

**White Paper**

# **Metadata in Microsoft Office and in PDF Documents**

Types, Export, Display and Removal

Copyright 2002 - 2009 soft Xpansion GmbH & Co. KG

Contents

Term Definitions and Types..... 2

    Definitions ..... 2

    Several Types ..... 2

Metadata in Office Files ..... 2

    Word ..... 4

    Excel ..... 5

    PowerPoint ..... 5

Metadata in PDF Documents ..... 6

    General Metadata..... 6

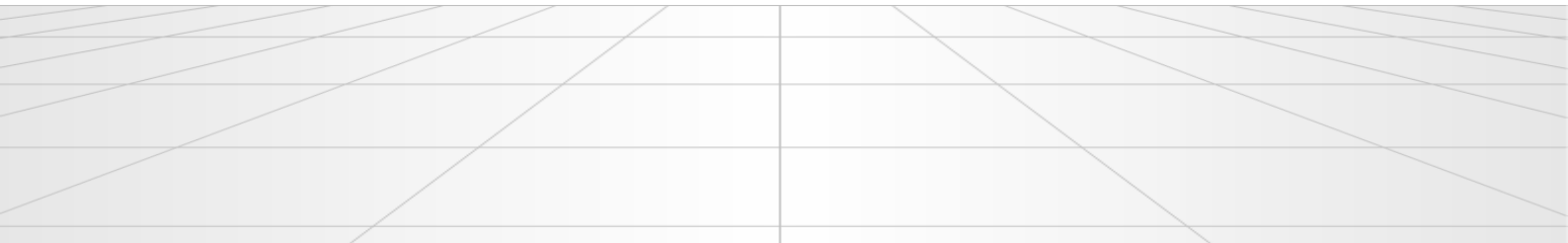
    Metadata from Office Documents ..... 6

    Examples of Conversion ..... 7

Presentation and Display ..... 9

Removing Metadata ..... 10

**Tip: All titles and names of products and companies mentioned in this document are trademarks or brand names of their respective manufacturers or legal owners.**



## Term Definitions and Types

### **Definitions**

According to the first quite general definition, the metadata is such data, which describes other data or gives information about this data. So, for example, letters or characters in a text are data. The number of letters in a text is the metadata, which can be also described as the additional information about the "source" data.

If a narrower sense is implied and the term is used in connection with file types, the metadata means, for example, such information as the name and the title of a file, its author, keywords to the contents of a file or the date of the saving. Then metadata or meta-information can be characterized in such a way that it is saved in a file, but as a rule is not visible at the first sight when someone opens the file in application software. Also at the printing of file contents (for example, of a document) on paper most metadata becomes lost. As this usually does not happen at the transfer of the files, they can pass on more information than their author actually wants to reveal.<sup>1</sup>



On the other hand, an obvious positive effect of metadata in files is: they allow cataloguing and browsing of data arrays according to certain general criteria in a simpler and more precise way.

### **Several Types**

In addition to the text documents, metadata can be also contained, for example, in music files, video files, photos or html pages. In music files, among the rest, it is the information about an artist and a title or special tags. In videos it includes, for example, record time, size and picture rate. In photos metadata is often saved in the form of picture information in Exif or IPTC format. Metadata in HTML pages (page title, keywords to the page contents etc.) is used to indicate pages in the result lists of search engines.

The following overview concerns only the use of metadata in Microsoft Office files and in PDF documents. At that the three Office applications – Word, Excel and PowerPoint will be considered.

## Metadata in Office Files

### **Office Documents in General**

In Office files a huge amount of meta-information is gathered at the saving of the document. On the one hand, it is the information, which can be specified in Word, Excel and PowerPoint as "Document Properties" and viewed (and changed if necessary) in all three document types in the same place of the application. On the other hand, it is the information, which can be accessed only within the respective application.

