Datalogics

Content Extraction From PDF Files

Matt Kuznicki
Chief Technical Officer
Datalogics

PDF Days – December 2014
PDF Association

Outline

- Image Files as Document Representations
- Advantages of PDF for Content Preservation
- Extracting Content: Examples
- Common Pitfalls
- Summary

In the Beginning… We Had Pictures

Image file formats have a long history in archiving and document management workflows:

- Images are the native format of scanners and digital cameras
- Image files are easy to produce and consume – little interpretation is needed in most cases
- Images are easy for humans to interpret visually
- Platform and environment compatibility is reasonable

Drawbacks of Images for Content Storage

Image Formats Have Several Key Drawbacks:

- The collection of dots (pixels) discards information in the original document

  Compression can exacerbate this dramatically

- File size can increase dramatically from the original
- Document content can no longer be edited
- Images capture only what a document *looks like*, not what a document *contains* or *means*
- ***Users – human and otherwise – who cannot see images cannot derive any value from them***

Images – They're Just Dots

Drawbacks of Images

Images – They're Just Dots

Drawbacks of Images

PDF – the Portable Document Format

PDF was created by Adobe Systems as a format to store content in an extractable manner, and to store information about that content.

PDF is an open, international standard (ISO 32000) and is freely implementable. Many tools, commercial and open source, are available.

Different flavors of PDF (PDF/A, PDF/UA) represent restrictions and best practices that enable more reliable and accurate content extraction.

PDF – the Portable Document Format

What can a PDF contain?

- Textual content: words, sentences, numeric values, tables
- Line art: curves, lines, shapes
- Pictures: raster images
- Multimedia and interactivity
- Markup and comments
- Display and navigation instructions: layers, bookmarks
- Attachments and embedded files

PDF – the Portable Document Format

PDFs can contain not just content, but information about the content within:

- Semantic meaning and alternate information (tagged PDF)
- Notes and annotations
- Metadata & properties
- Measurements and geospatial information
- Attachments, including source documents

PDFs are able but not guaranteed to store this information

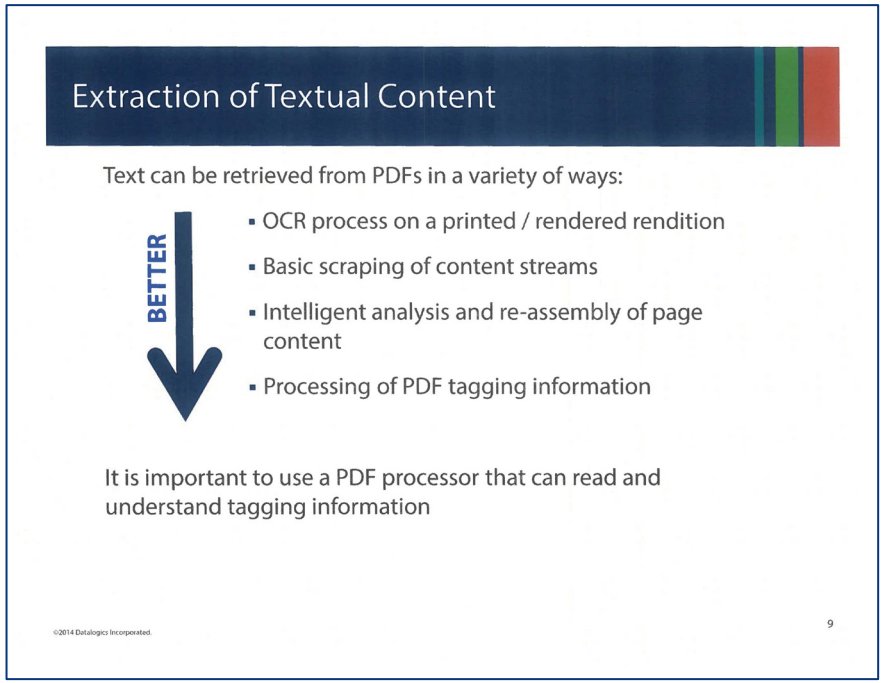
Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

