultimately, at least on this slide, we show that the processors also create rich presentations to supply to the consumers.

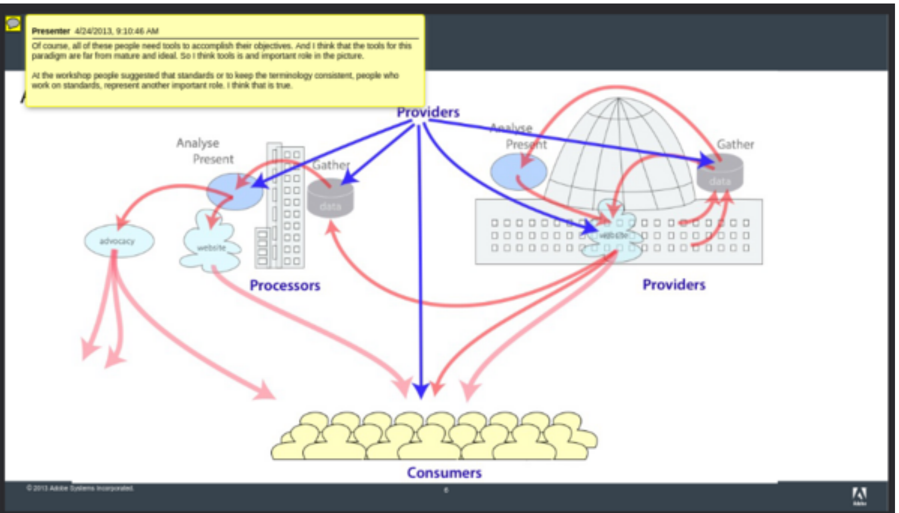


Presenter 4/24/2013, 9:10:29 AM

Of course, we still don't know all the uses for which open data may be used. I have lumped these activities into advocacy since I suspect that that is and will be one of the big uses of open data.

So I have introduced three roles in my open data paradigm: providers, consumers and processors.

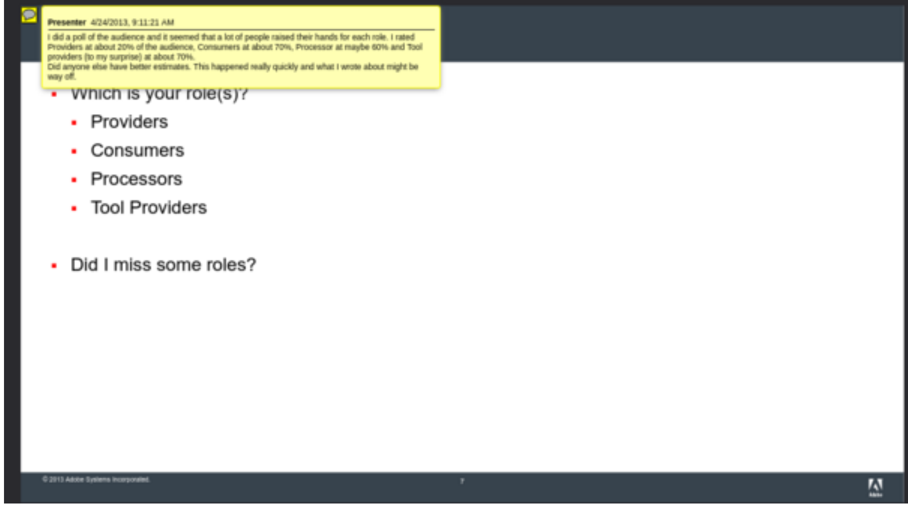
Other uses of Open Data



Presenter 4/24/2013, 9:10:46 AM

Of course, all of these people need tools to accomplish their objectives. And I think that the tools for this paradigm are far from mature and ideal. So I think tools is and important role in the picture.

At the workshop people suggested that standards or to keep the terminology consistent, people who work on standards, represent another important role. I think that is true.

