Xed: a new tool for eXtracting hidden structures from Electronic Documents

Karim Hadjar, Maurizio Rigamonti, Denis Lalanne and Rolf Ingold

DIUF, University of Fribourg
Chemin du Musée 3, 1700 Fribourg, Switzerland
{karim.hadjar, maurizio.rigamonti, denis.lalanne, rolf.ingold}@unifr.ch

Abstract

PDF became a very common format for exchanging printable documents. Further, it can be easily generated from the major documents formats, which make a huge number of PDF documents available over the net. However its use is limited to displaying and printing, which considerably reduces the search and retrieval capabilities. For this reason, additional tools have recently appeared that allow to extract the textual content. However their practical use is limited in the sense that the text's reading order is not necessary preserved, especially when handling multi-column documents, or in presence of complex layout. Our thesis is that those tools do not consider the hidden layout and logical structures of documents, which could greatly improve their results.

We propose a novel approach to overcome the document content extraction, by merging a) low-level extraction methods applied on PDF files with b) layout analysis performed on a synthetically generated TIFF image. The paper describes the various steps necessary to achieve this task. Finally, we present a first experiment on the restitution of the newspapers' reading order which shows encouraging results.

1. Introduction

Nowadays a huge number of documents are either created or converted in PDF format. PDF became the universal exchange format. The main advantage of this format is its portability, viewing and printing abilities. Although PDF allows to embed structural information, usual PDF converters don't use them. In fact, the converters generating PDF documents produce them in an automatic manner with the only goal of preserving the presentation. These converters use only PDF's low level based features such as font types, size, color, text position, etc. This tendency is mainly due to PDF's complex generation process. Nevertheless, PDF format has few drawbacks related mainly for search and extraction abilities. For instance, searching a word within the second column of a multi-column document using the search feature of Acrobat might be unsatisfactory; if the same word is located in the first column or else where outside the second column the Acrobat reader will return them all in order of appearances ignoring user's interest search area. Also if a user searches for an article talking about "dryness in California", a newspaper cover page containing an article about "dryness in Africa" and another article about "California dream", could be returned, which would not satisfy the request. So there is a need for extracting objects in a structured form and save the document in XML format in order to allow indexing, more accurate searching, and creating applications dealing with versioning and meta data [17].

Many researches and products have been developed in the past few years that are able to extract data from PDF documents and put it either into text or into XML format. There is an interesting research that deals with the extraction of the logical structure of PDF files [15]. However, most of them do not include analysis features, which permit to embed extra information related to physical and logical document structure, which would significantly enrich the original document. For this reason, we believe that running traditional document analysis algorithms on PDF documents will drastically improve PDF's content extraction. This will allow us to develop high-level applications such as the restitution of the reading order of a document, indexing, comparing, archiving, versioning and extracting meta data.

In this paper we present a new tool named Xed (e**X**tracting **e**lectronic **d**ocuments) for automatic extraction from PDF file. Xed extracts all the objects from a PDF document including text, images, and graphics. The output of the extracted objects is either in SVG (Scalable Vector Graphics) [20] or an "in-house" format described in appendix B. Xed also includes an internal common representation format, which allows the layout analysis methods to share their results with the PDF extraction tools, such as agents would collaborate through a black-board architecture. This collaboration significantly enriches Xed extraction with traditional analysis feature that highlights the physical structure and even the logical structure in the forthcoming versions. The main advantages of Xed are thus a) the adjunction of traditional analysis features to the PDF format and b) the improvement of the layout analysis results.

This paper is organized as follows: in section 2 we briefly introduce the PDF format. Then we make a brief survey of the state-of-the-art tools for PDF analysis in section 3. Section 4 presents the analysis feature of Xed and the restitution of the reading order algorithm. In section 5 we present detailed information about Xed's architecture. Finally, section 6 brings up the conclusion and future works.

2. PDF

PDF, acronym for Portable Document Format, is the universal desktop publishing format. In fact, PDF came after PostScript, which was the desktop publishing revolution. PostScript is a page description language that allows desktop printers to render complex text and graphics images. Despite PostScript was the desktop publishing revolution its drawbacks are: nearly impossible to edit, bothersome incompatibilities between dialects of PostScript, cross platform issues and risk of errors since PostScript is a programming language. But with the advent of networks and internet, there was a need for a universal document exchange format and in 1992 PDF was created. The first goal of PDF was to view computer file on screen outside of native application, which is currently done by Acrobat reader. When PDF was introduced, the faculty to render complex text and graphics images on the screen and the printer was straightforward. By this way a PDF document can be viewed or printed on any platform and several mobile devices with the same fidelity.

PDF files can be generated from nearly all standard applications (Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Adobe Photoshop, Adobe InDesign, Adobe Illustrator, Quark XPress, etc) by using distiller and PDF writer. These files can contain many types of components (images, fonts, drawings, etc) within a single file; this has encouraged the graphic art community to adopt PDF for file exchange. In most of the cases the PDF files obtained are smaller than the native file but it depends on the version of distiller and its options. In fact, PDF compresses every object with appropriate algorithm. Adobe has provided both a SDK and the Adobe PDF library for programmers. Every PDF specification is published and is publicly available on the Adobe web site [1, 2, 3].

Since the perfect file format doesn't exist, PDF has also its drawbacks mainly for search and extraction. It is not conceived for being editable; it can't be neither edited with a text editor nor indexed in a simple way. Another disadvantage is that applications generating PDF documents don't use the various document structures, even though this feature is included inside the source format. In the following section we review the available programs and libraries for extracting data from PDF files.

3. Tools for PDF analysis

Many programs and libraries permitting the extraction of data are available on the net. Most of them are commercial, except JPedal [14], which also exists in open source.

In the following subsections, we review the major state-of-the-art tools for PDF analysis followed by their comparison.

3.1 JPedal

JPedal exists both in a commercial and open source version. It is basically a library for Java development of PDF applications. The Java packages allow the extraction of text, images, annotations and forms. The JPedal web site offers to end-users the possibility to submit PDF documents and receive by e-mail the result of the extraction in XML format. It is also possible to have the extracted data sorted to reflect the reading order of the submitted document.

3.2 SVG Imprint

MatterCast [16] provides a commercial product called SVG Imprint. The suite consists of desktop applications for single and batch conversions of PDF documents to SVG. Besides the suite there is an SDK that enables the integration of the conversion engine into server based enterprise applications. The output in SVG retains the exact layout and integrates all the fonts embedded in PDF document. The raster images inside the PDF documents are extracted as separate file.