Some of this information (Document Properties) can be viewed after clicking on "Properties" in the "File" menu of Office versions 2003 and earlier. In Word, Excel and

---

<sup>1</sup>See Rost, Martin, Wallisch, Arnold: Was Office-Dateien verraten können, c't issue 2-2003, page 172ff. (in German language)

PowerPoint 2007 it is necessary to click on the Microsoft Office button first, then on "Prepare" and finally on "Properties". Alternatively the properties in Office 2007 can be also viewed at the opening or saving of a file, namely in the dialog boxes "Open" or "Save as". The majority of properties of a currently opened document can be not only viewed, but also changed. In addition to the default properties it is also possible to set the user-defined properties.<sup>2</sup>

The most important general document properties (explained on the example of Word 2007) are:

- ❖ Standard properties, which can be changed by the user in the display dialog. They are: author (the name, which was given by the user at the installation will be automatically inserted), title, subject, keywords, comments.

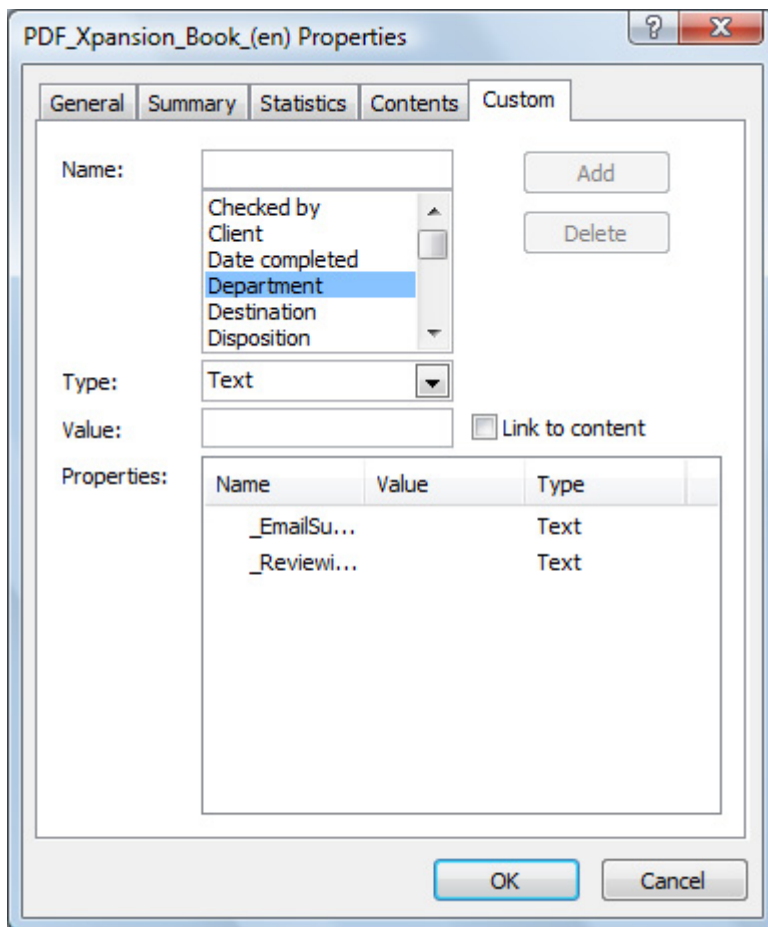
- ❖ Dynamically updated or updateable properties, which cannot be changed by the user in the display dialog. They are: saving and processing time, saving path, statistical data (number of pages, paragraphs, lines, words etc.).

Statistic name	Value
Pages:	39
Paragraphs:	218
Lines:	681
Words:	5217
Characters:	24834
Characters (with spaces):	29843
Bytes:	6368256

All auto text fields in general also belong to the dynamic meta-information, as well as "storages" for repeatedly used text or graphics, which should be used again.

<sup>2</sup> For details see the help and instructions on <http://office.microsoft.com/de-de/help/HA100475241031.aspx> (in German language)

- ❖ User-defined properties, which can be provided with digits, texts, dates or the value "Yes" / "No".



In addition to the general document properties the following meta-information is saved in the Office files and is specific for Word, Excel and PowerPoint respectively.

