



Quality Assurance in PDF

Roger Reeves

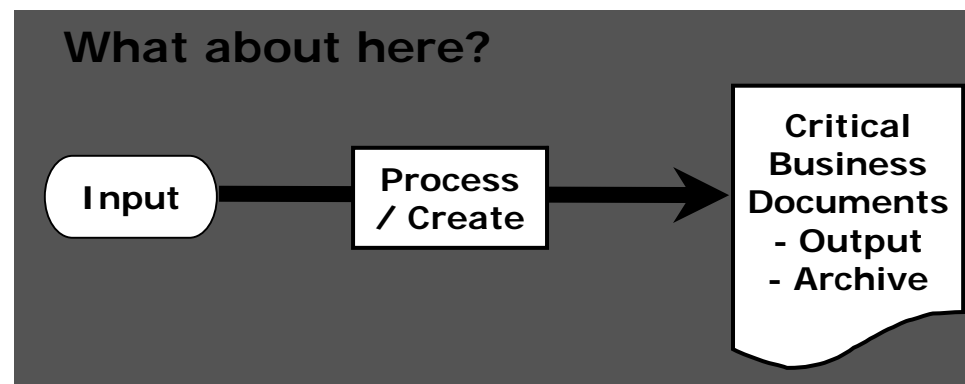
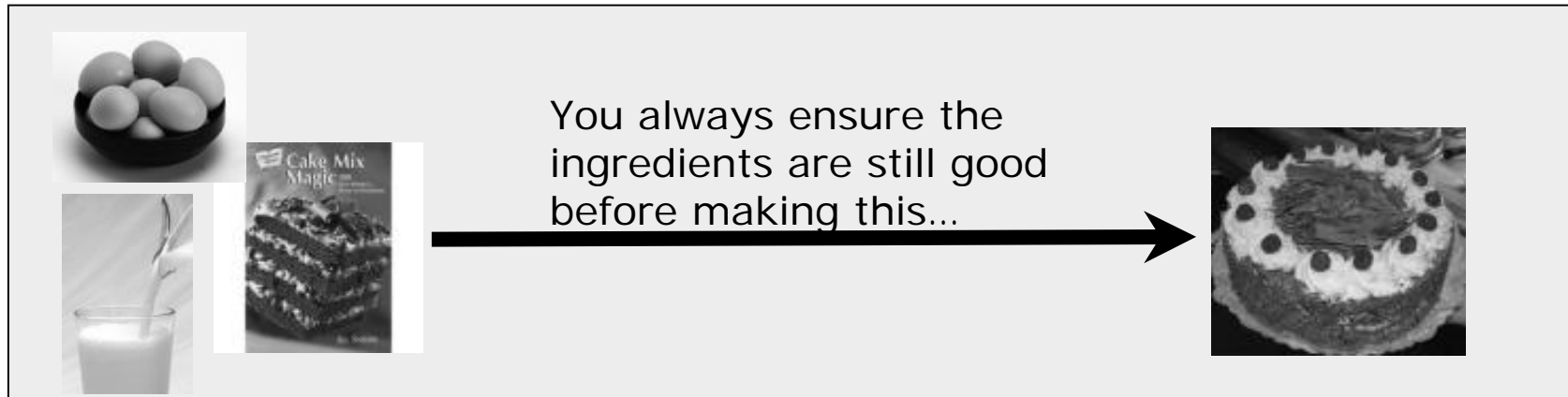
Dr. Hans-Rudolf Aschmann

PDF Tools AG, Switzerland

Topics

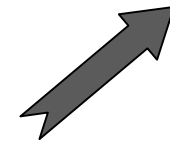
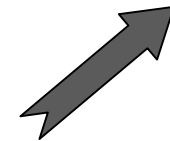
- Why Quality Assurance?
- The Adobe® PDF Reference and PDF/X
- What is corruption
- Common forms of corruption
- What happens if a PDF file is corrupt
- When should quality be analysed
- How to analyze PDF quality
- Summary

Why Quality Assurance?



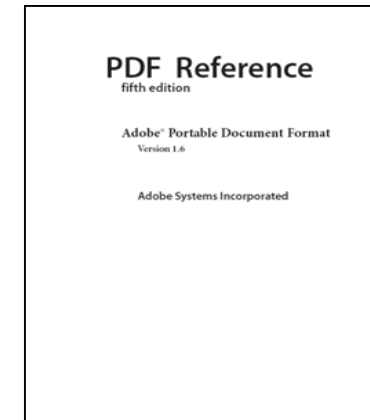
Quality Assurance is necessary

- Mature PDF “Creators” and “Processors” produce a much better quality than in the beginning
- Most PDFs created with newer versions of brand-name products (e.g. Adobe, etc.) should be error-free
- PDF products from questionable sources, older product versions, and customizing from less-experienced programmers are susceptible to error:
 - Improper validation techniques are used (validated against a specific Acrobat version, and not the PDF Reference)
 - PDF Reference is evolving
 - PDF Reference is not always fully understood



