## NAME

djvutoxml, djvuxmlparser - DjVuLibre XML Tools.

## SYNOPSIS

**djvutoxml [***options***]** *inputdjvufile* **[***outputxmlfile***]**
**djvuxmlparser [ -o** *djvufile* **]** *inputxmlfile*

## DESCRIPTION

The DjVuLibre XML Tools provide for editing the metadata, hyperlinks and hidden text associated with DjVu files. Unlike **djvused**(1) the DjVuLibre XML Tools rely on the XML technology and can take advantage of XML editors and verifiers.

## DJVUTOXML

Program **djvutoxml** creates a XML file *outputxmlfile* containing a reference to the original DjVu document *inputdjvufile* as well as tags describing the metadata, hyperlinks, and hidden text associated with the DjVu file.

The following options are supported:

**--page** *pagenum*
Select a page in a multi-page document. Without this option, **djvutoxml** outputs the XML corresponding to all pages of the document.

**--with-text**
Specifies the **HIDDENTEXT** element for each page should be included in the output. If specified without the **--with-anno** flag then the **--without-anno** is implied. If none of the **--with-text, --without-text, --with-anno,** or **--without-anno,** flags are specified, then the **--with-text** and **--with-anno** flags are implied.

**--without-text**
Specifies not to output the **HIDDENTEXT** element for each page. If specified without the **--without-anno** flag then the **--with-anno** flag is implied.

**--with-anno**
Specifies the area **MAP** element for each page should be included in the output. If specified without the **--with-text** flag then the **--without-text** flag is implied.

**--without-anno**

Specifies the area **MAP** element for each page should not be included in the output. If specified without the **--without-text** flag then the **--with-text** flag is implied.

## DJVUXMLPARSER

Files produced by **djvutoxml** can then be modified using either a text editor or a XML editor. Program **djvuxmlparser** parses the XML file *inputxmlfile* in order to modify the metadata of the corresponding DjVu file.

**-o** *djvufile*

In principle the target DjVu file is the file referenced by the *OBJECT* element of the XML file. This option provides the means to override the filename specified in the *OBJECT* element.

## DJVUXML DOCUMENT TYPE DEFINITION

The document type definition file (DTD)

**/usr/local/share/djvu/pubtext/DjVuXML-s.dtd**

defines the input and output of the DjVu XML tools.

The DjVuXML-s DTD is a simplification of the HTML DTD:

**http://www.w3c.org/TR/1998/REC-html40-19980424/sgml/dtd.html**

with a few new attributes added specific to DjVu. Each of the specified pages of a DjVu document are represented as **OBJECT** elements within the **BODY** element of the XML file. Each **OBJECT** element may contain multiple **PARAM** elements to specify attributes like page name, resolution, and gamma factor. Each **OBJECT** element may also contain one **HIDDENTTEXT** element to specify the hidden text (usually generated with an OCR engine) within the DjVu page. In addition each **OBJECT** element may reference a single area **MAP** element which contains multiple **AREA** elements to represent all the hyperlink and highlight areas within the DjVu document.

### PARAM Elements

Legal **PARAM** elements of a DjVu **OBJECT** include but are not limited to **PAGE** for specifying the

page-name, **GAMMA** for specifying the gamma correction factor (normally 2.2), and **DPI** for specifying the page resolution.

### HIDDENTEXT Elements

The **HIDDENTEXT** elements consists of nested elements of **PAGECOLUMNS, REGION, PARAGRAPH, LINE,** and **WORD.** The most deeply nested element specified, should specify the bounding coordinates of the element in top-down orientation. The body of the most deeply nested element should contain the text. Most DjVu documents use either **LINE** or **WORD** as the lowest level element, but any element is legal as the lowest level element. A white space is always added between **WORD** elements and a line feed is always added between **LINE** elements. Since languages such as Japanese do not use spaces between words, it is quite common for Asian OCR engines to use **WORD** as characters instead.

### MAP Elements

The body of the **MAP** elements consist of **AREA** elements. In addition to the attributes listed in

**http://www.w3.org/TR/1998/REC-html40-19980424/struct/objects.html#edef-AREA**,

the attributes **bordertype**, **bordercolor**, **border**, and **highlight** have been added to specify border type, border color, border width, and highlight colors respectively. Legal values for each of these attributes are listed in the DjVuXML-s DTD. In addition, the shape **oval** has been added to the legal list of shapes. An oval uses a rectangular bounding box.

## BUGS

Perhaps it would have been better to use CC2 style sheets with standard HTML elements instead of defining the **HIDDENTEXT** element.

## CREDITS

The DjVu XML tools and DTD were written by Bill C. Riemers <docbill@sourceforge.net> and Fred Crary.

## SEE ALSO

**djvu**(1), **djvused**(1), and **utf8**(7).