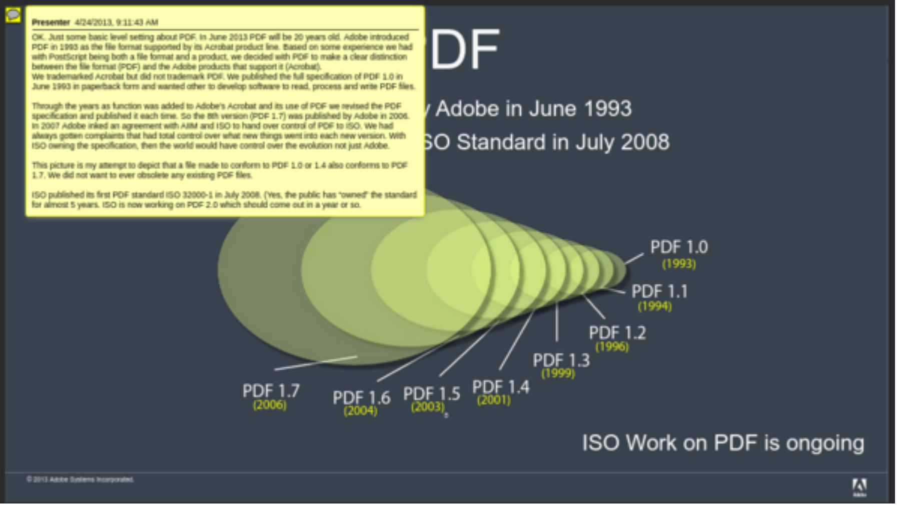


Presenter 4/24/2013, 9:11:21 AM

I did a poll of the audience and it seemed that a lot of people raised their hands for each role. I rated Providers at about 20% of the audience, Consumers at about 70%, Processor at maybe 60% and Tool providers (to my surprise) at about 70%.
Did anyone else have better estimates. This happened really quickly and what I wrote about might be way off.

- Which is your role(s)?
  - Providers
  - Consumers
  - Processors
  - Tool Providers

- Did I miss some roles?



Presenter 4/24/2013, 9:11:43 AM

OK. Just some basic level setting about PDF. In June 2013 PDF will be 20 years old. Adobe introduced PDF in 1993 as the file format supported by its Acrobat product line. Based on some experience we had with PostScript being both a file format and a product, we decided with PDF to make a clear distinction between the file format (PDF) and the Adobe products that support it (Acrobat).
We trademarked Acrobat but did not trademark PDF. We published the full specification of PDF 1.0 in June 1993 in paperback form and wanted other to develop software to read, process and write PDF files.

Through the years as function was added to Adobe's Acrobat and its use of PDF we revised the PDF specification and published it each time. So the 8th version (PDF 1.7) was published by Adobe in 2006. In 2007 Adobe inked an agreement with AIIM and ISO to hand over control of PDF to ISO. We had always gotten complaints that had total control over what new things went into each new version. With ISO owning the specification, then the world would have control over the evolution not just Adobe.

This picture is my attempt to depict that a file made to conform to PDF 1.0 or 1.4 also conforms to PDF 1.7. We did not want to ever obsolete any existing PDF files.

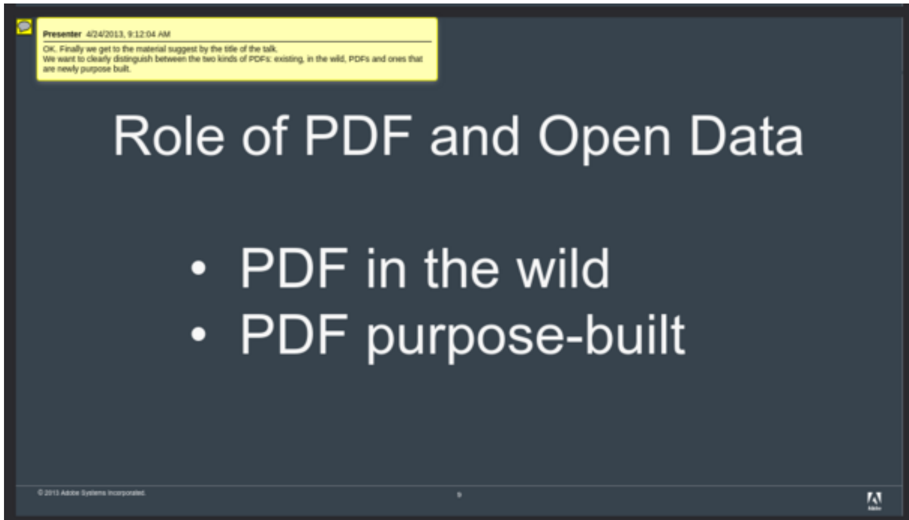
ISO published its first PDF standard ISO 32000-1 in July 2008. (Yes, the public has "owned" the standard for almost 5 years. ISO is now working on PDF 2.0 which should come out in a year or so.