The Adobe® PDF Reference

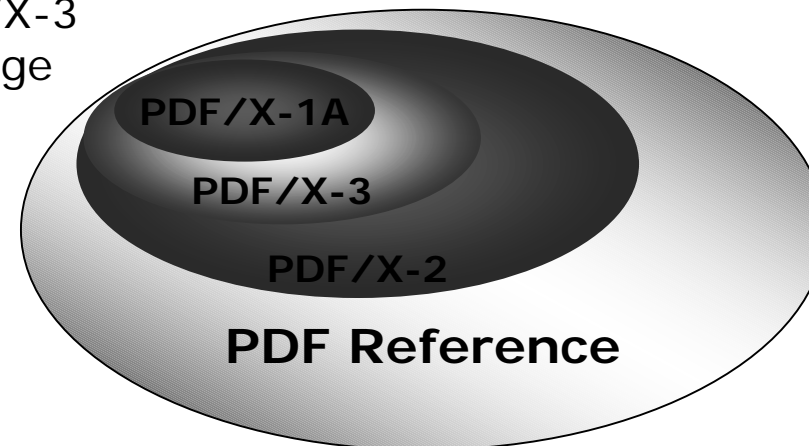
- The PDF “Bible”
- Defines requirements that a file must meet to be PDF
- Specifies how different functionality and objects (e.g. fonts, images, pages, etc.) are used to render (view, print) a PDF



PDF and PDF/X

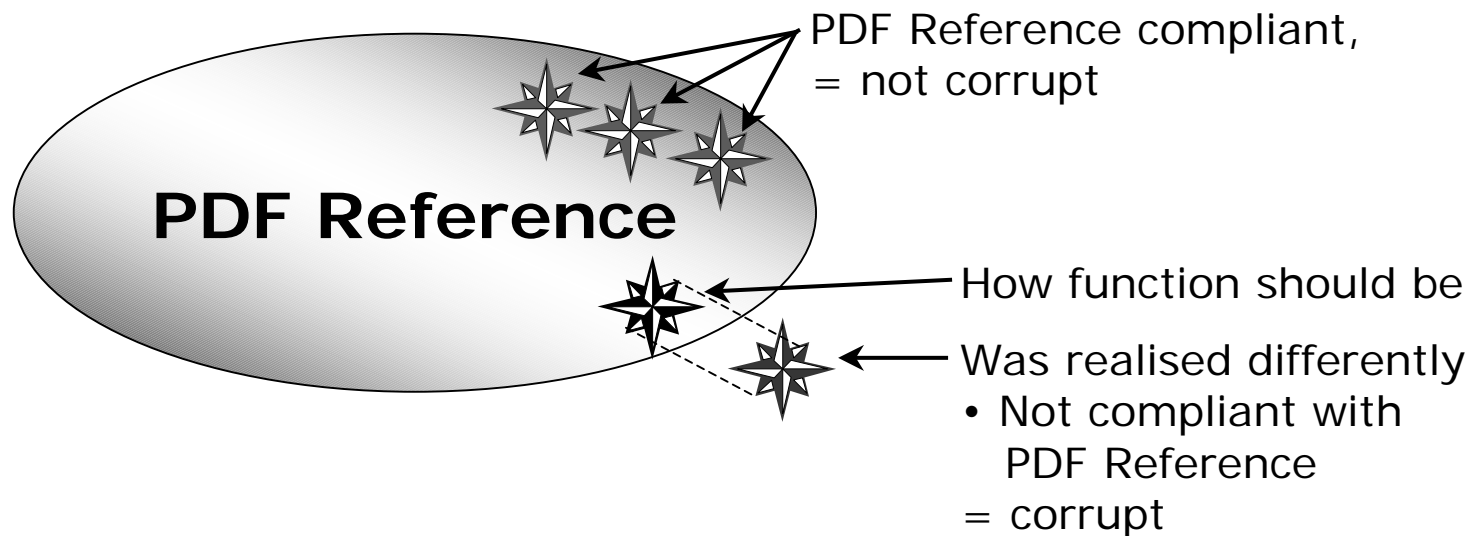
PDF/X is a subset of the PDF Reference, designed primarily for prepress data interchange (design goal: 100 % printer independent reproducibility)

- PDF/X-1A: CYMK, blind exchange (no additional technical info)
- PDF/X-3: superset of PDF/X-1A
CYMK, CIELab, RGB, blind exchange
- PDF/X-2: superset of PDF/X-3
not blind exchange



What is Corruption?

