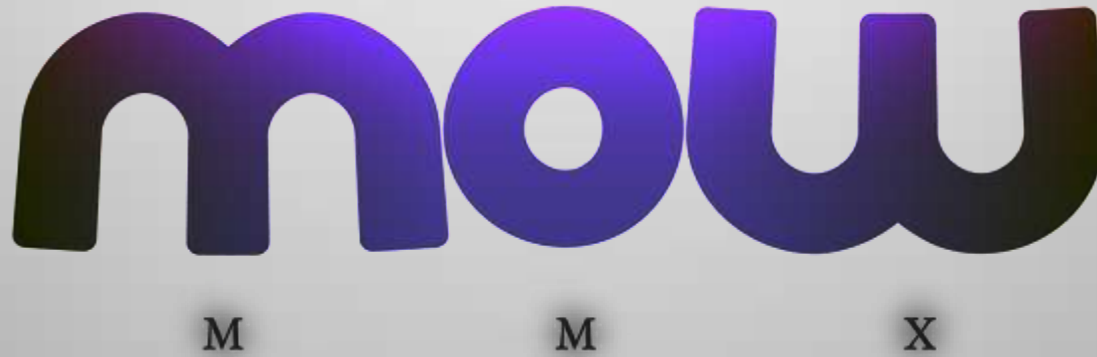


Dissecting PDF Documents



Miracle OpenWorld 2010

Mark S. Rasmussen – iPaper
mark@improve.dk

What Is This Session NOT About?

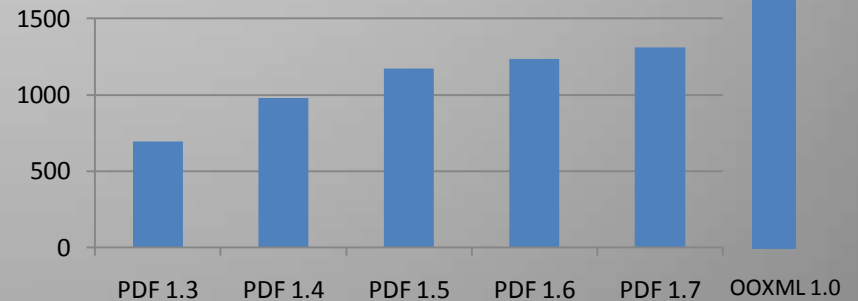
- Creating PDFs
- How to use Acrobat
- Transparency flattening options in InDesign

- So what is it about?
 - PDF documents
 - Tooling
 - Extracting data



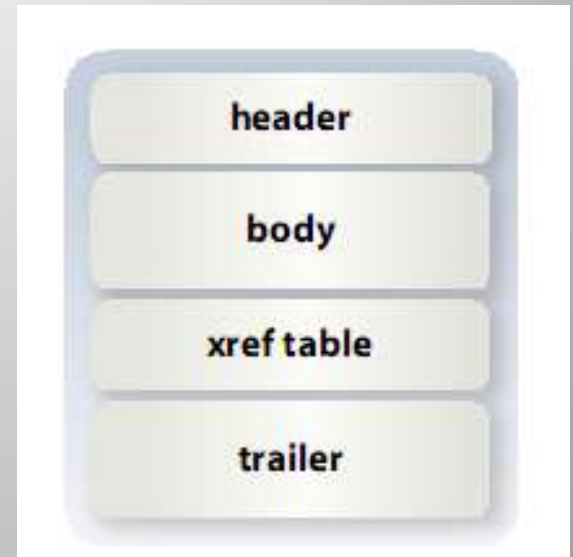
The PDF Format

- 1.0 released in 1993
- Open standard as of July 1st 2008
- Reference publicly available
 - http://www.adobe.com/devnet/pdf/pdf_reference_archive.html



PDF Structure

- Header
 - %PDF-1.4
 - % \AA (optional but common)
- Body
 - Objects
- Xref table
 - Index table containing pointers to objects
- Trailer
 - Pointers to Xref table, key objects
 - %%EOF