DF

/ Adobe in June 1993

SO Standard in July 2008

ISO Work on PDF is ongoing

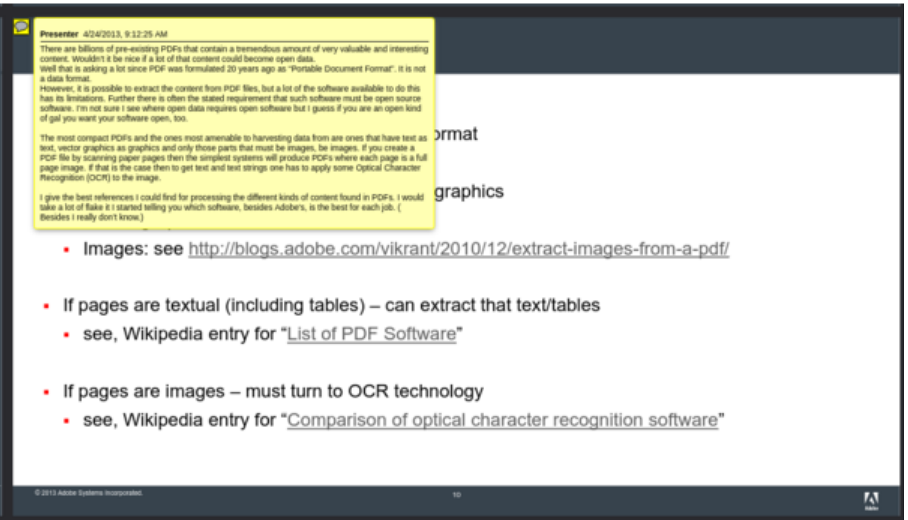


Presenter 4/24/2013, 9:12:04 AM

OK. Finally we get to the material suggest by the title of the talk.
We want to clearly distinguish between the two kinds of PDFs: existing, in the wild, PDFs and ones that are newly purpose built.

Role of PDF and Open Data

- PDF in the wild
- PDF purpose-built



Presenter 4/24/2013, 9:12:25 AM

There are billions of pre-existing PDFs that contain a tremendous amount of very valuable and interesting content. Wouldn't it be nice if a lot of that content could become open data.
Well that is asking a lot since PDF was formulated 20 years ago as "Portable Document Format". It is not a data format.
However, it is possible to extract the content from PDF files, but a lot of the software available to do this has its limitations. Further there is often the stated requirement that such software must be open source software. I'm not sure I see where open data requires open software but I guess if you are an open kind of gal you want your software open, too.

The most compact PDFs and the ones most amenable to harvesting data from are ones that have text as text, vector graphics as graphics and only those parts that must be images, be images. If you create a PDF file by scanning paper pages then the simplest systems will produce PDFs where each page is a full page image. If that is the case then to get text and text strings one has to apply some Optical Character Recognition (OCR) to the image.

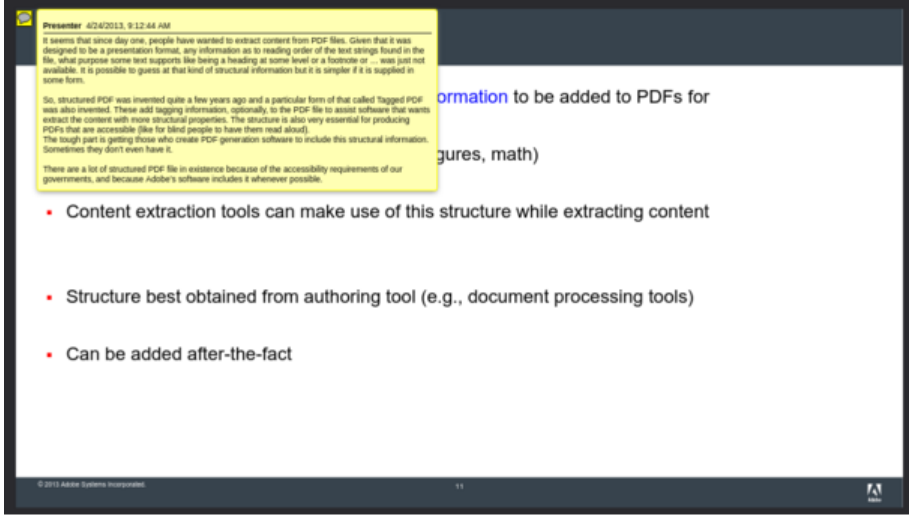
I give the best references I could find for processing the different kinds of content found in PDFs. I would take a lot of flake it I started telling you which software, besides Adobe's, is the best for each job. (Besides I really don't know.)

rmat

graphics

- Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/

- If pages are textual (including tables) – can extract that text/tables
  - see, Wikipedia entry for "List of PDF Software"
- If pages are images – must turn to OCR technology
  - see, Wikipedia entry for "Comparison of optical character recognition software"



Presenter 4/24/2013, 9:12:44 AM

It seems that since day one, people have wanted to extract content from PDF files. Given that it was designed to be a presentation format, any information as to reading order of the text strings found in the file, what purpose some text supports like being a heading at some level or a footnote or ... was just not available. It is possible to guess at that kind of structural information but it is simpler if it is supplied in some form.