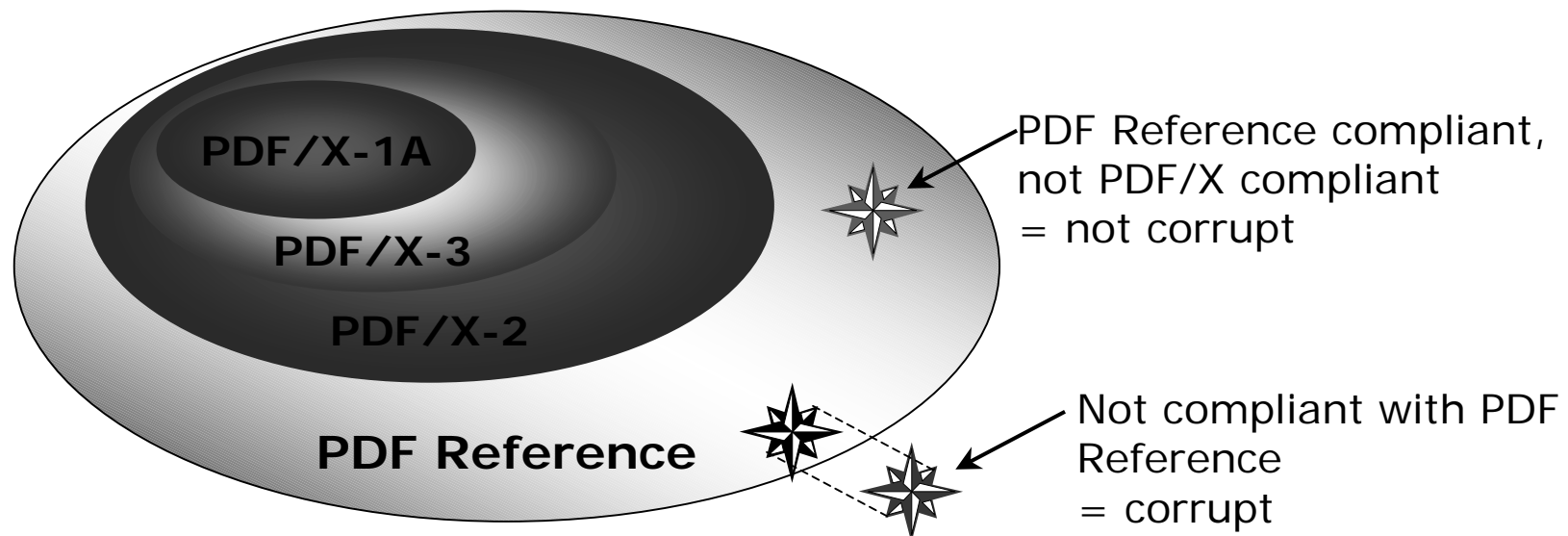
- If any element of a PDF file does not meet the requirements of the PDF Reference, the PDF file is corrupt.
- If embedded, non-PDF data elements (e.g. fonts, compressed images) are defective, the PDF file is corrupt.



What is corruption?