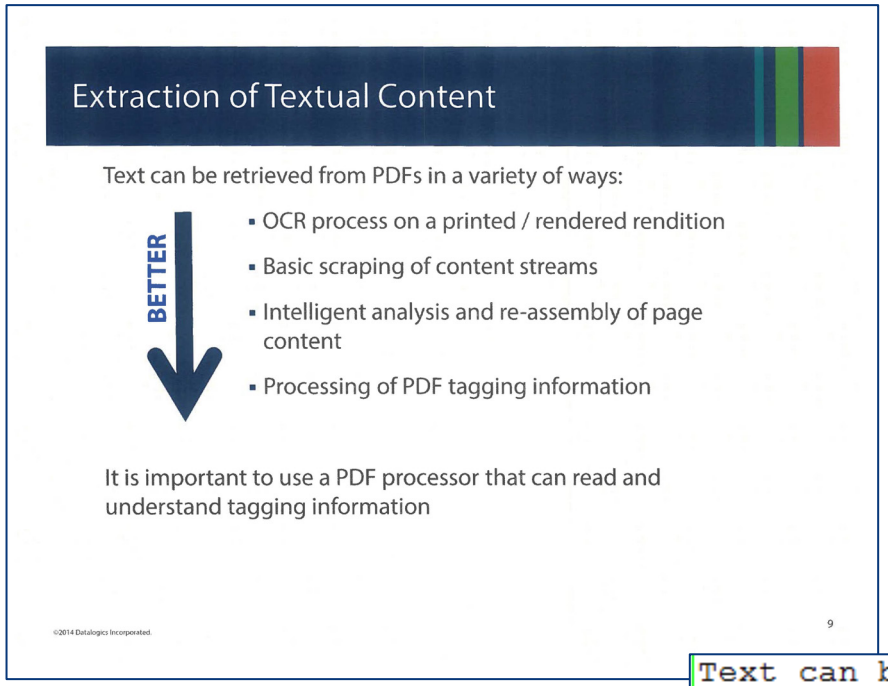
- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information

Example: Text Extraction via OCR



Example: Text Extraction via OCR

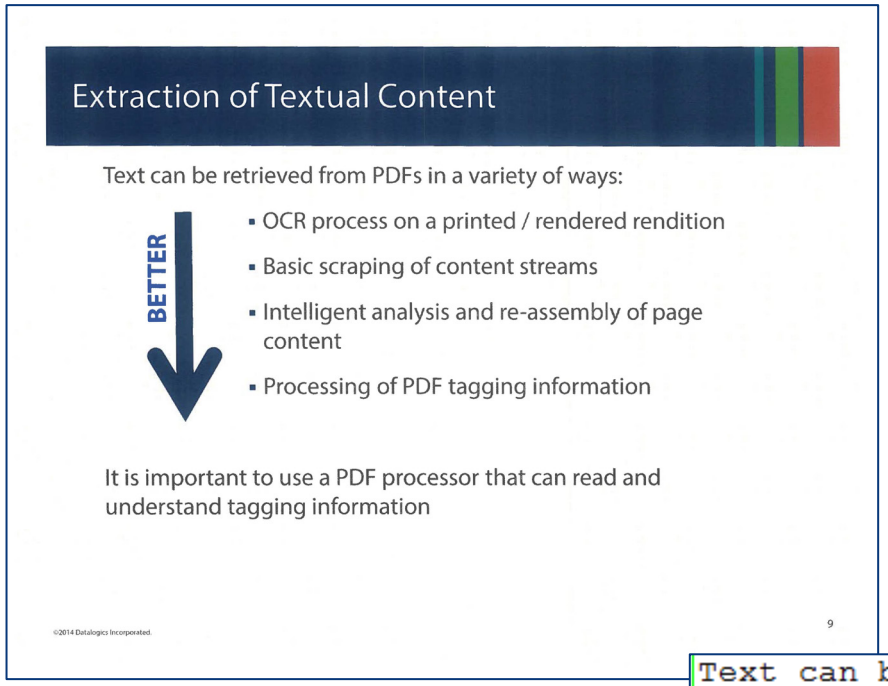



```
Text can be retrieved from PDFs in a variety of ways:

• OCR process on a printed I rendered rendition
• Basic scraping of content streams
• Intelligent analysis and re-assembly of page
content
• Processing of PDF tagging information
It is important to use a PDF processor that can read and
understand tagging information

<;;>2014 Datalog1cs Incorporated.
```