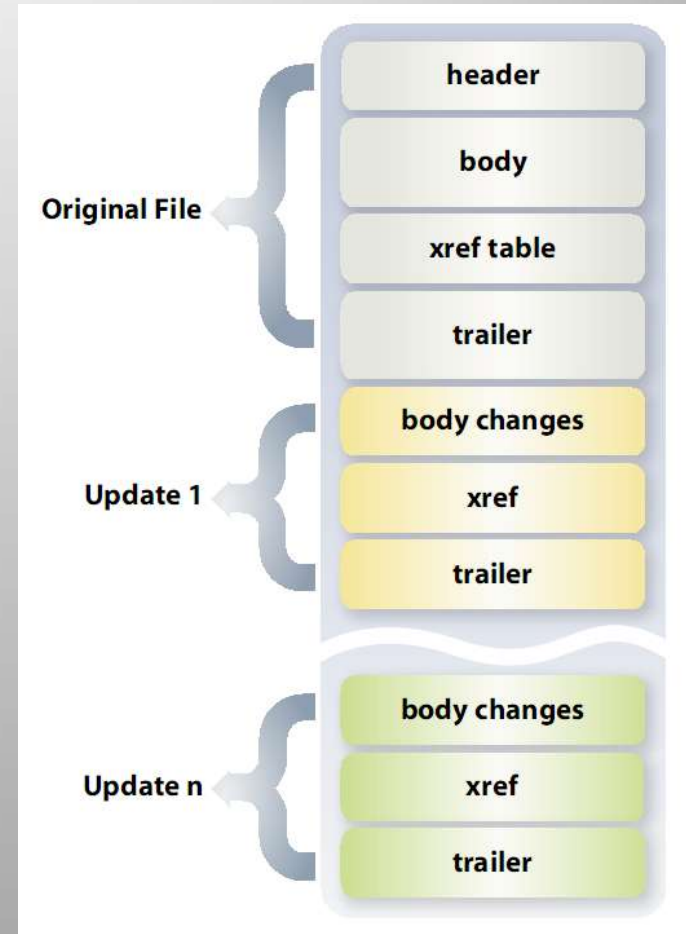
PDF Objects

"A PDF file should be thought of as a flattened representation of a data structure consisting of a collection of objects that can refer to each other in any arbitrary way."

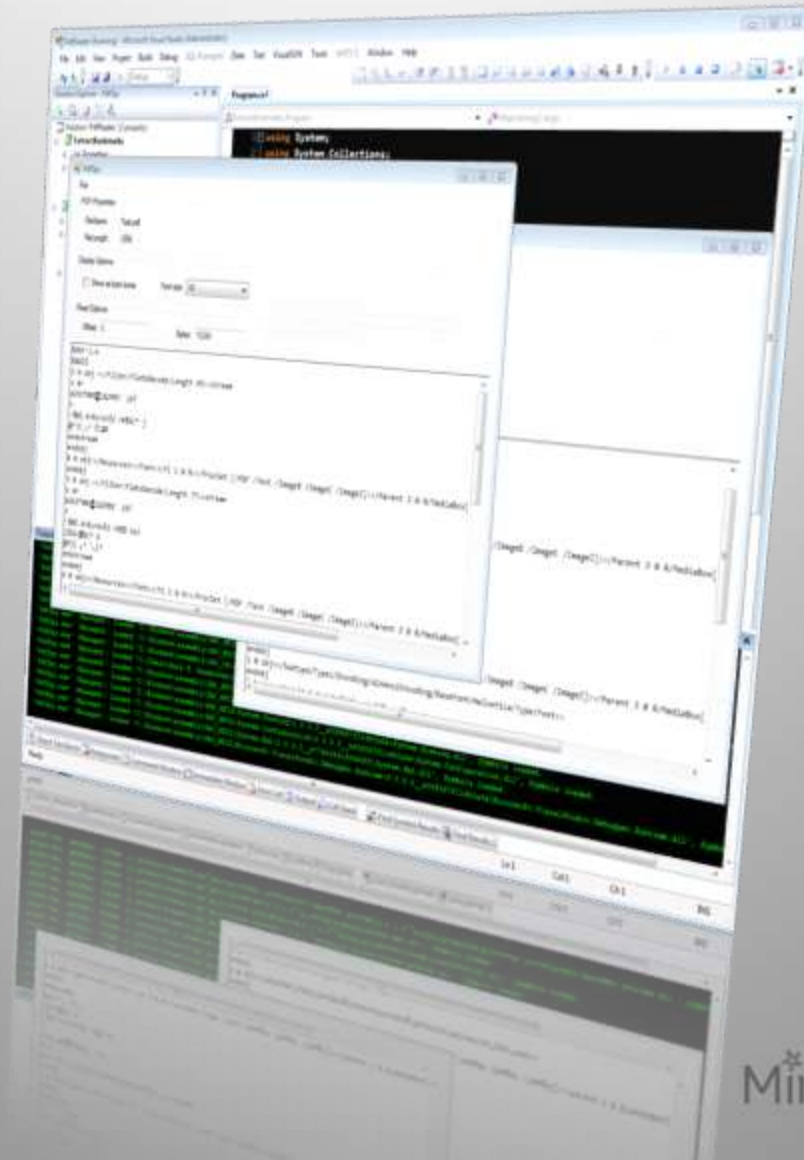
- Boolean, Number, String, Name, Array, Dictionary, Stream, Null
- Indirect & direct objects
- Random access