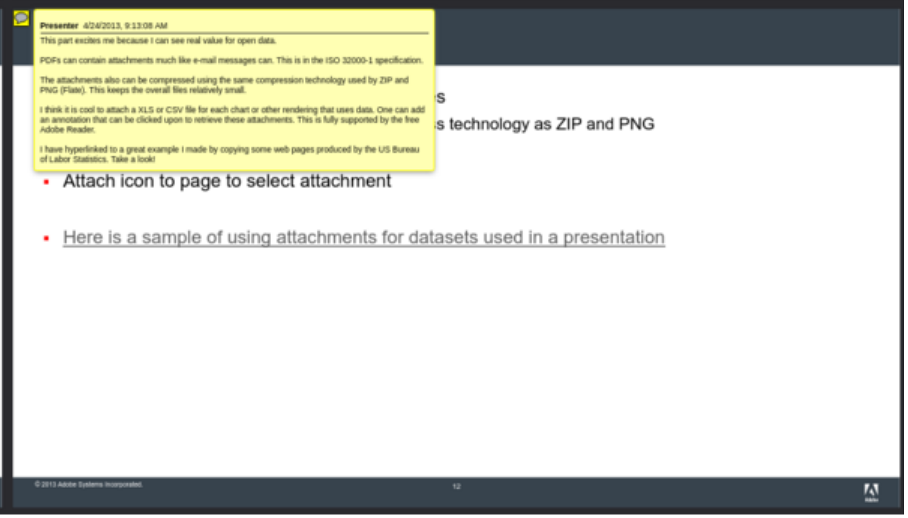
So, structured PDF was invented quite a few years ago and a particular form of that called Tagged PDF was also invented. These add tagging information, optionally, to the PDF file to assist software that wants extract the content with more structural properties. The structure is also very essential for producing PDFs that are accessible (like for blind people to have them read aloud).
The tough part is getting those who create PDF generation software to include this structural information. Sometimes they don't even have it.

There are a lot of structured PDF file in existence because of the accessibility requirements of our governments, and because Adobe's software includes it whenever possible.

ormation to be added to PDFs for

gures, math)

- Content extraction tools can make use of this structure while extracting content
- Structure best obtained from authoring tool (e.g., document processing tools)
- Can be added after-the-fact



Presenter 4/24/2013, 9:13:08 AM

This part excites me because I can see real value for open data.

PDFs can contain attachments much like e-mail messages can. This is in the ISO 32000-1 specification.

The attachments also can be compressed using the same compression technology used by ZIP and PNG (Flate). This keeps the overall files relatively small.

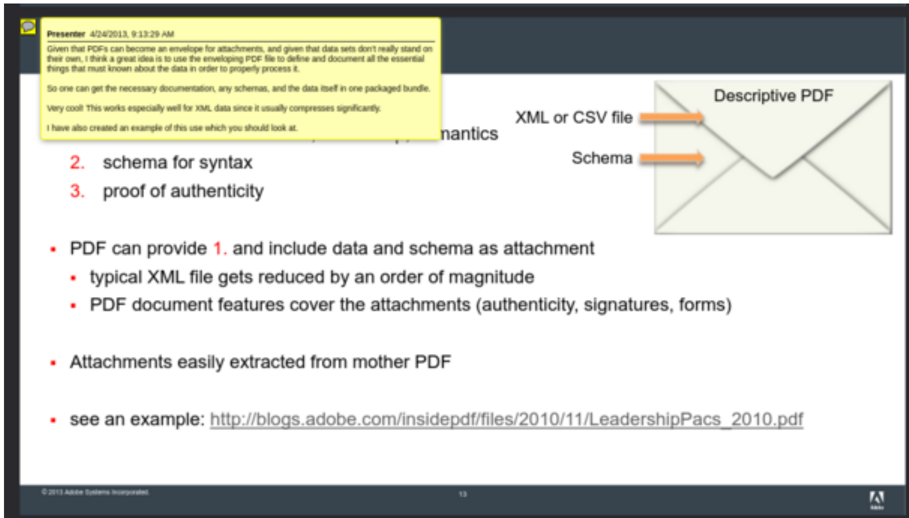
I think it is cool to attach a XLS or CSV file for each chart or other rendering that uses data. One can add an annotation that can be clicked upon to retrieve these attachments. This is fully supported by the free Adobe Reader.

