Highlight File Format

Technical Note #5172

Version : Acrobat 6.0

ADOBE SYSTEMS INCORPORATED
Corporate Headquarters
345 Park Avenue
San Jose, CA 95110-2704
(408) 536-6000
http://partners.adobe.com

July 2003

Contents

Contents

Highlight File Format

Introduction

This technical note describes the file format and URL specification that allows a Web server to highlight text in a PDF file being displayed by version 3.0 and above of the Adobe® Acrobat® viewers (Acrobat and Reader).

To use the highlight file, the Acrobat viewer must include the highlight plug-in and the External Window Handler (EWH) plug-in. In addition, the Web browser must use the Adobe Acrobat plug-in for Netscape Navigator — any Web browser that accepts this Netscape Navigator plug-in can be used. Highlight files can also be used with the ActiveX control for Internet Explorer 4.0 and later.

The highlight file is an ASCII file that is downloaded separately by the Web browser and used by Acrobat to highlight words in the PDF file.

If you received this technical note without obtaining the entire Acrobat Software Development Kit (SDK), you can get the complete SDK by visiting:

http://partners.adobe.com/asn/developer/acrosdk/main.html

File Format

Although the file begins with **<XML>** and is specified in the URL with **#xml**, the file format is not true XML and only supports the syntax and keywords as described in this document.

The highlight file must begin with the line:

```
<XML>
```

followed by the line:

```
<Body units=[characters |words ] version=verNum>
```

In this line, a vertical bar (|) within brackets separates choices from which one and only one must be used. The brackets should not be included in the highlight file.

The value of the **units** key indicates whether highlights are given as character offsets on the page (***characters***) or as word offsets on the page (***words***).

NOTE: Because of a bug in the word finding algorithm of the 3.01 Viewer (Reader and Exchange), it is strongly suggested that character offsets be used in the highlight file. If word offsets are used, incorrect words may be highlighted. This occurs when the word offsets are calculated using other (non 3.01) versions of the viewer, the text extraction toolkits or the PDF Library and then used to highlight text in PDF

files displayed in the 3.01 viewers. For a CJK text, it is necessary to use a character offset.

The value of the **version** key is used when **units** is set to words. It indicates the version of the Acrobat viewer's word finding algorithm for which word offsets are specified. Because the characters that make up a word may not be stored consecutively in the PDF file, an algorithm must be used to find words; different versions of the algorithm (corresponding to different versions of Acrobat) may give different word offsets. The version for Acrobat 3.0, 3.01, and 4.0 is 2.

NOTE: Previous versions of the Highlight File Format documented two other keys in this section of the file. These were color and mode. Color was documented as specifying the color of the highlight and mode was documented as determining whether the viewer would go to the first page containing highlighted text when opening the PDF file. This functionality was never implemented and therefore documentation of these keys has been removed. The highlights will always use the highlight color as set by the operating system and the viewer will always display the page containing the first highlighted text.

The next line must be:

```
<Highlight>
```

followed by one or more highlight ranges.

Each highlight range is a line of the form:

```
<loc pg=P pos=O len=L>
```

The value of **pg** is an integer that specifies the page on which the highlight is located. Pages are numbered sequentially, with the first page in a file having a page number of zero.

The value of **pos** is an integer that specifies the offset of the highlight on the page. Offsets are specified either in words or characters, depending on the value of **units** in the **Body** line described previously. The offset of the first character/word on a page is zero.

The value of **len** is an integer that specifies the number of words or characters to highlight. Note that version 3.0, 3.01 and 4.0 of the Acrobat viewer only support highlighting of entire words; the entire word is highlighted if at least one of its characters are highlighted.

Following the list of highlight ranges are the following lines:

```
</Highlight>
</Body>
</XML>
```

The example below shows a sample a highlight file. This file contains two highlights, specified as character offsets (**units=characters**). The Acrobat viewer will automatically display the first page that contains a highlight. No word finding algorithm is specified because the file uses character offsets. The first highlight range contains 10 characters, beginning with the first character on the document's first page. The second

highlight range contains two characters, beginning with the fifth character on the document's second page.

```
<XML>
<Body units=characters>
<Highlight>
<loc pg=0 pos=0 len=10>
<loc pg=1 pos=4 len=2>
</Highlight>
</Body>
</XML>
```

URL Format

The URL that specifies the PDF file must also specify the highlight file, as illustrated in the following example:

```
http://www.adobe.com/a.pdf#xml=http://www.adobe.com/a.txt
```

The URL consists of two pieces, separated by a pound sign (#). The first part of the URL specifies the PDF file (`a.pdf` in this example). The second part specifies the highlight file (`a.txt`) and the auxiliary data handler to which its contents should be sent (`xml`, the name of the highlight plug-in's auxiliary data handler).

Debugging Highlight Files

If you have a Web server, the easiest and most thorough way to debug a highlight file is to place the files on your server and specify an extended URL. You can, however, write a simple plug-in for Acrobat which will pass a highlight file to the highlight plug-in. Because this method bypasses the EWH plug-in, it does not enable the **Next Highlight** and **Previous Highlight** toolbar buttons.

The following code reads in data from a highlight file and uses it to highlight text in the currently displayed document.

```
{
     ASBool status;
     AVDoc avdoc = AVAppGetActiveDoc();
     FILE   *stream;
     char auxData[255];
     ASInt32 numchar;

     char * filename = "C:\\test.txt";
     stream = fopen (filename, "r");
     numchar = fread (auxData, sizeof(char),255,stream);
     status = AVDocSendAuxData (avdoc,
            ASAtomFromString ("xml"), auxData, numchar);
}
```