## **Word**

Application-specific metadata in Word files is, for example,

- ❖ Security options (information about encryption and password protection)
- ❖ References (links): Navigation information about jump points of internal links between the pages of a Word document and external or Internet links
- ❖ Table of contents (TOC) with jump points information about a page, which should be opened by clicking on a heading in the table of contents
- ❖ Information about form fields, common control elements and web tools
- ❖ (Hidden) annotations and comments
- ❖ Footnotes, endnotes
- ❖ Tips about updates
- ❖ (Hidden) markup tips

- ❖ Hidden texts and (invisible) page backgrounds
- ❖ Index

## ***Excel***

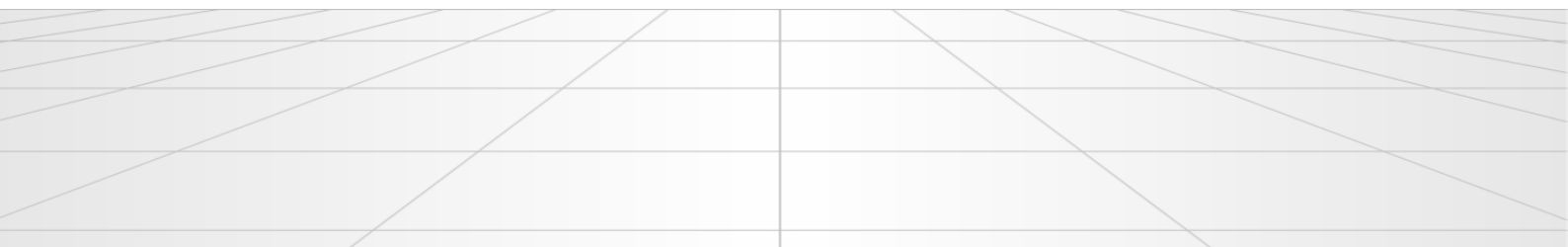
Application-specific metadata in Excel worksheets is, for example,

- ❖ Security options (information about encryption and password protection)
- ❖ References (links): Navigation information about jump points of internal links between the sheets of an Excel worksheet and external or Internet links
- ❖ Information about available form fields or common control elements
- ❖ Comments
- ❖ Grid lines
- ❖ Lines and columns

## ***PowerPoint***

Application-specific metadata in PowerPoint presentations is, for example,

- ❖ Security options (information about encryption and password protection)
- ❖ References (links): Navigation information about jump points of internal links between the slides of a presentation and external or Internet links
- ❖ Information about available form fields or common control elements
- ❖ Notes/comments
- ❖ Grid lines
- ❖ Lines and columns
- ❖ (Hidden) slides
- ❖ Slide borders
- ❖ Adjusted grayscale representation



## Metadata in PDF Documents

### **General Metadata**<sup>3</sup>

Independently of the file format of the original application, PDF files usually contain general metadata, such as title and author of a document and information about its creation and modifications. In PDF files metadata can be changed or added by the user or by the plug-in extensions, at that the representation in PDF readers by Adobe is limited by the document properties on maximum 255 bytes.<sup>4</sup> Since version 1.4 of the PDF specification, metadata can be also defined for single components of a document. In a PDF document the metadata is saved either in a document information directory or in metadata streams.

### **Metadata from Office Documents**

Provided that the source file is an Office file of Word, Excel or PowerPoint, the application-specific meta-information, listed above for the three programs, can be transferred from the Office files to the PDF file by PDF converters, which control the conversion of metadata. Such converters are, for example, [Office to PDF Premium](#), [Word to PDF](#) or [Excel to PDF](#) by soft Xpansion.

