Introduction

1 At a glance
   1.1 Requirements .................................................. 4
   1.2 Similar libraries .............................................. 4
   1.3 In use .......................................................... 4

Index 51
**pikepdf** is a Python library allowing creation, manipulation and repair of PDFs. It provides a Pythonic wrapper around the C++ PDF content transformation library, QPDF.

Python + QPDF = “py” + “qpdf” = “pyqpdf”, which looks like a dyslexia test and is no fun to type. But say “pyqpdf” out loud, and it sounds like “pikepdf”.

Fig. 1: A northern pike, or *esox lucius*.\(^1\)

---

\(^1\) Public domain image.
CHAPTER 1

At a glance

pikepdf is a library intended for developers who want to create, manipulate, parse, repair, and abuse the PDF format. It supports reading and write PDFs, including creating from scratch. Thanks to QPDF, it supports linearizing PDFs and access to encrypted PDFs.

```python
# Rotate all pages in a file by 180 degrees
import pikepdf
my_pdf = pikepdf.Pdf.open('test.pdf')
for page in my_pdf.pages:
    page.Rotate = 180
my_pdf.save('test-rotated.pdf')
```

It is a low level library that requires knowledge of PDF internals and some familiarity with the PDF specification. It does not provide a user interface of its own.

pikepdf would help you build apps that do things like:

- **Copy pages** from one PDF into another
- **Split** and **merge** PDFs
- Extract content from a PDF such as text or images
- Replace content, such as replacing an image without altering the rest of the file
- Repair, reformat or **linearize** PDFs
- Change the size of pages and reposition content
- Optimize PDFs similar to Acrobat’s features by downsampling images, deduplicating
- Calculate how much to charge for a scanning project based on the materials scanned

---

4 PDF 32000-1:2008.
2 CC0 image.

Fig. 1: Pike fish are tough, hard-fighting, aggressive predators.
• Alter a PDF to meet a target specification such as PDF/A or PDF/X
• Add or modify PDF metadata
• Create well-formed but invalid PDFs for testing purposes

What it cannot do:
• Rasterize PDF pages for display (that is, produce an image that shows what a PDF page looks like at a particular resolution/zoom level) – use Ghostscript instead
• Convert from PDF to other similar paper capture formats like epub, XPS, DjVu, Postscript – use MuPDF or PyMuPDF
• Print to paper

If you only want to generate PDFs and not read or modify them, consider reportlab (a “write-only” PDF generator).

1.1 Requirements

pikepdf currently requires Python 3.5+. There are no plans to backport to 2.7 or older versions in the 3.x series.

1.2 Similar libraries

Unlike similar Python libraries such as PyPDF2 and pdfrw, pikepdf is not pure Python. Both were designed prior to Python wheels which has made Python extension libraries much easier to work with. By leveraging the existing mature code base of QPDF, despite being new, pikepdf is already more capable than both in many respects – for example, it can read compress object streams, repair damaged PDFs in many cases, and linearize PDFs. Unlike those libraries, it’s not pure Python: it is impure and proud of it.

1.3 In use

pikepdf is used by the same author’s OCRmyPDF to inspect input PDFs, graft the generated OCR layers on to page content, and output PDFs. Its code contains main practical examples, particular in pdfinfo.py, _weave.py, and optimize.py. pikepdf is also used in the test suite.

1.3.1 Installation

Basic installation

Most users on Linux, macOS or Windows with x64 systems should use pip to install pikepdf in their current Python environment (such as your project’s virtual environment).

3 CC-BY-SA 2.0 image.
This command installs binary wheels. 64-bit wheels are available for Windows, Linux and macOS. (Except Alpine Linux.)

32-bit wheels are available for Windows, for use with the 32-bit version of Python (regardless of the bitness of Windows). 32-bit wheels for Linux will be added if anyone uses them.

Binary wheels should work on most systems work on Linux distributions 2007 and newer, macOS 10.11 and newer (for Homebrew), Windows 7 and newer.

Use `pip install --user pikepdf` to install the package for the current user only. Use `pip install pikepdf` to install to a virtual environment.

### Platform support

Some platforms include versions of pikepdf that are distributed by the system package manager (such as `apt`). These versions may lag behind the version distributed with PyPI, but may be convenient for users that cannot use binary wheels.

#### Debian, Ubuntu and other APT-based distributions

`apt install pikepdf`

#### Fedora 29

`dnf install python-pikepdf`

#### ArchLinux

Available in ArchLinux User Repository.