Incremental Changes

- Fast saves, but not for free
- Undo & history
- Save vs Save As
- Single-pass writing
- Linearization



Linearization & Xref Chaining



mow

M M X

Miracle OpenWorld 2010

ABCpdf

- Commercial
- Excellent .NET API
- ObjectSoup is a valuable friend
- Good image rendering
- Useless SWF rendering
- Unstable rendering
- Decent support
- <http://www.websupergoo.com/secret.htm>

```
using (var doc = new Doc())
{
    doc.Read("Test.pdf");
    doc.PageNumber = 1;
    doc.Rect.String = doc.CropBox.String;
    doc.Rendering.Save("Page_1.png");
}
```

Attempted to read or write protected memory. This is often an indication that other memory is corrupt.

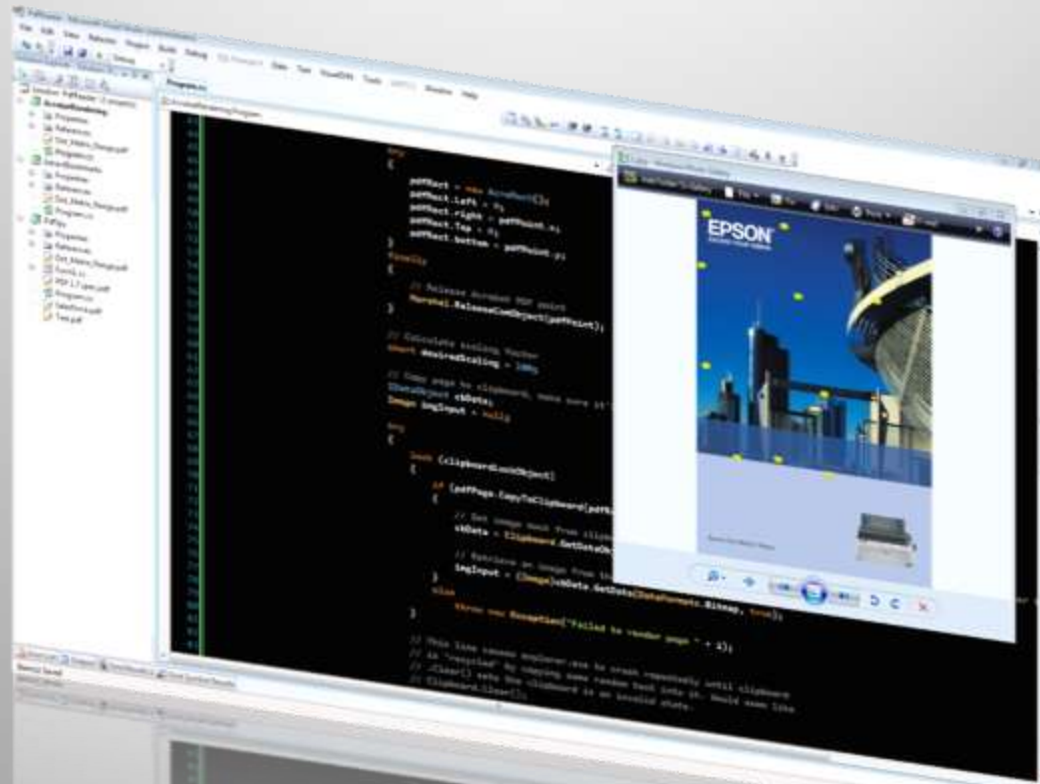


Acrobat

- Commercial (tricky license)
- No COM libraries after 7.x
- Surprisingly stable and fast
- Ugly API



Rendering Using Acrobat



mow

M M X

Miracle OpenWorld 2010

Xpdf

- Open source (GPL)
- Pdffonts, pdfimages, pdfinfo, pdftops, pdftotext
- Basis for many other libraries & tools
- Commercial license & COM library available at www.glyphandcog.com
- <http://www.foolabs.com/xpdf/>



PDF Font Management

- Client must have fonts used in PDF document
- However...
 - Complete font can be embedded
 - Or a subset
 - 14 standard fonts (Courier, Helvetica, Times + ITC Zapf & Dingbats)
 - Font replacement

Text In PDF

- No concept of text, just characters
- Flow order not guaranteed
- Requires guesstimation to extract text
- Extraction may require embedded fonts
- Lots of tools, some better than others

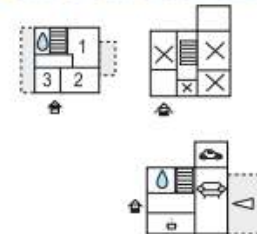
Text According To ABCpdf

1 Ostseeküste

2 Saisonzeiten auf Seite 101



PPO 652 - KOLCZEWO ★