3.3 Glance

Glance [7] provides a set of products and API all dealing with PDF. The commercial products are command line programs. Among these programs there is pdimgcomp which is a PDF image compression, conversion and extraction tool. Img2pdf and text2pdf respectively convert images and text into PDF documents. 3 heights is the inverse of img2pdf its output is images either in TIFF or JPEG or BMP. CLE, which stands for command line tool enhanced edition, includes the following programs: all of them have either in input or output PDF documents. The first one: pdcat useful for concatenation. The second one: pdsplit deals with splitting. The third one: pdsel allowing the selection of pages. The fourth one: pdw permitting the extraction of text in ASCII or Unicode. The last one: pdwebl useful for adding link annotations.

3.4 BCL

BCL [5] provides three products BCL Drake, BCL Freebird and BCL Jade, all of them have either in input or output PDF documents. The first one is useful for conversion into RTF for viewing and editing in Microsoft Word. The second one is a plug-in for Adobe Acrobat allowing the conversion into graphics (Tiff, Jpeg and Bmp). The last one: the Acrobat plug-in BCL Jade allows the extraction of text, tabular data and graphics of mouse selection. BCL does not extract the images found inside the page of a document; it only extracts the image of the whole page.

3.5 Comparison

The above products can help produce or extract features of PDF files, either with command line programs or through programming. For instance Glance provides a program named "pdw" which extracts the whole text stored inside a PDF document. We applied this program on the front page of the International Herald Tribune and we obtained the output presented in figure 1 and 2.

```
X     Y      FontSize Width   Angle   FontName:  Text
81.4  1113.7  9.0     50.5    0       BeBl:      BAGHDAD:
81.4  1095.1  8.9     150.3   0       PoRo:      Powell urges
                                                 Iraqis to agree
                                                 to a speedier
                                                 timetable
81.4  1085.9  8.9     150.2   0       PoRo:      toward self-
                                                 government in
                                                 Iraq, cau-
81.4  1076.6  8.9     59.8    0       PoRo:      tioned Sunday
147.7 1076.6  8.9     34.0    0       PoRo:      that the
188.1 1076.6  8.9     43.6    0       PoRo:      process of
81.4  1067.3  8.9     150.2   0       PoRo:      restoring
                                                 sovereignty had
                                                 to be carried
81.4  1058.1  8.9     150.2   0       PoRo:      out in stages
                                                 and might not
                                                 be seen as
81.4  1048.8  8.9     150.2   0       PoRo:      legitimate if the
                                                 pace were
                                                 overly rap-
81.4  1039.6  8.9     9.6     0       PoRo:      id.
…
```

Figure 1. From the pdw output, the manually identified first paragraph.

```
X     Y      FontSize Width   Angle   FontName:  Text
725.4 1215.6  9.0     62.5    0       BeBl:      STOCKHOLM:
796.7 1215.6  8.9     22.4    0       PoRo:      After
823.9 1215.6  8.9      6.5    0       PoRo:      a
835.3 1215.6  8.9      40.4   0       PoRo:      passionate
725.4 1206.3  8.9     150.2   0       PoRo:      campaign made
                                                 uncertain to the
                                              last
725.4 1197.1  8.9     150.2   0       PoRo:      moment by the
                                                 assassination of
                                                 Foreign
725.4  1187.8  8.9    33.4    0       PoRo:      Minister
764.0  1187.8  8.9    51.4    0       PoRo:      Anna Lindh,
820.6  1187.8  8.9    28.5    0       PoRo:      Swedes
854.3  1187.8  8.9    21.3    0       PoRo:      voted
725.4  1178.5  8.9    150.2   0       PoRo:      overwhelmingly
                                                 in a
                                                 referendum to
                                                 re-
725.4  1169.3  8.9    14.2    0       PoRo:      ject
747.0  1169.3  8.9    50.8    0       PoRo:      membership
803.0  1169.3  8.9    7.9     0       PoRo:      in
818.2  1169.3  8.9    12.3    0       PoRo:      the
837.9  1169.3  8.9    37.7    0       PoRo:      European
725.4  1160.0  8.9   150.2    0       PoRo:      single currency,
                                                 according to
                                                 final re-
725.4  1150.7  8.9    93.2    0       PoRo:      sults tallied late
                                                 Sunday.
733.8  1141.5  8.9    141.9   0       PoRo:      The margin
                                                 exceeded the
                                                 expecta-
725.4  1132.2  8.9    19.2    0       PoRo:      tions
753.4  1132.2  8.9   7.7      0       PoRo:      of
769.9  1132.2  8.9    23.8    0       PoRo:      many
800.2  1132.2  8.9    26.1    0       PoRo:      people
835.1  1132.2  8.9    17.1    0       PoRo:      here
861.0  1132.2  8.9   14.6     0       PoRo:      and
725.4  1122.9  8.9    150.2   0       PoRo:      provided a
                                                 stunning defeat
                                                 for Prime
725.4  1113.7  8.9    150.2   0       PoRo:      Minister Goran
                                                 Persson.
                                                 Analysts said
725.4  1104.4  8.9    150.2   0       PoRo:      the impact
                                                 would be felt
                                                 across Europe,
….
```

Figure 2. From the pdw output, the manually identified second paragraph.

When analyzing the figures we notice that the first and the second paragraph of text of the extracted data, corresponds respectively to the bloc number 5 and 6 in figure 3. This means that the reading order is not respected. Figure 3 represents the re-constructed reading order from the pdw output over the front page of the International Herald Tribune newspaper.

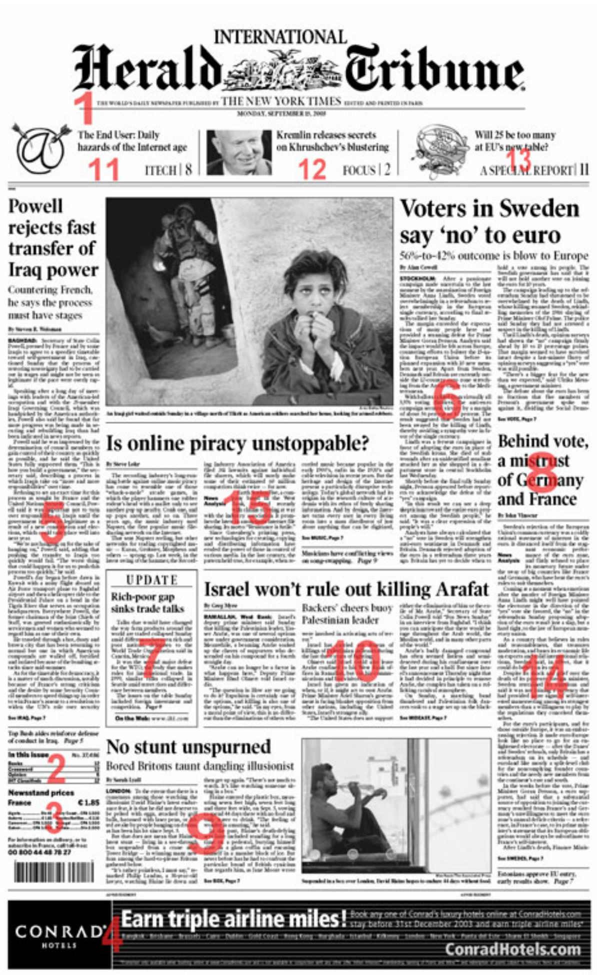


Figure 3. The reading order of the International Herald Tribune front page from the pdw output.