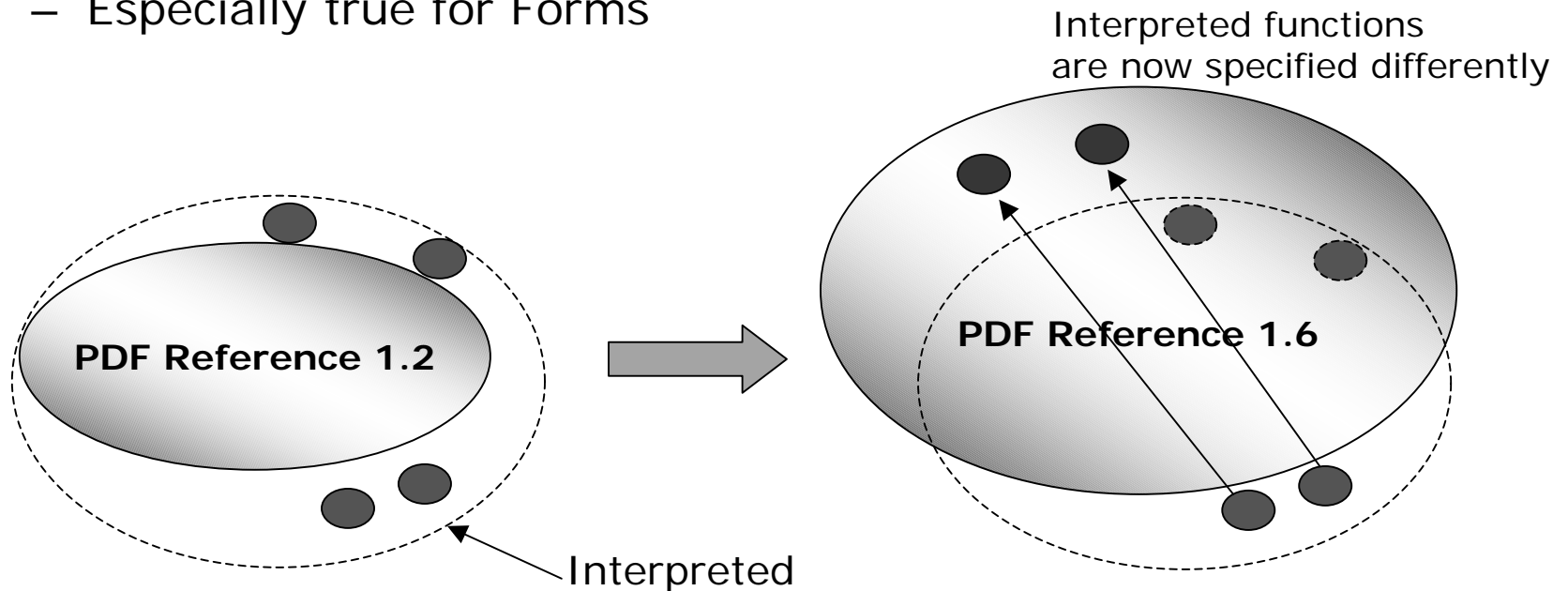
PDF/X and corruption

- A file that doesn't meet the PDF/X requirements is not corrupt, if the Adobe PDF Reference is fulfilled.
- Compliance with PDF/X and compliance with the PDF Reference are different issues



Common forms of corruption

- PDF Reference is evolving
 - Reference has(had) holes that are being filled
 - Unfilled holes are interpreted, sometimes incorrectly
 - Especially true for Forms





PDF not recognized as Binary 1/2

Sample PDF Code (PDF viewed with Notepad)

```
%PDF-1.2
```

```
%âãïó
```



Non-printing characters
PDF Reference recommends

```
913 0 obj<</Length
```

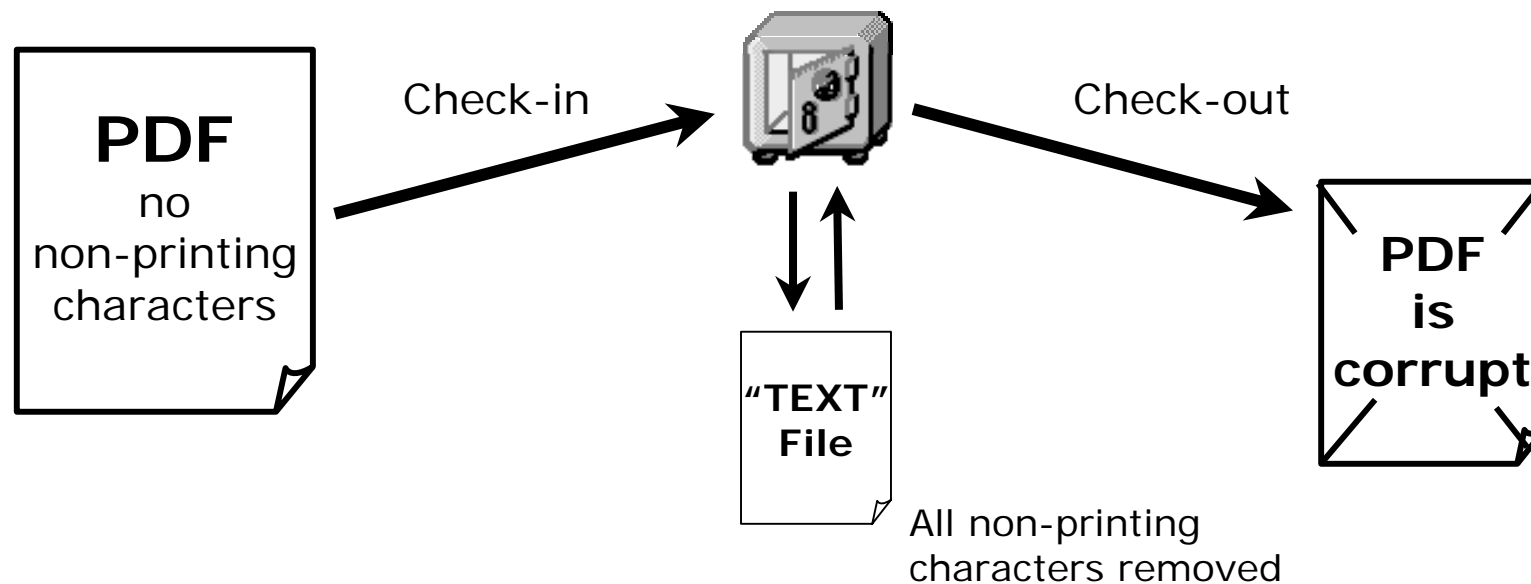
```
2756/Filter/FlateDecode/Type/ObjStm/First 1774/N
```

```
200>>stream
```

```
9®EœË°hú Ìü ç Yr0šÅÅïXÐ (
```