 Karte: A2 1994/04 8 P. 180/725 m²
 Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzyzdroje (Misdroy) und etwa 30 km von Swinoujście (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage.
 Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp., 200 L Gefrierfach, Waschmaschine, Zentralheizung Kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche. 2 Badezimmer: (WC, Wanne.)+(WC, Dusche.)
 3 Schlafräume: (1D)+(1D)+(1D) Wohnzimmer: (1D)
 In der Nähe: Golfplatz.



80 80 300 6 30

 A 870 B 740 C 610 D 429 E 359

- 20
- 2 Saisonzeiten auf Seite 101
- 5 Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzyzdroje (Misdroy) und etwa 30 km von Swinoujście (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage. Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp., 200 L Gefrierfach, Waschmaschine, Zentralheizung Kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche. 2 Badezimmer: (WC, Wanne.)+(WC, Dusche.) 3 Schlafräume: (1D)+(1D)+(1D) Wohnzimmer: (1D) In der Nähe: Golfplatz.
- 3 PPO 652 - KOLCZEWO
- 4 Karte: A2 1994/04 8 P. 180/725 m²

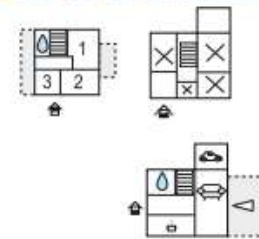
Text According To Xpdf

1 Ostseeküste

2 Saisonzeiten auf Seite 101



PPO 652 - KOLCZEWO ★
 Karte: A2 1994/04 8 P. 180/725 m²
 Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzyzdroje (Misdroj) und etwa 30 km von Swinoujscie (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage.
 Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp. 200 L Gefrierfach, Waschmaschine, Zentralheizung kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche, 2 Badezimmer: (WC, Wanne.)+(WC, Dusche.)
 3 Schlafräume: (1D)+(1D)+(1D) Wohnzimmer: (1D)
 In der Nähe: Golfplatz.