Example: Text Extraction via OCR



This is a best-case example. More complicated pages lead to more errors.

```
Text can be retrieved from PDFs in a variety of ways:

• OCR process on a printed I rendered rendition
• Basic scraping of content streams
• Intelligent analysis and re-assembly of page
content
• Processing of PDF tagging information
It is important to use a PDF processor that can read and
understand tagging information

<;;>2014 Datalog1cs Incorporated.
```

Example: Text Extraction via Content Stream

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information

8

Example: Text Extraction via Content Stream

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information


```
It is important to use a PDF processor that can read and
understand tagging information

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

BETTER

§OCR §Basic §Intelligent §Processing
process scraping of
on a of analysis PDF
printed / content and re- tagging
rendered streams assembly information
rendition of page
```

Example: Text Extraction via Content Stream

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information

The order of text in the content stream may be quite different from how the human eye interprets the page

```
It is important to use a PDF processor that can read and
understand tagging information

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

BETTER

§OCR §Basic §Intelligent §Processing
process scraping of
on a of analysis PDF
printed / content and re- tagging
rendered streams assembly information
rendition of page
```

Example: Text Extraction, Tagged PDF

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information

8

Example: Text Extraction, Tagged PDF

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

BETTER

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information


```
Extraction of Textual Content


Text can be retrieved from PDFs in a variety of ways:

A list of text content extraction methods, in increasing capability:
§OCR process on a printed / rendered rendition
§Basic scraping of content streams
§Intelligent analysis and re-assembly of pagecontent
§Processing of PDF tagging information




It is important to use a PDF processor that can read and
understand tagging information
```

Example: Text Extraction, Tagged PDF

Extraction of Textual Content

Text can be retrieved from PDFs in a variety of ways:

BETTER

- OCR process on a printed / rendered rendition
- Basic scraping of content streams
- Intelligent analysis and re-assembly of page content
- Processing of PDF tagging information

It is important to use a PDF processor that can read and understand tagging information

Text is Extracted in the Reading Order Specified by the PDF Creator.

Extracted Alternate Text from the Arrow Graphic

```
Extraction of Textual Content


Text can be retrieved from PDFs in a variety of wa

A list of text content extraction methods, in increasing capability:
§OCR process on a printed / rendered rendition
§Basic scraping of content streams
§Intelligent analysis and re-assembly of pagecontent
§Processing of PDF tagging information



It is important to use a PDF processor that can read and
understand tagging information
```