`pacman -S pikepdf`

#### Installing on FreeBSD

`pkg install py36-pikepdf`

To attempt a manual install, try something like:

`pkg install python3 lang/python3
pkg install py36-lxml qpdf
pip install --user pikepdf`

This procedure is known to work on FreeBSD 11.2. It has not been tested on other versions.
Building from source

Requirements
pikepdf requires:

• a C++11 compliant compiler - GCC (4.8 and up) and clang (3.3 and up); C++14 is recommended and will produced smaller binaries
• pybind11
• libqpdf 8.4.2 or higher from the QPDF project.
• defusedxml - Python package

On Linux the library and headers for libqpdf must be installed because pikepdf compiles code against it and links to it.

Check Repology for QPDF to see if a recent version of QPDF is available for your platform. Otherwise you must build QPDF from source. (Consider using the binary wheels, which bundle the required version of libqpdf.)

Compiling with GCC or Clang

• clone this repository
• install libjpeg, zlib and libqpdf on your platform, including headers
• pip install .

Note: pikepdf should be built with the same compiler and linker as libqpdf; to be precise both must use the same C++ ABI. On some platforms, setup.py may not pick the correct compiler so one may need to set environment variables CC and CXX to redirect it. If the wrong compiler is selected, import pikepdf._qpdf will throw an ImportError about a missing symbol.

On Windows (requires Visual Studio 2015)

pikepdf requires a C++11 compliant compiler (i.e. Visual Studio 2015 on Windows). See our continuous integration build script in .appveyor.yml for detailed and current instructions. Or use the wheels which save this pain.

These instructions require the precompiled binary qpdf.dll. See the QPDF documentation if you also need to build this DLL from source. Both should be built with the same compiler. You may not mix and match MinGW and Visual C++ for example.

Running a regular pip install command will detect the version of the compiler used to build Python and attempt to build the extension with it. We must force the use of Visual Studio 2015.

1. Clone this repository.
2. In a command prompt, run:

```bash
%VS140COMNTOOLS%\..\..\VC\vcvarsall.bat" x64
set DISTUTILS_USE_SDK=1
set MSSdk=1
```
3. Download qpdf-8.4.2-bin-msvc64.zip from the QPDF releases page.
4. Extract bin\qpdfXX.dll from the zip file above, where XX is the version of the ABI, and copy it to the src/pikepdf folder in the repository.
5. Run pip install . in the root directory of the repository.
Note: The user compiling pikepdf to must have registry editing rights on the machine to be able to run the vcvarsall.bat script.

Note: If you are attempting to build pikepdf because you want to use OCRmyPDF, OCRmyPDF is not supported on Windows at this time.

Windows runtime requirements

On Windows, the Visual C++ 2015 redistributable packages are a runtime requirement for this project. It can be found here.

Building the documentation

Documentation is generated using Sphinx and you are currently reading it. To regenerate it:

```
pip install -r requirements/docs.txt
cd pikepdf/docs
make html
```

1.3.2 Release notes

pikepdf releases use the semantic versioning policy.

The pikepdf API (as provided by import pikepdf) is stable and is in production use. Note that the C++ extension module pikepdf._qpdf is a private interface within pikepdf that applications should not access directly, along with any modules with a prefixed underscore.

v1.6.5

- Fixed build settings to support Python 3.8.
- pybind11 2.4.3 is now required, to support Python 3.8.

v1.6.4

- When images were encoded with CCITTFaxDecode, type G4, with the /EncodedByteAlign set to true (not default), the image extracted by pikepdf would be a corrupted form of the original, usually appearing as a small speckling of black pixels at the top of the page. Saving an image with pikepdf was not affected; this problem only occurred when attempting to extract images. We now refuse to extract images with these parameters, as there is not sufficient documentation to determine how to extract them. This image format is relatively rare.
v1.6.3

- Fixed compatibility with libqpdf 9.0.0. - A new method introduced in libqpdf 9.0.0 overloaded an older method, making a reference to this method in pikepdf ambiguous.
  - A test relied on libqpdf raising an exception when a pikepdf user called `Pdf.save(..., min_version='invalid')`. libqpdf no longer raises an exception in this situation, but ignores the invalid version. In the interest of supporting both versions, we defer to libqpdf. The failing test is removed, and documentation updated.
- Several warnings, most specific to the Visual C++ compiler, were fixed.
- The Windows CI scripts were adjusted for the change in libqpdf ABI version.
- Wheels are now built against libqpdf 9.0.0.
- libqpdf 8.4.2 and 9.0.0 are both supported.