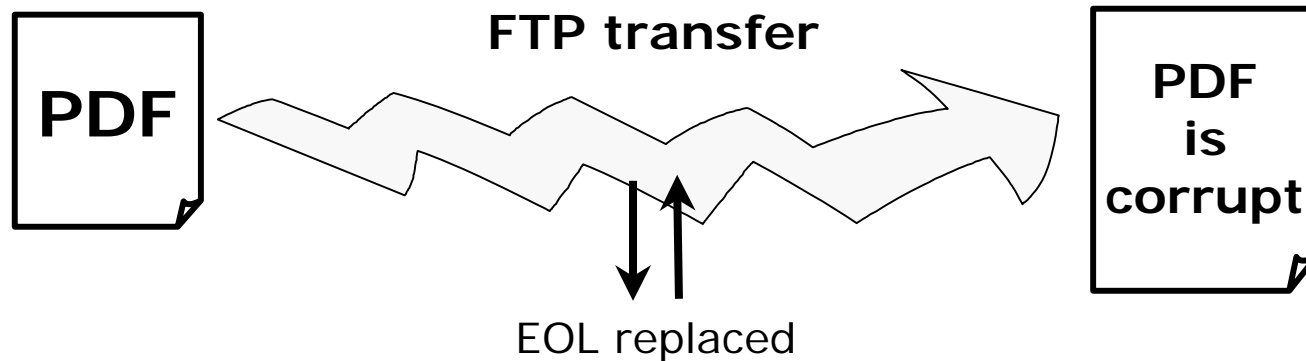
PDF not recognized as Binary 2/2

- If non-printing characters are not present in line 2, Visual SourceSafe may interpret the file as a text file when checking it in (automatic file recognition). **All** non-printing characters will then be deleted.
- Once deleted, they cannot be recovered and the file is corrupt.



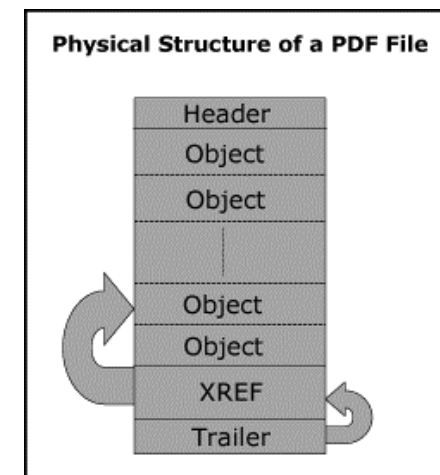
Incorrect use of FTP

- Many FTP (File Transfer Protocol) applications don't transmit binary files unless they are specifically configured for it. Default = Text.
- End-of-Line (EOL) is replaced through Carriage Return (CR) and/or vice-versa
- This automatically (and systematically) corrupts the PDF file



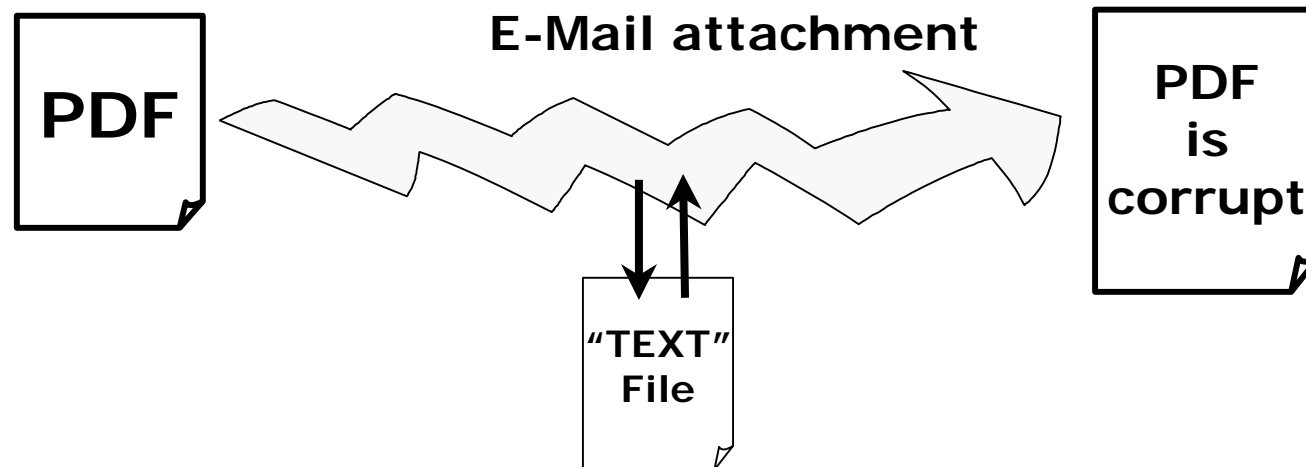
Truncated File Transfers

- PDF physical and logical file structures are different
- Physical structure shown in diagram
- Logically the header is read, then the trailer. The trailer points to the XREF table, which in turns point to the objects.
- If the file was truncated during transmission, the trailer and possibly the XREF table are incomplete and the PDF file is corrupt