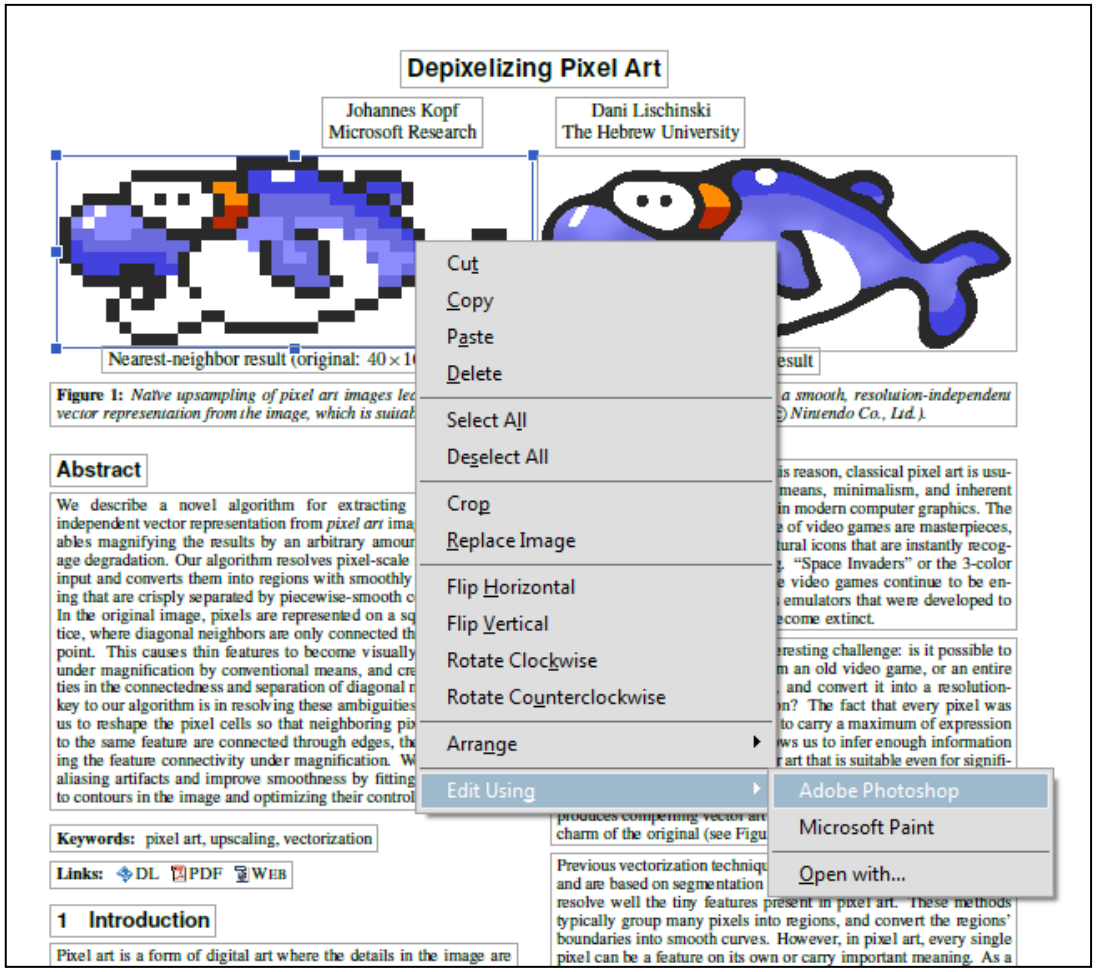
Extraction of Images

Images can exist in a variety of data formats in PDFs:

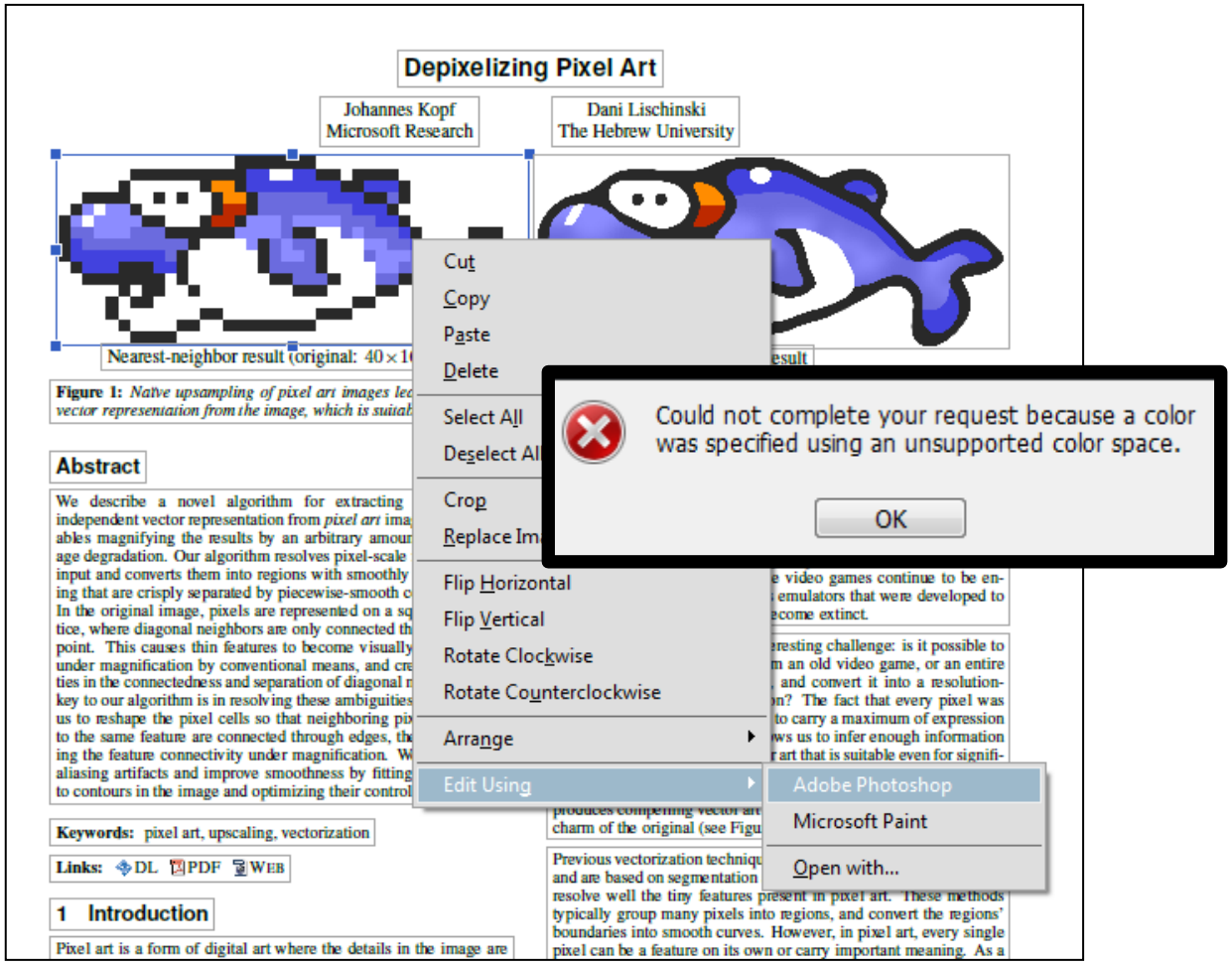
- Some images may be directly exported
- Different images may be in different color spaces, even on the same page
- Most images in a PDF require some transformation in order to be repurposed
- Some images in a PDF may need color conversion, resampling or other changes for downstream processes
- Images can contain metadata and alternate text, but this is not required

It's imperative to use a PDF processor that can understand and properly process all types of images in a PDF file