80 80 300 6 30

 A 870 B 740 C 610 D 429 E 359

1 Ostseeküste
 3 PPO 652 - KOLCZEWO
 4 Karte: A2 1994/04 8 P. 180/725 m² Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzyzdroje (Misdroj) und etwa 30 km von Swinoujscie (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage. Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp. 200 L Gefrierfach, Waschmaschine, Zentralheizung kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche, 2 Badezimmer: (WC, Wanne.)+(WC, Dusche.) 3 Schlafräume: (1D)+(1D)+(1D) wohnzimmer: (1D) In der Nähe: Golfplatz.
 80 80 300 6 30

2 Saisonzeiten auf Seite 101



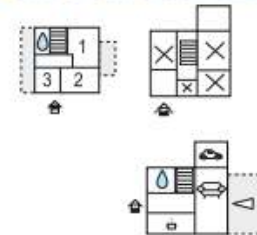
Physical Text According To Xpdf

1 Ostseeküste

2 Saisonzeiten auf Seite 101



PPO 652 - KOLCZEWO 
 Karte: A2 1994/04 8 P. 180/725 m²
 Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzzydroje (Misdra) und etwa 30 km von Swinoujście (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage.
 Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp., 200 L Gefrierfach, Waschmaschine, Zentralheizung kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche. 2 Badezimmer: (WC, Wanne,)+(WC, Dusche.)
 3 Schlafräume: (1D)+(1D)+(1D) Wohnzimmer: (1D)
 In der Nähe: Golfplatz.



80 80 300 6 30

 A 870 B 740 C 610 D 429 E 359

ostseeküste

saisonzeiten auf Seite 101

1

- 3
- 4
- 5

2

PPO 652 - KOLCZEWO
 Karte: A2 1994/04 8 P. 180/725 m²
 Dieses hübsche und gut ausgestattete Ferienhaus mit zwei Balkonen liegt etwa 2,5 km von einem Ostseestrand, 15 km von Miedzzydroje (Misdra) und etwa 30 km von Swinoujście (Swinemünde) an der Ostseeküste entfernt. Von der Terrasse haben Sie Aussicht auf einen nahe gelegenen See mit guten Angelmöglichkeiten. Die Umgebung und ein Nationalpark laden zu einem aktiven und gesunden Ferienaufenthalt ein. Drei Fahrräder und ein Gartengrill stehen Ihnen zur Verfügung. Auf dem geschlossenen Grundstück befinden sich zwei Parkplätze und eine Garage.
 Hausinfo: 4 TV, Sat-TV, Tel. f. eing. Gespr. Kamin, E-Herd, Mikrowelle, Geschirrsp., 200 L Gefrierfach, waschmaschine, Zentralheizung kinderfreundlich, 3 Garagen, Seeblick, Gartenmöbel, Grill, Schaukel, Sandkasten, Ruderboot mietbar, 40 EUR/Woche. 2 Badezimmer: (WC, Wanne,)+(WC, Dusche.)
 3 Schlafräume: (1D)+(1D)+(1D) wohnzimmer: (1D)
 In der Nähe: Golfplatz.

6

80 80 300 6 30
 A 870 B 740 C 610 D 429 E 359

SWFTTools

- Open source (GPL)
- PDF2SWF converts PDF files to SWF format
 - Based on Xpdf
 - Active mailing list
 - Author actively working on project
 - Use dev snapshots / git repo
 - Stable, but some kinks
- <http://www.swftools.org>