In order to evaluate the products listed above, we have downloaded the trial version of BCL, Glance, Mattercast and the java open source library JPedal.

Six criteria have been used to compare the output features provided by the products. Three of them are relative to the extraction of text, graphics and images. Two are relative to the format of output: XML or SVG. And finally, the last criterion corresponds to the restitution of the document reading order.

A small corpus of PDF documents has been used. It was composed of various newspapers, PowerPoint slides, graphics, and papers. The test consisted in seeking whether the criterion was supported or not. After testing all the products we obtained the following table (table 1).

Table 1. The results of the comparison.

		JPpedal	BCL	Glance	SVG Imprint
Extraction	Text	√	√	√	√
	Graphics	-	-	-	√
	Images	√	√	√	√
SVG		-	-	-	√
XML		√	-	-	-
Reading order		√	-	-	-

Although all tested tools provided text and XML output formats, only JPedal respects the reading order.

The reading order problem observed with existing tools is certainly related to the intrinsic features of the PDF format. In fact, PDF files are generated with the only goal of restituting the document's presentation accurately. To do so, a structured internal representation is not necessary, even if this feature is provided by the source file format. We observed numerous amazing results in the tools outputs. For instance, a single word missing in a paragraph reappeared as an isolated word at the end of the document. The reason we suspect to be behind such a situation, is the internal representation of the text processing system used to produce the document, in which a last minute correction is probably represented as an incremental change and thus appended to the end of the file.

The general problem we have to deal with can be stated as follows: the order in which text strings are recorded within a PDF file has nothing to do, neither with the reading order, nor with the order in which the text appears on the two dimensional rendering plane (paper or screen). Developers of extracting tools must be aware of that. However, we believe that most of the current tools try to solve the problem by defining complicated internal sorting rules. Instead of using such abstract rules, we believe that layout analysis algorithms, as developed for many years by the document analysis community, can provide additional useful information to guide this sorting process, which we call topological sorting. The next section presents this method and gives a detailed illustration of that important step.

4. Layout structure analysis

Layout analysis is an essential part both in the field of document image analysis and in recognition systems. Among the different stages of layout analysis there is the layout structure extraction. The goal of this stage is to describe the structure of the layout in terms of topological properties and also in terms of the function of each region. The first one is known as physical layout structure and the second one as logical layout structure.

There is no doubt that there is an improvement in this field but this one is insufficient to consider layout analysis domain as a solved problem. Complex layout structures, such as newspapers are still imperfectly analyzed [4, 8, 9] and still requires human interventions. The algorithms are generally applied on scanned images which are often of poor quality; such images may have degradations such as noise, optical deformations and so on. Furthermore they are often skewed. In our case, we handle high quality images un-skewed and noise free generated from PDF files.

In the literature a variety of proposed algorithms for geometric layout analysis of document images are proposed such as morphology or smearing based approaches, projection profiles, texture-based analysis, analysis of the background structure, and others [18]. We notice that the first layout analysis methods were focused on simple document structures [11]. Some recent works show a great interest in complex layout analysis. Currently known approaches rely on document models [13] and interactive incremental learning which is one of the main goals of the CIDRE[1] project [10]. These document models are either set up by hand or generated automatically in a previous learning step that needs a lot of ground-truthed data.

In the next subsections we explain all the methods used in Xed. First we explain the layout analysis algorithms; then we present the matching between the extracted objects from the PDF document and the result of layout analysis. Finally we review the topological sorting algorithm with an example reading order restitution.

4.1 Layout analysis algorithms

In computer vision, the aim of image segmentation is to separate the given image into homogenous region. Each region is indicated by a meaningful property. In fact our segmentation algorithm contains the following steps: thread extraction, frame extraction, image text separation, text line extraction and line merging into blocks. Our segmentation algorithm is modeled as a set of tools that can be used separately.

A bottom-up approach based on connected components is used for image, thread and frames extraction. Connected components are also used for text line extraction after the use of RLSA. The line merging into blocks is done according to rules.

The input of our segmentation algorithm is a TIFF image generated from PDF file, whereas the output is an XML file. This XML describes the segmentation results concerning different components such as threads, images, frames, text lines extraction and line merging into blocks. A sample of the XML segmentation output is illustrated in figure 4. The corresponding DTD is presented in the appendix A.

```
<?xml version="1.0" encoding="UTF-8"?>
<segmentation image="IHT_15_09_2003.tif">
   <Threads>
    <Thread x="827" y="955" w="3054" h="3" />
    <Thread x="339" y="1462" w="3981" h="3" />
    <Thread x="3022" y="2076" w="627" h="4" />
    <Thread x="339" y="2501" w="627" h="3" />
  ….
  </Threads>
   <Images>
    <Image x="361" y="1054" w="419" h="318" />
    <Image x="1010" y="1566" w="1968" h="1429" />
  ….
  </Images>
   <Texts>
    <Text x="413" y="15" w="8" h="1" />
    <Text x="1087" y="17" w="1" h="1" />
    <Text x="597" y="19" w="26" h="1" />
    <Text x="1631" y="19" w="5" h="1" />
  ….
  </Texts>
   <Frames>
    <Frame x="1010" y="4114" w="627" h="1082" />
   </Frames>
  <Blocks>
    <Block x="1010" y="4114" w="627" h="1082" />
    <Block x="339" y="207" w="1" h="53" />
    <Block x="1751" y="412" w="911" h="97" />
    <Block x="3029" y="546" w="68" h="63" />
    <Block x="1911" y="973" w="606" h="38" />
  ….
  </Blocks>
</segmentation>
```

Figure 4. A sample of the XML segmentation output.

[1] CIDRE stands for Cooperative and Interactive Document Reverse Engineering and is supported by the Swiss National Fund for Scientific Research, code 2000-059356.99-1.