v1.6.2

- Fixed another build problem on Alpine Linux - musl-libc defines `struct FILE` as an incomplete type, which breaks pybind11 metaprogramming that attempts to reason about the type.
- Documentation improved to mention FreeBSD port.

v1.6.1

- Dropped our one usage of QPDF’s C API so that we use only C++.
- Documentation improvements.

v1.6.0

- Added bindings for QPDF’s page object helpers and token filters. These enable: filtering content streams, capturing pages as Form XObjects, more convenient manipulation of page boxes.
- Fixed a logic error on attempting to save a PDF created in memory in a way that overwrites an existing file.
- Fixed `Pdf.get_warnings()` failed with an exception when attempting to return a warning or exception.
- Improved manylinux1 binary wheels to compile all dependencies from source rather than using older versions.
- More tests and more coverage.
- libqpdf 8.4.2 is required.

v1.5.0

- Improved interpretation of images within PDFs that use an ICC colorspace. Where possible we embed the ICC profile when extracting the image, and profile access to the ICC profile.
- Fixed saving PDFs with their existing encryption.
- Fixed documentation to reflect the fact that saving a PDF without specifying encryption settings will remove encryption.
- Added a test to prevent overwriting the input PDF since overwriting corrupts lazy loading.
• Object.write(filters=, decode_parms=) now detects invalid parameters instead of writing invalid values to Filters and DecodeParms.
• We can now extract some images that had stacked compression, provided it is /FlateDecode.
• Add convenience function Object.wrap_in_array().

v1.4.0
• Added support for saving encrypted PDFs. (Reading them has been supported for a long time.)
• Added support for setting the PDF extension level as well as version.
• Added support converting strings to and from PDFDocEncoding, by registering a "pdfdoc" codec.

v1.3.1
• Updated pybind11 to v2.3.0, fixing a possible GIL deadlock when pikepdf objects were shared across threads. (#27)
• Fixed an issue where PDFs with valid XMP metadata but missing an element that is usually present would be rejected as malformed XMP.

v1.3.0
• Remove dependency on defusedxml.lxml. Unfortunately this module of defusedxml is deprecated because the issue it worked around have been resolved in the underlying lxml library. In the absence of other options for XML hardening we have reverted to standard lxml.
• Fixed an issue where PdfImage.extract_to() would write a file in the wrong directory.
• Eliminated an intermediate buffer that was used when saving to an IO stream (as opposed to a filename). We would previously write the entire output to a memory buffer and then write to the output buffer; we now write directly to the stream.
• Added Object.emplace() as a workaround for when one wants to update a page without generating a new page object so that links/table of contents entries to the original page are preserved.
• Improved documentation. Eliminated all arg0 placeholder variable names.
• Added PageList.remove(p=1), so that it is possible to remove pages using counting numbers.

v1.2.0
• Implemented Pdf.close() and with-block context manager, to allow Pdf objects to be closed without relying on del.
• PdfImage.extract_to() has a new keyword argument fileprefix=, which to specify a filepath where an image should be extracted with pikepdf setting the appropriate file suffix. This simplifies the API for the most common case of extracting images to files.
• Fixed an internal test that should have suppressed the extraction of JPEGs with a nonstandard ColorTransform parameter set. Without the proper color transform applied, the extracted JPEGs will typically look very pink. Now, these images should fail to extract as was intended.
• Fixed that Pdf.save(object_stream_mode=...) was ignored if the default fix_metadata_version=True was also set.

1.3. In use
• Data from one Pdf is now copied to other Pdf objects immediately, instead of creating a reference that required source PDFs to remain available. Pdf objects no longer reference each other.

• libqpdf 8.4.0 is now required

• Various documentation improvements

v1.1.0

• Added workaround for macOS/clang build problem of the wrong exception type being thrown in some cases.

• Improved translation of certain system errors to their Python equivalents.

• Fixed issues resulting from platform differences in datetime.strftime(#25)

• Added Pdf.new, Pdf.add_blank_page and Pdf.make_stream convenience methods for creating new PDFs from scratch.

• Added binding for new QPDF JSON feature: Object.to_json.

• We now automatically update the XMP PDFVersion metadata field to be consistent with the PDF’s declared version, if the field is present.

• Made our Python-augmented C++ classes easier for Python code inspectors to understand.

• Eliminated use of the imghdr library.

• Autoformatted Python code with black.