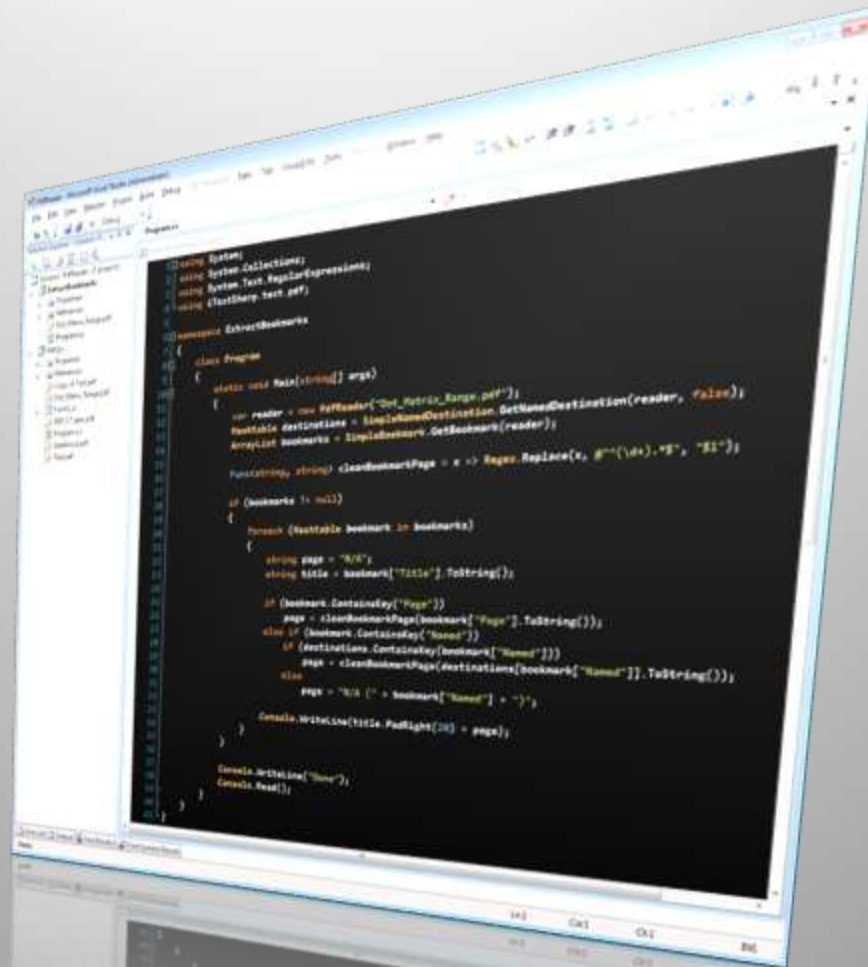


iTextSharp

- Open source (5.0 – AGPL(!), 4.1 - LGPL)
- Commercial license available
- .NET port of iText
- Very stable
- Excellent for creating & modifying PDFs
- No rendering capabilities
- <http://itextsharp.sourceforge.net/>
- <http://itextpdf.com/>