In the following section we review the matching of the extracted objects from the PDF document with the layout analysis results.

4.2 Matching Xed's extracted objects with layout analysis results

Before doing the matching of the extracted objects from the PDF document with the layout analysis results, we extract PDF objects. The goal of this extraction is to obtain the various document content objects such as text, graphics, image and other related information. This task is delegated to Xed and is described in detail in the following section.

After obtaining the extracted content and the output of the layout analysis, we match them in the manner shown in figure 5.

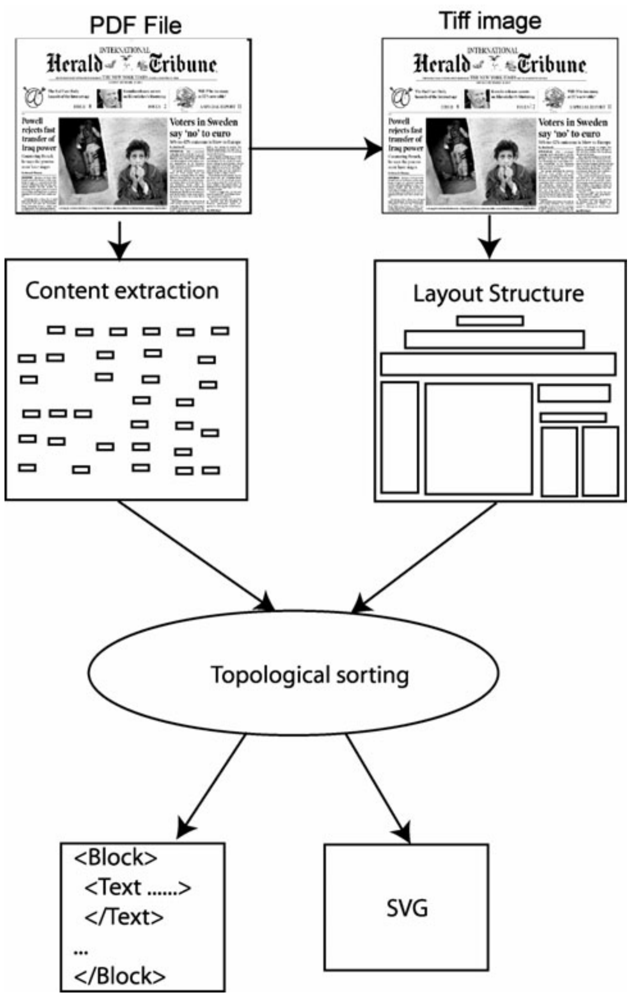


Figure 5. Topological sorting

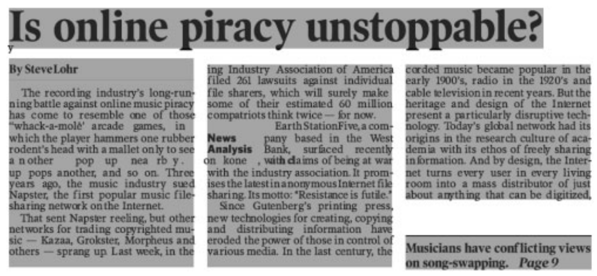
Many applications, such as restitution of the words, lines and blocks, can be built after structuring the content and matching it with the output of layout analysis algorithms. In this section we focus only on blocks.

The aim of the matching is to assign the set of text to each block. This resembles applying an OCR. The steps of the matching algorithm are the following:

1. For each extracted text element, compute its bounding box,
2. Check if it intersects with an existing text block,
3. If it intersects a block, add the text to this block; if not, the block is not correctly segmented and the corresponding text element is added to an array of non-resolved blocks.

The array of non-resolved blocks can be used to correct the missing blocks.

Figure 6 illustrates the result of the matching algorithm.



Is online piracy unstoppable?

y

By Steve Lohr

The recording industry's long-running battle against online music piracy has come to resemble one of those "whack-a-molé' arcade games, in which the player hammers one rubber rodent's head with a mallet only to see a n other pop up nea rb y . up pops another, and so on. Three years ago, the music industry sued Napster, the first popular music file-sharing network on the Internet.

That sent Napster reeling, but other networks for trading copyrighted music — Kazaa, Grokster, Morpheus and others — sprang up. Last week, in the

ing Industry Association of America filed 261 lawsuits against individual file sharers, which will surely make some of their estimated 60 million compatriots think twice — for now.

EarthStationFive, a company based in the West Bank, surfaced recently on kone , with claims of being at war with the industry association. It promises the latest in anonymous Internet file sharing. Its motto: "Resistance is futile."

News Analysis

Since Gutenberg's printing press, new technologies for creating, copying and distributing information have eroded the power of those in control of various media. In the last century, the

corded music became popular in the early 1900's, radio in the 1920's and cable television in recent years. But the heritage and design of the Internet present a particularly disruptive technology. Today's global network had its origins in the research culture of academia with its ethos of freely sharing information. And by design, the Internet turns every user in every living room into a mass distributor of just about anything that can be digitized.

Musicians have conflicting views on song-swapping. ***Page 9***

Figure 6. The result of the matching algorithm

In the next section we describe the algorithm for restituting the reading order.

4.3 The algorithm for restituting the reading order

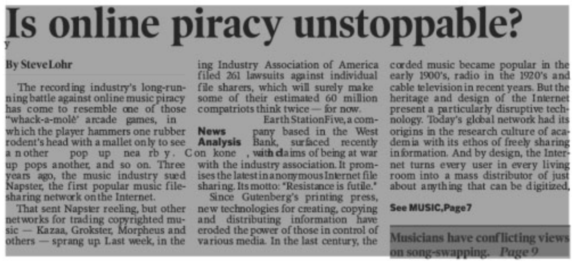
After the matching of the blocks with the n corresponding set of text, we reestablish the reading order. In fact, we merge the blocks belonging to the same article. This will help ordering the blocks correctly and fully obtain the logical zones, such as articles in the newspaper front page. Then, we sort them according to their spatial coordinates and their relationships.

The merging algorithm is composed of 3 steps:

1. Compute the mean of the width of all the blocks. The resulting value corresponds approximately to the width of columns,

2. Extract the blocks exceeding the mean width (in most of the case, they correspond to titles or subtitles),

3. For each of this block: a) fix the threshold to the height of the block; b) seek in the south direction the blocks that are nearer than the threshold and merge them (see figure 7).


Is online piracy unstoppable?