Corruption from Mail Applications

- Older mail programs did not recognize PDF format and treated the files as text (non-printing characters removed)
- You probably don't have any current PDFs that have been corrupted by mail programs
- You may have older PDFs in your archive which were

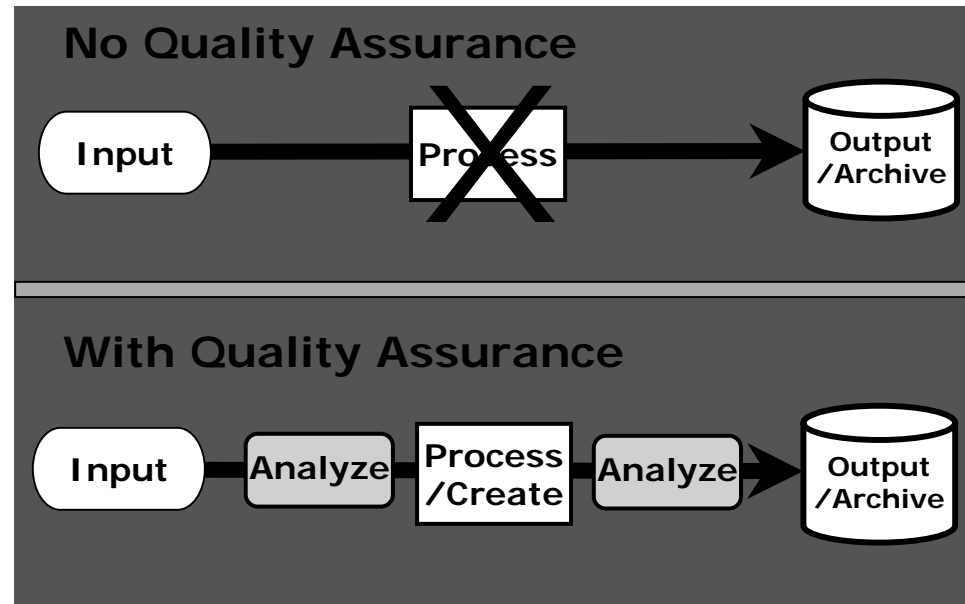


What happens if a PDF file is corrupt?

- If you're lucky, the file can be fully repaired 
- Viewer may correct the file "on-the-fly" so that it can be displayed:
 - File is not necessarily repaired, corruption remains
 - Processing the file later may worsen the corruption
 - If newer viewer versions adhere closer to the PDF Reference, the file may not be viewable in the future
- Viewer may display the PDF file, without recognizing that it is corrupt:
 - Only certain portions are not displayed (e.g. bottom of page 105)
 - File does not get repaired or is repaired incorrectly
- Worst case: it won't open and is irreparable 

When should quality be analyzed?