Extracting Bookmarks



```
using System;
using System.Collections;
using System.Text.RegularExpressions;
using iTextSharp.text.pdf;

namespace ExtractBookmarks
{
    class Program
    {
        static void Main(string[] args)
        {
            var reader = new PdfReader("Dot_Matrix_Range.pdf");
            PdfDestination destination = SimplePdfDestination.GetNamedDestination(reader, false);
            ArrayList bookmarks = SimpleBookmark.GetBookmarks(reader);

            PdfContentStream contentStream = reader.GetContentStream();
            PdfDictionary dictionary = contentStream.GetDictionary();
            PdfArray array = dictionary.GetArray(PdfName.Names);

            if (array != null)
            {
                foreach (PdfObject bookmark in array)
                {
                    string page = "N/A";
                    string title = bookmark.GetString(PdfName.Title).ToString();

                    if (bookmark.ContainsKey(PdfName.Page))
                    {
                        page = SimplePdfDestination.GetNamedDestination(reader, false).ToString();
                    }
                    else if (bookmark.ContainsKey(PdfName.Named))
                    {
                        PdfDictionary dictionary = bookmark.GetDictionary(PdfName.Named);
                        PdfArray array = dictionary.GetArray(PdfName.Names);
                        page = SimplePdfDestination.GetNamedDestination(reader, false).ToString();
                    }
                    else
                    {
                        page = "N/A (" + bookmark.GetString(PdfName.Named) + ")";
                    }

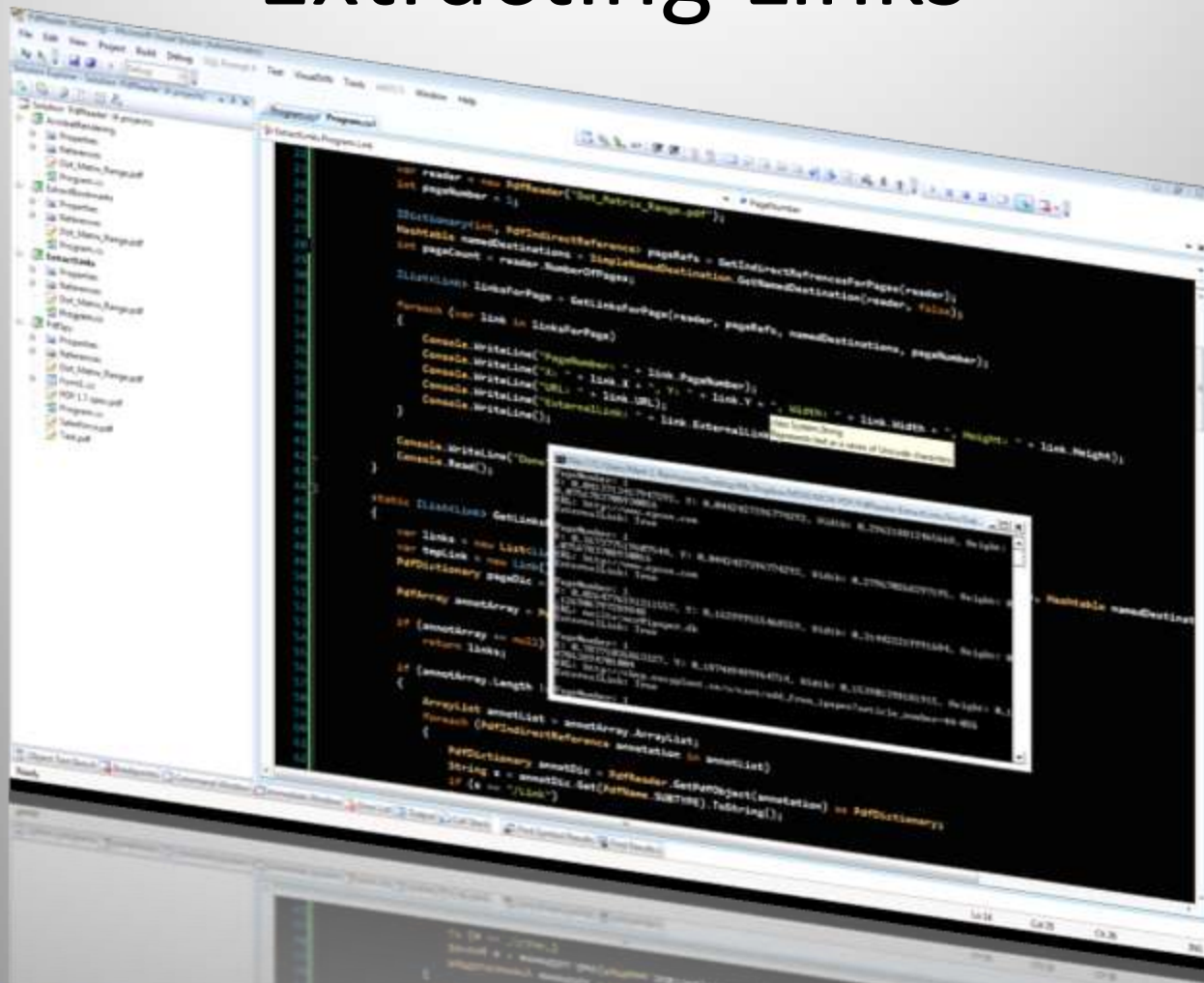
                    Console.WriteLine(title.PadRight(20) + page);
                }
            }

            Console.WriteLine("Done");
            Console.ReadLine();
        }
    }
}
```



M M X

Extracting Links



Thank you!

For attending this session



M

M

X

Miracle^{*} OpenWorld 2010

Email

mark@improve.dk

Twitter

@improvedk

Blog

improve.dk



M

M

X

Miracle^{*} OpenWorld 2010