By Steve Lohr

The recording industry's long-running battle against online music piracy has come to resemble one of those "whack-a-mole" arcade games, in which the player hammers one rubber rodent's head with a mallet only to see another pop up nearby. Con kone , with claims of being at war up pops another, and so on. Three years ago, the music industry sued Napster, the first popular music file-sharing network on the Internet.

That sent Napster reeling, but other networks for trading copyrighted music — Kazaa, Grokster, Morpheus and others — sprang up. Last week, in the ing Industry Association of America filed 261 lawsuits against individual file sharers, which will surely make some of their estimated 60 million compatriots think twice — for now.

News Analysis

EarthStationFive, a company based in the West Bank, surfaced recently with the industry association. It promises the latest in anonymous Internet file sharing. Its motto: "Resistance is futile."

Since Gutenberg's printing press, new technologies for creating, copying and distributing information have eroded the power of those in control of various media. In the last century, the corded music became popular in the early 1900's, radio in the 1920's and cable television in recent years. But the heritage and design of the Internet present a particularly disruptive technology. Today's global network had its origins in the research culture of academia with its ethos of freely sharing information. And by design, the Internet turns every user in every living room into a mass distributor of just about anything that can be digitized.

See MUSIC, Page 7

Musicians have conflicting views on song-swapping. *Page 9*

Figure 7. Merging blocks into articles

Finally, we sort the merged blocks to obtain the reading order. First we sort all the blocks with the x criteria. Second we resort all the blocks belonging to the same column with the y criteria. The reference value of the axis x and y is the centre of gravity of the block.

We finally obtain the reading order as illustrated in figure 8.

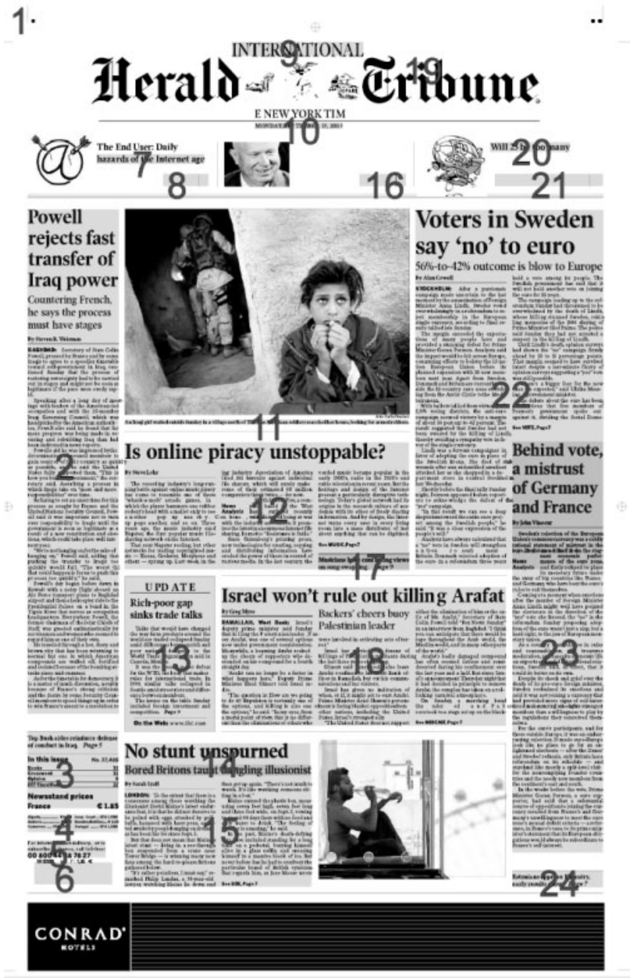

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
INTERNATIONAL Herald Tribune
Powell rejects fast transfer of Iraq power
Voters in Sweden say 'no' to euro
Is online piracy unstoppable?
Behind vote, a mistrust of Germany and France
UPDATE
Rich-poor gap sinks trade talks
Israel won't rule out killing Arafat
No stunt unspurned
CONRAD HOTELS

Figure 8. The restitution of the reading order of IHT frontpage

In the next section we review the architecture of Xed.

5. Architecture of Xed

After the discussion of the role of Xed in the algorithm for restituting the reading order, this section presents how the application operates to extract data from the PDF document.

We propose an architecture that is composed of three different phases: the reading, the extraction and the analysis phase (see figure 9).

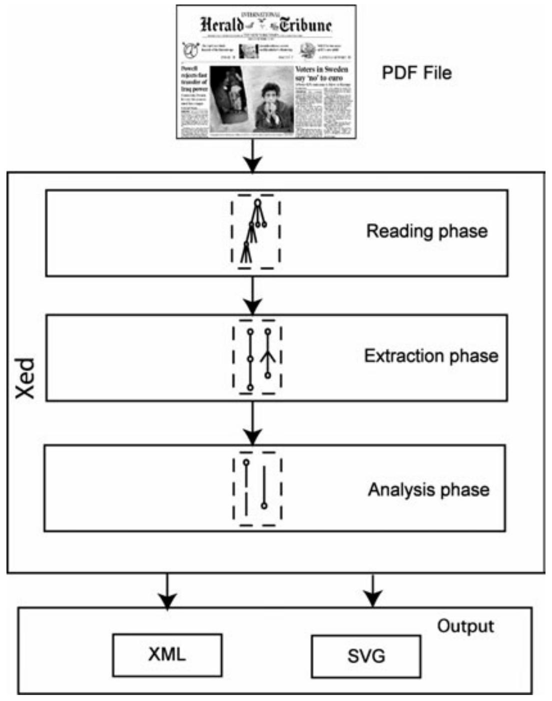


Figure 9. The phases of Xed

The reading phase is the low-level phase. It reads the PDF file and it creates an equivalent tree. The aim of this phase is to prepare the PDF data for the extraction. Every object in the PDF file is represented in Xed with an equivalent object, which contains exactly the same information as the original. The main difference between the original object and its double is the representation. In fact, most of the PDF objects can be represented in different manners: for example, a block of text can be described either in a binary stream or in an encoded binary stream. Similarly, a string has two different formats: literal or hexadecimal. The reading phase homogenizes the different representations of the PDF data: Xed allows only a possible format for a specific typology of object. As stated in the previous examples, Xed decodes the encoded blocks and it represents strings as characters.

The extraction phase follows the reading; it reconstructs and restitutes the entire information embedded in the PDF document. In order to catch up this objective, the extraction phase creates an instance of the PDF document. It works such as a PDF viewer but it reproduces the document in an inner representation instead of rendering it on the screen. In the rest of this section, this inner representation is called *abstract document*.