As it concerns the Office documents on the one hand and the PDF files on the other hand, these are the file types developed separately of each other, so the conversion of a file type does not occur "identically" for all types of metadata. That's why the following circumstances are to be followed concerning the conversion:

1. Not all metadata from Office files can be imported in PDF documents. So technical requirements for the conversion of some metadata in the object model of the PDF format, for example, the metadata of the auto text fields, are missing. However, this restriction has no effect on the field contents visible in the document text.
2. Vice versa in some cases the conversion in a PDF file offers more functionality than there is in the Office file. For example, it is technically possible to import an index from a Word file in a PDF file, so that the single entries will be indicated in the directory of the PDF file as active references. Then it is possible to click on these references to jump to a page given in the index.

It this case it can be a problem, that in the PDF format, as opposed to the Word document, there is no possibility to perform a "field update" command in the index of a PDF file automatically after adding or deleting some PDF pages to bring them to the newest state.

---

<sup>3</sup> See Adobe Systems Incorporated, PDF Reference, version 1.7, pages 843 - 847, on the basics of metadata in PDF files

<sup>4</sup> *ibid.*, page 1125, annotation 160

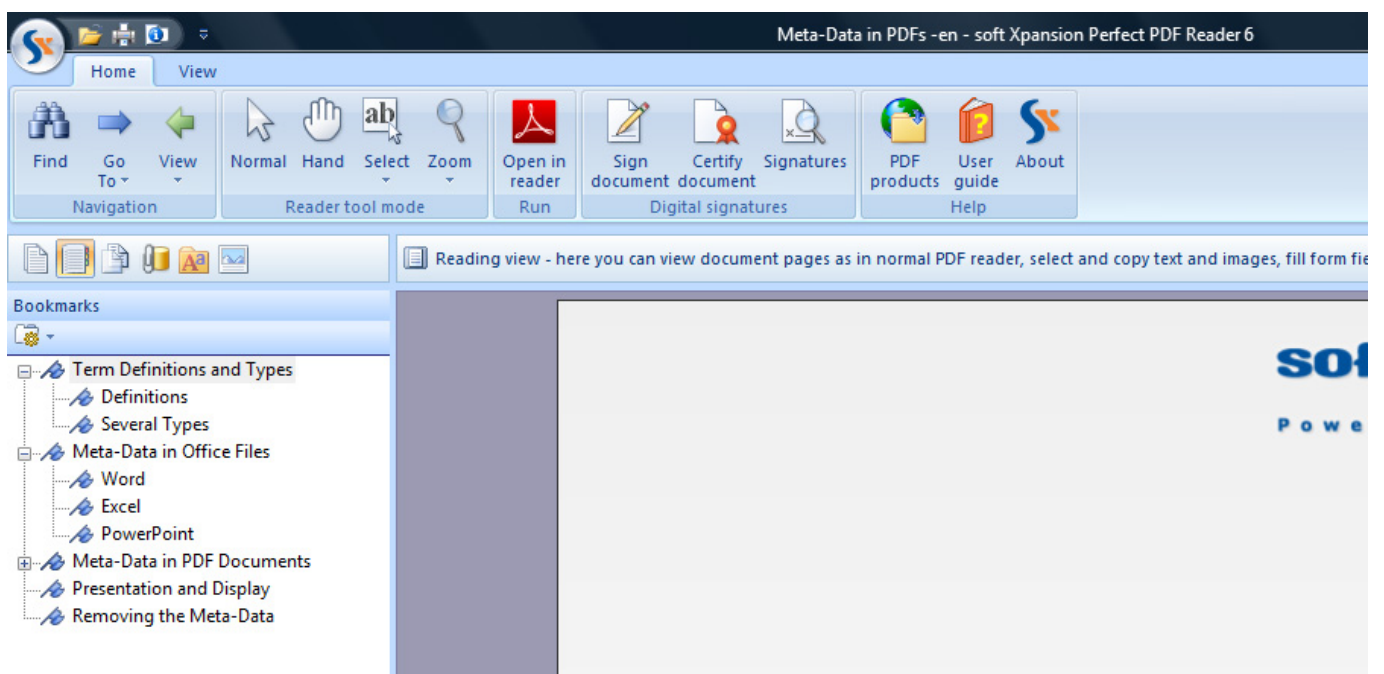
3. Finally, at the conversion there are degrees of freedom of the transfer for the developers of PDF converters on how this can be performed specifically.