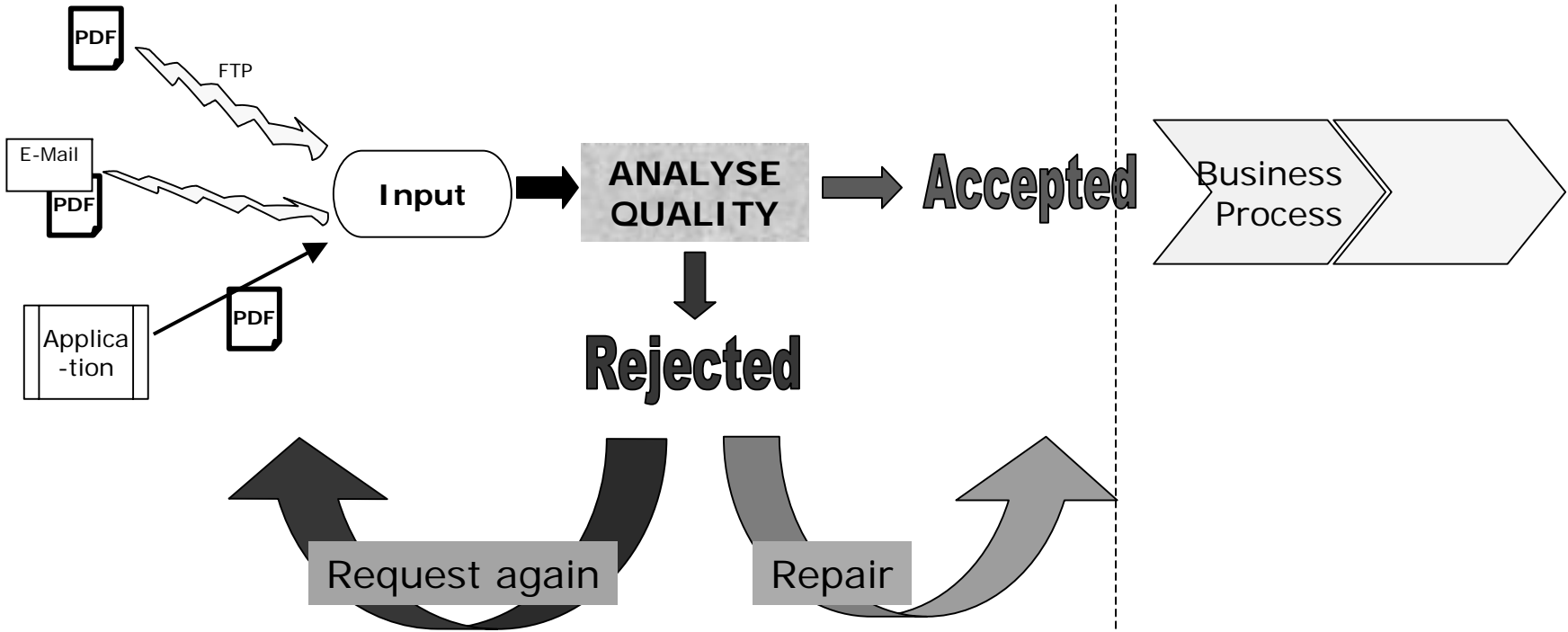
- Before introducing PDF files into your business process
- After creating or processing PDF files, before they are archived or distributed as output files



When should Quality be Analysed

Analyse Quality: before Processing

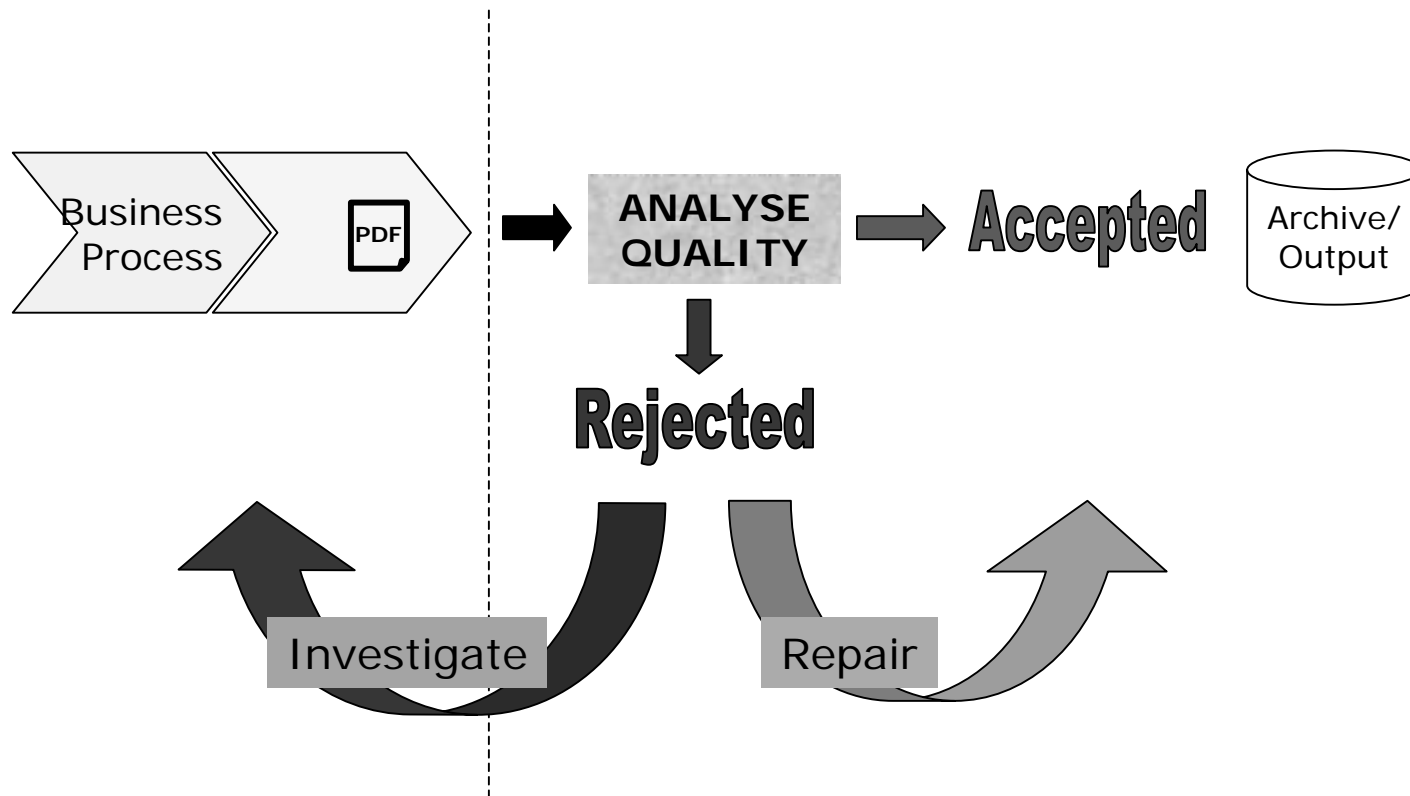
- Confirm quality of PDF at start of business process