The abstract document is created page after page. First, Xed loads the resources used to compose the page (color spaces, images, fonts and extended graphical states) and then it constructs the page. A page is composed of three kinds of primitives: text, images and graphics (for example, lines or paths). During this process, Xed maintains a graphical state, which allows to determine the current drawing position, the stroke style, the color and other graphical properties of any object [1, 2, 3]. The abstract document is essential because every object depends on the graphical state at any moment. The object itself does not contain all the information describing its properties. For example, the rendering of a path depends either on the shape or on the appearance. This latter depends both of the color and the stroke, which are only defined in the current graphical state. The aim of the abstract document is to enrich and complete the information explicitly contained in the original PDF document.

The figure 10 resumes the construction of the abstract document.

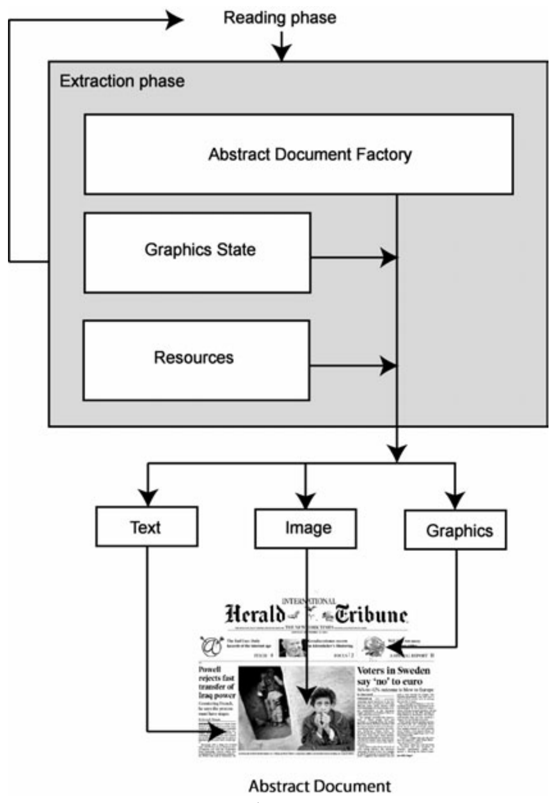


Figure 10. Abstract Document

The last phase, which follows the extraction phase, consists in analyzing the abstract document in order to transform it according to the applications needs. This is the case of the method for rebuilding the reading order described in the previous section.

Finally, Xed transforms the information collected in the abstract document in a XML language. Xed is implemented in Java and uses two different libraries, which encapsulates respectively the reading and the extraction phase. The analysis phase is embedded inside the main application. The existing prototype can output a result either in SVG language (Scalable Vector Graphics) [20] or in a specific XML language. The corresponding DTD is presented in the appendix B.

5.1 Using Xed

In order to illustrate the capabilities of Xed, this sub-section presents the results of the information extraction from a PDF file. The considered document is the front page of the French newspaper "Le Monde", published the 1st April 2003 and shown in figure 11.

Le Monde

www.lemonde.fr

MARDI 1er AVRIL 2003

SUPPLÉMENT

ÉCONOMIE

La France a-t-elle encore les moyens de rebondir ?

EMPLOI

Les Français et la guerre en Irak

CUBA
77 opposants arrêtés, dont le journaliste Raul Rivero (photo) p. 12

PNEUMONIE ATYPIQUE
Plus de 50 morts recensés par l'OMS p. 16

RUSSIE
Naissance de « Russie unie », le parti du pouvoir p. 13

MAIRIE DE PARIS
Un proche de Chirac mis en examen p. 15

CGT
Bernard Thibault au « Grand Jury RTL-Le Monde-LCI » p. 25 et notre éditorial p. 22

PARIS
Les artistes réclament le droit au squat p. 17

► Un sondage Ipsos pour TF1 et « Le Monde »

► 78 % des Français désapprouvent la guerre américaine

► Un tiers seulement se disent du côté des Américains et des Britanniques

► 74 % approuvent l'opposition de Chirac à l'action de Bush

► Reportages : Bassora, Bagdad, Kurdistan

► Rumsfeld répond aux critiques des militaires

Justice : les surprises du procès Elf

► Les premiers jours ont déstabilisé les principaux prévenus

► Débuts d'aveux à l'audience

► Portrait du président, Michel Desplan

SPORTS

Grand chelem pour l'Angleterre en attendant la Coupe du monde

Figure 11. The original PDF document

In this example, Xed extracts primitive objects (i.e. text, graphics and images). The figures 12-14 show an example of the output XML language.

```
<text
      x="266"
      y="214"
      length="167"
      rotate="0"
      fontname="L-SemiBold"
      fontsize="54"
      letterspacing="0.0"
      fill="rgb(0.0,0.0,0.0)"
      stroke="none" >
            Fran&#x00E7;ais
</text>
```

Figure 12. Text element

```
<image
      filename="images/img1.png"
      x="225.5"
      y="22.59440000000002"
      width="450.0985"
      height="102.0794"
      rotate="0" />
```

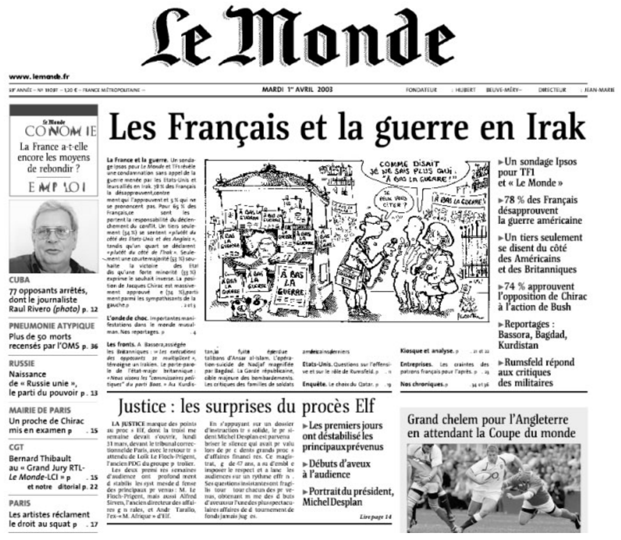
Figure 13. Image element

```
<path>
   <subpath closepath="true">
      <moveto x="43.0"
              y="135.0" />
      <hline  endx="816.0" />
      <vline  endy="1.0" />
      <hline  endx="-816.0" />
   </subpath>

   <style stroke="none"
          fill="rgb(0.0,0.0,0.0)"
          fillrule="nonzero"
          strokewidth="1.0"
          strokelinecap="butt"
          strokelinejoin="miter"
          strokemiterlimit="10.0" />
<path/>
```

Figure 14. Path element

The output of Xed can be observed and validated with xmillum, a framework for cooperative and interactive analysis of document, that allows to visualize and to edit document recognition results expressed in any XML language [12, 19].
In this example, xmillum loads the data extracted from the PDF document and then an XSLT stylesheet is applied. In xmillum, the transformed XML data associates a graphic representation to each element (text, path and image). Figure 15 shows a screenshot of the extracted data from PDF and visualized with xmillum.