## ***Examples of Conversion***

The following information about the table of contents/TOC, footnotes/endnotes and text processing shows on an example that the transfer and representation of metadata can occur in different ways.

### **1. Table of Contents/TOC**

In Word tables of contents can be automatically generated from the chapter headers, paragraphs and subparagraphs. Such table of contents is placed in the beginning of a Word document as a collection of active links. With a mouse click on a link it is possible to reach the place in the document where the respective part of text begins.



The view of table of contents in a PDF viewer (here [Perfect PDF Reader](#)) as bookmarks

In a PDF file the table of contents can have the same function as in the Word file, so it is possible to proceed directly from the respective shortcuts and jump to a certain text header. In addition, there is still a possibility to embed the table as a collection of bookmarks in a form of a list (directory tree) that is displayed not in the document text, but is separated in an own window area of the PDF viewers.

### **2. Footnotes/Endnotes**

Footnotes and endnotes are embedded in Word in a way that the footnote or endnote text is displayed above the footnote/endnote character in an info field, when a mouse pointer moves to the footnote/endnote character in a document. If



the character is clicked twice, the view jumps directly to the accompanying footnote or endnote.

Some of this information (Document Properties) can be viewed after clicking on "Properties" in the "File" menu of Office versions 2003 and earlier. In Word, Excel and PowerPoint 2007 it is necessary to click on the Microsoft Office button first, then on "Properties". Alternatively the properties in Office 2007 can be also viewed of a file, namely in the dialog boxes "Open" or "Save as". The currently opened document can be not only viewed, but also edited. In the default properties it is also possible to set the user-defined properties.<sup>2</sup>

For details see the help and instructions on <http://office.microsoft.com/de-de/help/HA100475241031.aspx>

In PDF files footnotes and endnotes are either simply imported, or in such a way that only one variant is selected. For example, with [Office to PDF Premium and Word to PDF](#) by soft Xpansion only the info field is transferred. A double click here has another function: in these programs the field is provided with an icon and can be shown or hidden with a double click.

printing of file contents (for example, of a document) on paper most meta-data becomes lost. As this usually does not happen at the transfer of the files, they can pass on more information than their author actually wants to reveal.<sup>1</sup>

On the other hand, an ob-  
cataloguing and browsing  
and more precise way.

### Several Types

In addition to the text do-  
files, video files, photos c-  
about an artist and a title  
size and picture rate. In p-  
in Exif or IPTC format. M-  
etc.) is used to indicate p-

See Rost, Martin; Wallisch, Arnold, Was Office-Dateien verraten können, c't issue 2-2003, page 172ff. (in German language)

The following overview concerns only the use of meta-data in Microsoft Office files and in PDF documents. At that the three Office applications – Word, Excel and PowerPoint will be considered.

## 3. Text Processing (Tips to Updates, Markup Tips)

As it is known, in Word text processing can be tracked, the concrete updates can be captured, indicated in color, shown on the edge of a page in a form of speech bubbles and displayed in an info field ("Deleted", "Pasted"), and even underlined in addition.

3. Finally, at the conversion there are degrees of freedom of the transfer for the developers of PDF converters on how this can be performed specifically.

### Examples *of* Conversion

The following information about the table of contents/TOC, footnotes/endnotes and text processing shows on an example that the transfer and representation of meta-data can occur in different ways.