Extraction of Images – Problem Example



Extraction of Images – Problem Example



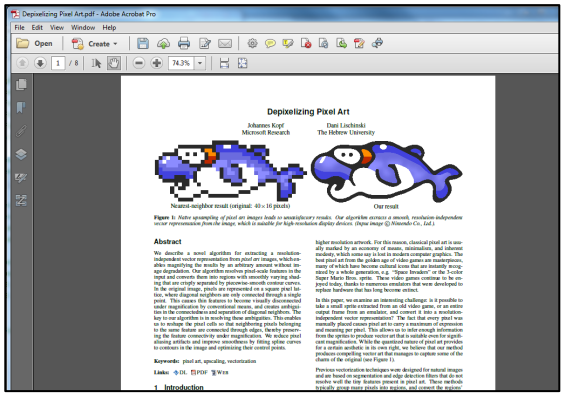
Extraction of Metadata and Properties

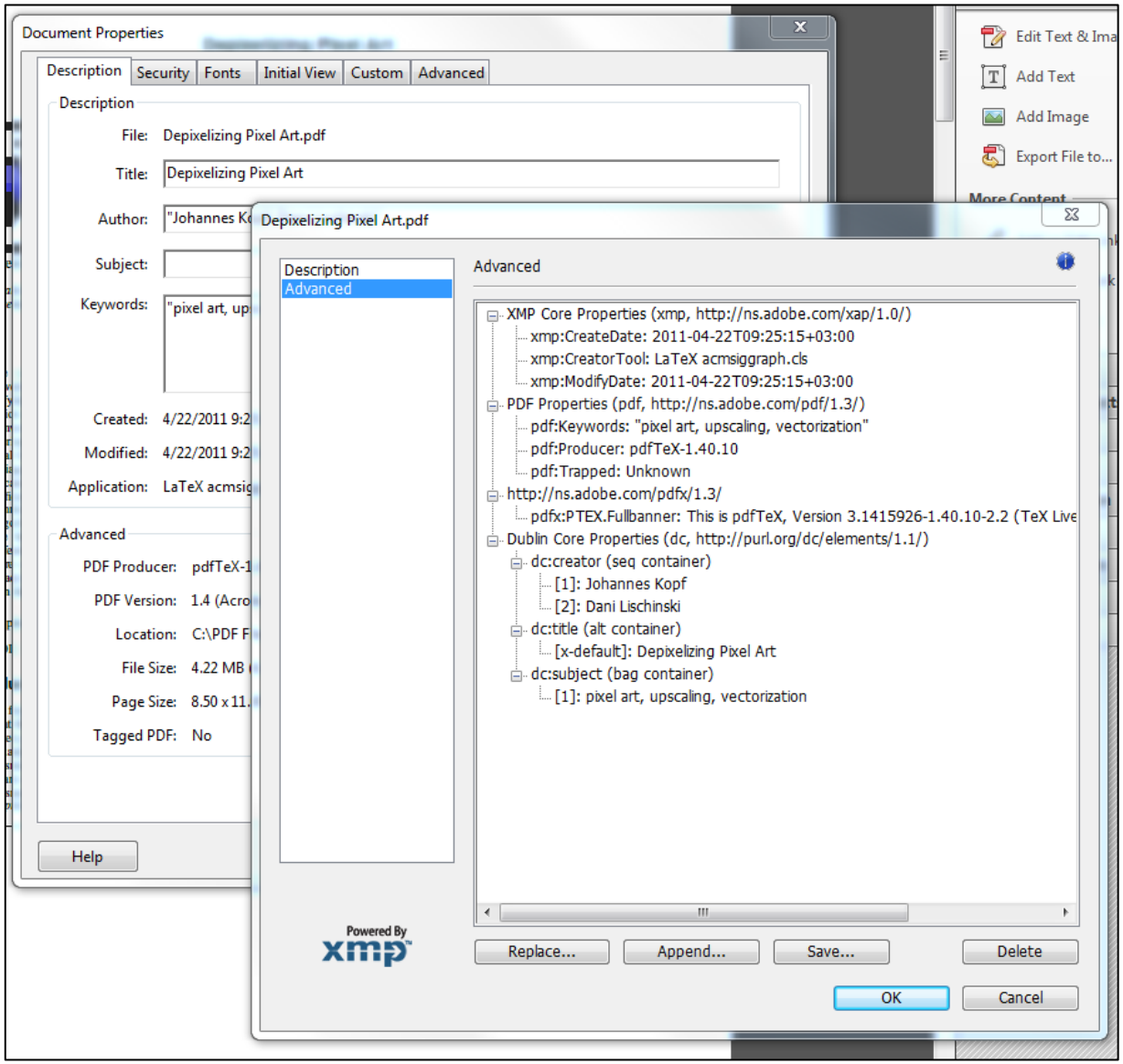
Metadata and properties can exist for a PDF and for the elements stored within:

- Document level information: author, title, ISBN, etc.
- Metadata about images: copyright information, camera information, etc.
- Alternate text descriptions for images and figures
- Creator or workflow specific metadata for any piece of extractable content

Information about how content is structured is contained within Tagged PDF

Example: Document Metadata





Common Pitfalls

PDF files are able to contain information that can reliably be extracted – but they do not have to be "nice":

- Text may be represented in a semantically different way in the PDF from its visual appearance

  In extreme cases, text may accidentally or intentionally be represented in a way that cannot be reliably extracted

- What looks to be one image may be several (or thousands!) placed close together

- Invisible content may be contained within a file

PDF files that are created well are much easier for processors to use for content extraction.

Summary

PDF is a format for documents and information – not just visual images

- PDF files can be edited and used in workflows, they are not just for end-stage archiving or storage
- PDF files contain reusable, extractable information – they are not just pictures
- PDF files are not guaranteed to be well-behaved or perfectly constructed for content extraction
- PDF standards like PDF/A and PDF/UA help make content extraction more predictable and reliable
- Tools that understand PDF tagging and structure make for better results in content extraction

Thank you for your time!