Le Monde

www.lemonde.fr

59e ANNÉE – Nº 18097 – 1,20 € – FRANCE MÉTROPOLITAINE – MARDI 1er AVRIL 2003 FONDATEUR : HUBERT BEUVE-MÉRY– DIRECTEUR : JEAN-MARIE

Le Monde CONOM IE La France a-t-elle encore les moyens de rebondir ? EMP 1OI

Les Français et la guerre en Irak

La France et la guerre. Un sondage Ipsos pour Le Monde et TF1 révèle une condamnation sans appel de la guerre menée par les Etats-Unis et leurs alliés en Irak. 78 % des Français la désapprouvent,contre ment qui l'approuvent et 5 % qui ne se prononcent pas. Pour 65 % des Français,ce sont les portent la responsabilité du déclenchement du conflit. Un tiers seulement (34 %) se sentent « plutôt du côté des Etats-Unis et des Anglais », tandis qu'un quart se déclarent « plutôt du côté de l'Irak ». Seulement une courte majorité (53 %) souhaite la victoire des Etat dis qu'une forte minorité (33 %) exprime le souhait inverse. La position de Jacques Chirac est massivement approuvé e (74 %),parti ment parmi les sympathisants de la gauche.p . 2 et 3

L'onde de choc. Importantes manifestations dans le monde musulman. Nos reportages. p . 4

Les fronts. A Bassora,assiégée les Britanniques : « Les exécutions des opposants se multiplient », témoigne un Irakien. Le porte-parole de l'état-major britannique : « Nous visons les "commissaires politiques" du parti Baas. » Au Kurdistan,la fuite éperdue talibans d'Ansar al-Islam. L'opération-suicide de Nadjaf magnifiée par Bagdad. La Garde républicaine, cible majeure des bombardements. Les critiques des familles de soldats américainsderniers

Etats-Unis. Questions sur l'offensive et sur le rôle de Rumsfeld. p . 11

Enquête. Le choix du Qatar. p . 19

Kiosque et analyse. p . 21 et 22

Entreprises. Les craintes des patrons français pour l'après. p . 25

Nos chroniques. p . 34 et 36

►Un sondage Ipsos pour TF1 et « Le Monde »

►78 % des Français désapprouvent la guerre américaine

►Un tiers seulement se disent du côté des Américains et des Britanniques

►74 % approuvent l'opposition de Chirac à l'action de Bush

►Reportages : Bassora, Bagdad, Kurdistan

►Rumsfeld répond aux critiques des militaires

CUBA
77 opposants arrêtés, dont le journaliste Raul Rivero (photo) p. 12

PNEUMONIE ATYPIQUE
Plus de 50 morts recensés par l'OMS p. 36

RUSSIE
Naissance de « Russie unie », le parti du pouvoir p . 13

MAIRIE DE PARIS
Un proche de Chirac mis en examen p . 15

CGT
Bernard Thibault au « Grand Jury RTL-Le Monde-LCI » p . 15
et notre ditorial p. 22

PARIS
Les artistes réclament le droit au squat p . 17

Justice : les surprises du procès Elf

LA JUSTICE marque des points au proc s Elf, dont la troisi me semaine devait s'ouvrir, lundi 31 mars, devant le tribunal correctionnel de Paris, avec le retour tr s attendu de Loïk Le Floch-Prigent, l'ancien PDG du groupe p trolier.
Les deux premi res semaines d'audience ont profond ment d stabilis les syst mes de d fense des principaux pr venus : M. Le Floch-Prigent, mais aussi Alfred Sirven, l'ancien directeur des affaires g n rales, et Andr Tarallo, l'ex-« M. Afrique » d'Elf.

En s'appuyant sur un dossier d'instruction tr s solide, le pr sident Michel Desplan est parvenu briser le silence qui avait pr valu lors de pr c dents grands proc s d'affaires financi res. Ce magistrat, g de 47 ans, a su d'embl e imposer le respect et a lanc les audiences sur un rythme effr n . Ses questions insistantes ont fragilis tour tour chacun des pr venus, obtenant m me des d buts d'aveux sur l'une des plus spectaculaires affaires de d tournement de fonds jamais jug es.

►Les premiers jours ont déstabilisé les principaux prévenus

►Débuts d'aveux à l'audience

►Portrait du président, Michel Desplan

Lire page 14

Grand chelem pour l'Angleterre en attendant la Coupe du monde

Figure 15. Xed output shown in xmillum.

6. Conclusion

Extracting high-level document structures from PDF files is a very challenging problem with many potential practical applications. But, such an operation must rely on reliable low-level extraction tools. In our opinion, this problem has been largely underestimated so far.

In this paper we have described a novel approach: Xed for improving this essential step. The originality of our method consists in combing PDF symbol analysis with traditional document image processing techniques. By considering complex layout analysis results as input to drive a topological sorting algorithm for text pieces, we have brought a significant contribution to improve the reliability of low level text extraction from PDF files. Currently we are extracting words and lines from the bounding boxes of the characters, substrings and words. Our future work on Xed will focus on the improvement of the functionality of Xed in order to fix some problems, especially with words reconstitution. After that we will focus in details on the logical structure of the extracted document.

7. References

[1] Adobe Developer Support, The Compact Font Format Specification, Technical Note #5176, Version 1.0, 16 March 2000, http://partenrs.adobe.com

[2] Adobe Developer Support, The Type 2 Charstring Format, Technical Note #5177, 16 March 2000, http://partenrs.adobe.com

[3] Adobe PDF reference,
http://partners.adobe.com/asn/tech/pdf/specifications.jsp

[4] A. Antanacopoulos, B. Gatos and D. Karatzas, "ICDAR 2003 Page Segmentation Competition", *ICDAR2003,* Edinburgh (Scotland), August 2003, pp. 688-692.

[5] BCL, http://www.bcltechnologies.com/document/index.asp

[6] R. Cattoni, T. Coianiz, S. Messelodi and C.M. Modena, "Geometric layout analysis techniques for document image understanding a review", *Technical report, IRST*, Trento, Italy, 1998.

[7] Glance, http://www.pdf-tools.com/en/home.asp

[8] K. Hadjar, O. Hitz and R. Ingold, "Newspaper Page Decomposition using a Split and Merge Approach", *ICDAR'01*, Seattle (USA), September 2001, pp. 1186-1189.