When should Quality be Analysed

Analyse Quality: after Processing

- Confirm quality of PDF at end of business process



Analyse Quality: How often?

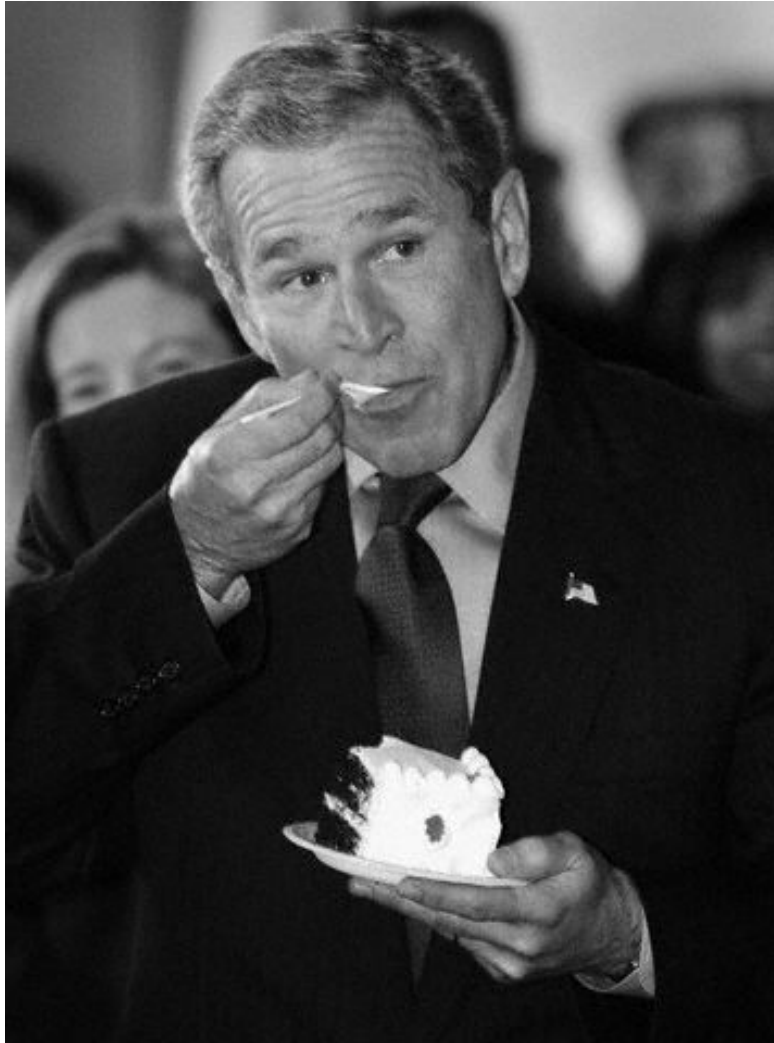
- **Input files** from stable internal data source:
 - Check once using sampling, and after each change to data source
- **Input files** from external or uncertain data source:
 - Check each file or run batch check on a regular basis
- **Output files** with repetitive processing using the same data type, same application:
 - Check once using sampling, and after each change to application / product or data source
- **Output files** with complicated processing or non-certified application:
 - Check each file or run batch check on a regular basis

Analyse Quality: How?

- **Open with current version of Adobe® Acrobat**
 - Good but not great – Acrobat only complains when encountering certain types of corruptions during rendering
 - Acrobat is currently optimized to work around corruption, and not to analyze for it
- **Analyze using a PDF analysis and repair tool**
 - “3-Heights™ PDF Repair Tool” from PDF Tools AG
 - “Advanced PDF Repair (APDFR)” from PDF Repair, Inc.
- **Analyze using a PDF subset compliance check tool**
 - “Colour Chameleon” (cleans incoming Postscript before it hits Acrobat Distiller)

Summary

- Corruption is still prevalent in PDF
- Viewable today does not mean viewable tomorrow
- Use professional PDF creation and processing tools
- Analyse input, output and archived PDF files
- Design your business processes accordingly



Ensure that your
PDF documents
are as good
as your cake was

Questions?