Deleted: for

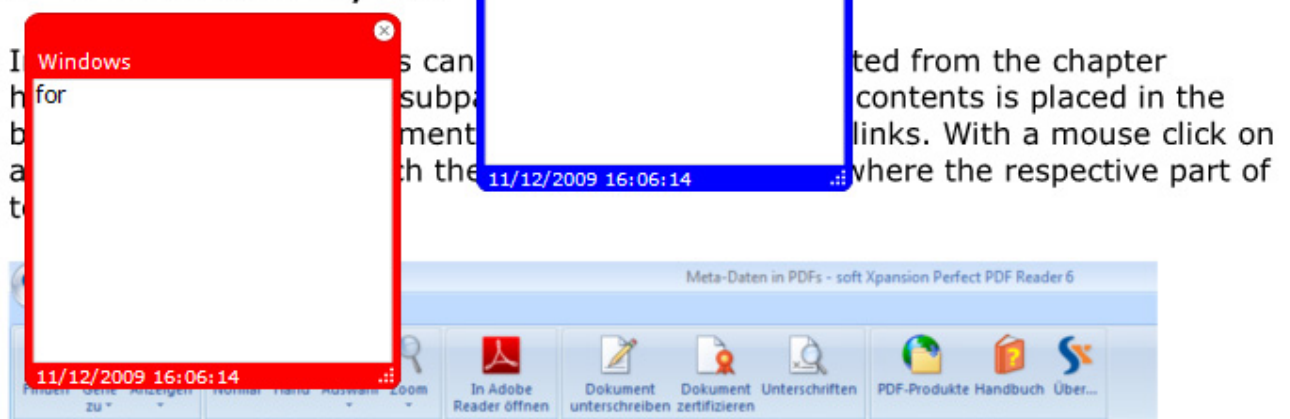
The corresponding PDF file can display processed text likewise or in another way. In [Office to PDF Premium and Word to PDF](#) by soft Xpansion red question marks and framed info fields are usually used as tips to the deleted text. Additions to text can be recognized by blue symbols and info fields framed by the same color. They contain additional information compared with the representation in Word with the update data, time and the computer name on which the update was performed.

3. Finally, at the conversion there are degrees of freedom of the transfer for the developers of PDF converters on how this can be performed specifically.

### Examples *for* Conversion

The following information about the table of contents/TOC, footnotes/endnotes and text processing shows on an example that the transfer and representation of meta-data can occur in different ways.


#### 1. Table of Contents/TOC




## Presentation and Display

Provided that a conversion is performed with the creation of a file, such meta-information as author, title, subject or keywords in PDF files is usually displayed in dialog boxes, which are named "Document Properties" or similar. Other metadata as for example annotations, notes and comments, grid lines, slides or slide borders, which is hidden in the original document, will be transferred to the PDF document and displayed there, if a conversion is performed or allowed with the creation of a PDF file.

## Removing Metadata

It was explained above, what metadata from Office files can be transferred in PDF and how it can be performed. However, in certain cases it makes sense to remove all or some of the contained metadata, before the conversion in PDF is performed and a file is proceed. Namely when an author of the document does not want to share it. There were significant cases, when an author missed such information in a document about the weapons of mass destruction supposedly located in Iraq, which should have justified the interference of Great Britain in the Iraq war<sup>5</sup>  w to remove a huge amount of metadata from the Office files comfortably, is explained below.

To delete the metadata in the original applications Word, Excel and PowerPoint there are the following ways available:

1. First there is a possibility to use program settings within Word, Excel and PowerPoint to remove the thoroughly examined data from the file properties.<sup>6</sup> 
2. Besides, the special tools provided by Microsoft can be used to delete some of the metadata:

For Office version 2003 and lower it is explained on page

<http://www.microsoft.com/downloads/details.aspx?FamilyId=144E54ED-D43E-42CA-BC7B-5446D34E5360&displaylang=en>

In Office 2007 removing can be performed with the help of a function called *Document Inspector*, see

<http://office.microsoft.com/en-us/help/HA100375931033.aspx>

In addition, there is a freeware and a commercial tool of another manufacturer, the *Document Trace Remover* and the *Metadata Analyzer*.

---

<sup>5</sup> [http://www.chip.de/artikel/Verraeterisches-Office-2\\_12875097.html](http://www.chip.de/artikel/Verraeterisches-Office-2_12875097.html) (in German language)

<sup>6</sup> [http://www.chip.de/artikel/Verraeterisches-Office-4\\_12875104.html](http://www.chip.de/artikel/Verraeterisches-Office-4_12875104.html) (in German language)