[9] K. Hadjar and R. Ingold, "Arabic Newspaper Page Segmentation", *ICDAR'03*, Edinburgh (Scotland), August 2003, pp. 895-899.

[10] K. Hadjar, O. Hitz, L. Robadey and R. Ingold, "Configuration REecognition Model for Complex Reverse Engineering Methods: 2(CREM)", *DAS'02*, Princeton, NJ (USA), August 2002, pp. 469-479.

[11] R.M. Haralick, "Document image understanding: Geometric and logical layout", *Proc. Internet. Conf. On Computer Vision and Pattern Recognition*, 1994, pp. 385-390.

[12] O. Hitz, L. Robadey and R. Ingold, "An architecture for editing documents recognition results using xml technology", *DAS'2000*, Rio de Janeiro (Brazil), December 2000, pp. 385-396.

[13] J. Hu, R. Kashi, D. Lopresti, G. Nagy and G. Wilfong, "Why table ground truthing is hard", *ICDAR'01*, Seattle (USA), September 2001, pp. 129-133.

[14] JPEDAL, http://www.jpedal.org

[15] W.S. Lovegrove and D.F. Brailsford, "Document analysis of PDF files: methods, results and implications", *Electronic Publishing, Vol. 8(2 & 3)*, June & September 1995, pp. 207-220.

[16] MatterCast, http://www.mattercast.com/default.aspx

[17] B. Mohit, "Meta Data Extraction from PDF Research Papers",
http://www.sims.berkeley.edu/~behrangm/cs294/final/finalWriteup.pdf.

[18] G. Nagy, "Twenty Years of Document Image Analysis in PAMI", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No 1: January 2000, pp. 38-62.

[19] M. Rigamonti, O. Hitz and R. Ingold, "A Framework for Cooperative and Interactive Analysis of Technical Documents", *GREC 2003*, Barcelona (Spain), July 2003, pp. 407-414.

[20] SVG: Scalable Vector Graphics,
http://www.w3.org/TR/SVG/

Appendix A

DTD of the layout analysis

```
<!ELEMENT segmentation (Threads, Images, Texts, Frames, Blocks)>
<!ATTLIST segmentation
      image             CDATA #REQUIRED
>

<!ELEMENT Threads (Thread+)>
<!ELEMENT Thread EMPTY>
<!ATTLIST Thread
      x                 CDATA #REQUIRED
      y                 CDATA #REQUIRED
      w                 CDATA #REQUIRED
      h                 CDATA #REQUIRED
>

<!ELEMENT Images (Image+)>
<!ELEMENT Image EMPTY>
<!ATTLIST Image
      x                 CDATA #REQUIRED
      y                 CDATA #REQUIRED
      w                 CDATA #REQUIRED
      h                 CDATA #REQUIRED
>

<!ELEMENT Texts (Text+)>
<!ELEMENT Text EMPTY>
<!ATTLIST Text
      x                 CDATA #REQUIRED
      y                 CDATA #REQUIRED
      w                 CDATA #REQUIRED
      h                 CDATA #REQUIRED
>

<!ELEMENT Frames (Frame+)>
<!ELEMENT Frame EMPTY>
<!ATTLIST Frame
      x                 CDATA #REQUIRED
      y                 CDATA #REQUIRED
      w                 CDATA #REQUIRED
      h                 CDATA #REQUIRED
>

<!ELEMENT Blocks (Block+)>
<!ELEMENT Block EMPTY>
<!ATTLIST Block
      x                 CDATA #REQUIRED
      y                 CDATA #REQUIRED
      w                 CDATA #REQUIRED
      h                 CDATA #REQUIRED
>
```

Appendix B

DTD of Xed primitives

```
<!ELEMENT XedDoc (text | image | path | font | clippingpath | mask)+>

<!ELEMENT text (#PCDATA)>
<!ATTLIST text
      x                  CDATA #REQUIRED
      y                  CDATA #REQUIRED
      length             CDATA #REQUIRED
      rotate             CDATA #IMPLIED
      fontname           CDATA #REQUIRED
      fontsize           CDATA #REQUIRED
      letterspacing      CDATA #IMPLIED
      fill               CDATA #IMPLIED
      stroke             CDATA #IMPLIED
>

<!ELEMENT image EMPTY>
<!ATTLIST image
      filename           CDATA #REQUIRED
      x                  CDATA #REQUIRED
      y                  CDATA #REQUIRED
      width              CDATA #REQUIRED
      height             CDATA #REQUIRED
      rotate             CDATA #REQUIRED
>
<!ELEMENT path (subpath+, style)>
<!ELEMENT  subpath(moveto,  (moveto  |  lineto  |  hlineto  |  vlineto  |  cubicto)+,
closepath)>
<!ELEMENT moveto EMPTY>
<!ATTLIST moveto
      x                  CDATA #REQUIRED
      y                  CDATA #REQUIRED
>
<!ELEMENT lineto EMPTY>
<!ATTLIST lineto
      endx               CDATA #REQUIRED
      endy               CDATA #REQUIRED
>
<!ELEMENT hlineto EMPTY>
<!ATTLIST hlineto
      endx               CDATA #REQUIRED
>
<!ELEMENT vlineto EMPTY>
<!ATTLIST vlineto
      endy               CDATA #REQUIRED
>
<!ELEMENT cubicto EMPTY>
<!ATTLIST cubicto
      crt1x              CDATA #REQUIRED
      crt1y              CDATA #REQUIRED
      crt2x              CDATA #REQUIRED
      crt2y              CDATA #REQUIRED
      endx               CDATA #REQUIRED
      endy               CDATA #REQUIRED
>
```

```
<!ELEMENT closepath (0 | 1)>

<!ELEMENT style EMPTY>
<!ATTLIST style
      stroke            CDATA #IMPLIED
      fill              CDATA #IMPLIED
      fillrule          CDATA #IMPLIED
      strokewidth       CDATA #IMPLIED
      strokelinecap     CDATA #IMPLIED
      strokelinejoin    CDATA #IMPLIED
      strokemiterlimit  CDATA #IMPLIED
>

<!ELEMENT font EMPTY>
<!ATTLIST font
      family            CDATA #REQUIRED
      weight            CDATA #IMPLIED
      style             CDATA #IMPLIED
>
<!ELEMENT clippingpath (path)>
<!ATTLIST clippingpath
      id                CDATA #REQUIRED
>

<!ELEMENT mask (image)>
<!ATTLIST mask
      id                CDATA #REQUIRED
>
```