I have hyperlinked to a great example I made by copying some web pages produced by the US Bureau of Labor Statistics. Take a look!

s

s technology as ZIP and PNG

- Attach icon to page to select attachment
- Here is a sample of using attachments for datasets used in a presentation



Presenter 4/24/2013, 9:13:29 AM

Given that PDFs can become an envelope for attachments, and given that data sets don't really stand on their own, I think a great idea is to use the enveloping PDF file to define and document all the essential things that must known about the data in order to properly process it.

So one can get the necessary documentation, any schemas, and the data itself in one packaged bundle.

Very cool! This works especially well for XML data since it usually compresses significantly.

I have also created an example of this use which you should look at.

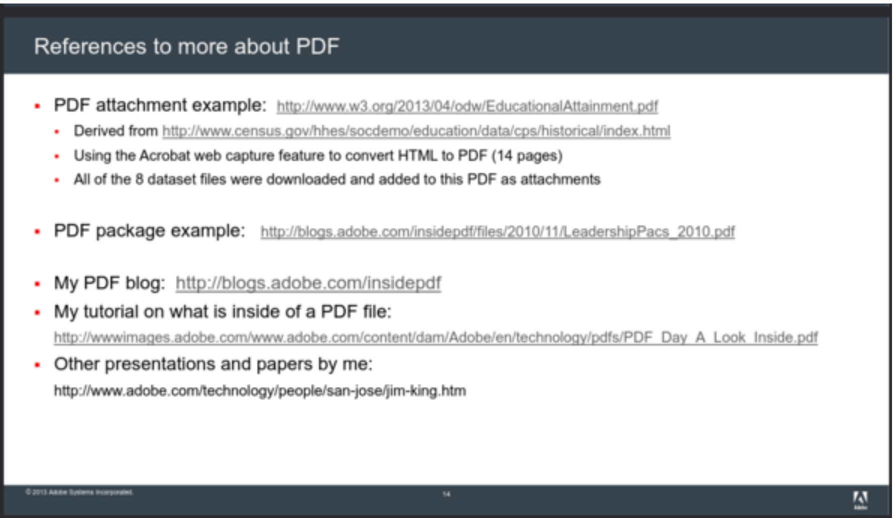
nantics

2. schema for syntax
3. proof of authenticity

- PDF can provide 1. and include data and schema as attachment
  - typical XML file gets reduced by an order of magnitude
  - PDF document features cover the attachments (authenticity, signatures, forms)

- Attachments easily extracted from mother PDF

- see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

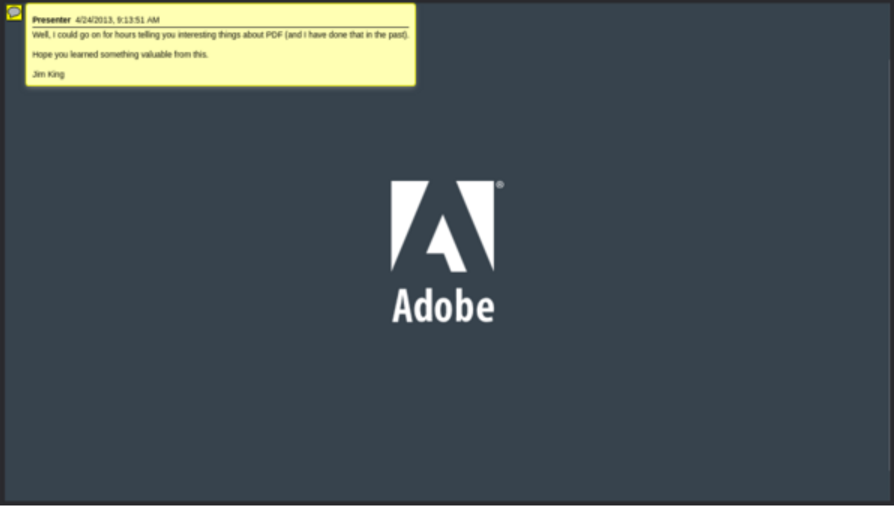


References to more about PDF

- PDF attachment example: http://www.w3.org/2013/04/odw/EducationalAttainment.pdf
  - Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html
  - Using the Acrobat web capture feature to convert HTML to PDF (14 pages)
  - All of the 8 dataset files were downloaded and added to this PDF as attachments

- PDF package example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

- My PDF blog: http://blogs.adobe.com/insidepdf
- My tutorial on what is inside of a PDF file: http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.pdf
- Other presentations and papers by me: http://www.adobe.com/technology/people/san-jose/jim-king.htm



Presenter 4/24/2013, 9:13:51 AM

Well, I could go on for hours telling you interesting things about PDF (and I have done that in the past).

Hope you learned something valuable from this.

Jim King