• Fixed handling of XMP metadata that omits the standard <x:xmpmeta> wrapper.

v1.0.5

• Fixed an issue where an invalid date in XMP metadata would cause an exception when updating DocumentInfo. For now, we warn that some DocumentInfo is not convertible. (In the future, we should also check if the XMP date is valid, because it probably is not.)

• Rebuilt the binary wheels with libqpdf 8.3.0. libqpdf 8.2.1 is still supported.

v1.0.4

• Updates to tests/resources (provenance of one test file, replaced another test file with a synthetic one)

v1.0.3

• Fixed regression on negative indexing of pages.

v1.0.2

• Fixed an issue where invalid values such as out of range years (e.g. 0) in DocumentInfo would raise exceptions when using DocumentInfo to populate XMP metadata with .load_from_docinfo.

v1.0.1

• Fixed an exception with handling metadata that contains the invalid XML entity &#0; (an escaped NUL)
v1.0.0

• Changed version to 1.0.

v0.10.2

Fixes

• Fixed segfault when overwriting the pikepdf file that is currently open on Linux.
• Fixed removal of an attribute metadata value when values were present on the same node.

v0.10.1

Fixes

• Avoid canonical XML since it is apparently too strict for XMP.

v0.10.0

Fixes

• Fixed several issues related to generating XMP metadata that passed veraPDF validation.
• Fixed a random test suite failure for very large negative integers.
• The lxml library is now required.

v0.9.2

Fixes

• Added all of the commonly used XML namespaces to XMP metadata handling, so we are less likely to name something ‘ns1’, etc.
• Skip a test that fails on Windows.
• Fixed build errors in documentation.

v0.9.1

Fixes

• Fix `Object.write()` accepting positional arguments it wouldn’t use
• Fix handling of XMP data with timezones (or missing timezone information) in a few cases
• Fix generation of XMP with invalid XML characters if the invalid characters were inside a non-scalar object
v0.9.0

Updates

- New API to access and edit PDF metadata and make consistent edits to the new and old style of PDF metadata.
- 32-bit binary wheels are now available for Windows
- PDFs can now be saved in QPDF’s “qdf” mode
- The Python package defusedxml is now required
- The Python package python-xmp-toolkit and its dependency libexempi are suggested for testing, but not required

Fixes

- Fixed handling of filenames that contain multibyte characters on non-UTF-8 systems

Breaking

- The `Pdf.metadata` property was removed, and replaced with the new metadata API
- `Pdf.attach()` has been removed, because the interface as implemented had no way to deal with existing attachments.

v0.3.7

- Add API for inline images to unparshe themselves

v0.3.6

- Performance of reading files from memory improved to avoid unnecessary copies.
- It is finally possible to use `for key in pdfobj` to iterate contents of PDF Dictionary, Stream and Array objects. Generally these objects behave more like Python containers should now.
- Package API declared beta.

v0.3.5

Breaking

- `Pdf.save(...stream_data_mode=...)` has been dropped in favor of the newer `compress_streams=` and `stream_decode_level` parameters.

Fixes

- A use-after-free memory error that caused occasional segfaults and “QPDFFakeName” errors when opening from stream objects has been resolved.
v0.3.4

Updates

- pybind11 vendoring has ended now that v2.2.4 has been released

v0.3.3

Breaking

- libqpdf 8.2.1 is now required

Updates

- Improved support for working with JPEG2000 images in PDFs
- Added progress callback for saving files, `Pdf.save(..., progress=)`
- Updated pybind11 subtree

Fixes

- `del obj.AttributeName` was not implemented. The attribute interface is now consistent
- Deleting named attributes now defers to the attribute dictionary for Stream objects, as get/set do
- Fixed handling of JPEG2000 images where metadata must be retrieved from the file

v0.3.2

Updates

- Added support for direct image extraction of CMYK and grayscale JPEGs, where previously only RGB (internally YUV) was supported
- `Array()` now creates an empty array properly
- The syntax `Name.Foo in Dictionary()`, e.g. `Name.XObject in page.Resources`, now works

v0.3.1

Breaking

- `pikepdf.open` now validates its keyword arguments properly, potentially breaking code that passed invalid arguments
- libqpdf 8.1.0 is now required - libqpdf 8.1.0 API is now used for creating Unicode strings
- If a non-existent file is opened with `pikepdf.open`, `a FileNotFoundError` is raised instead of a generic error
- We are now temporarily vendoring a copy of pybind11 since its master branch contains unreleased and important fixes for Python 3.7.
Updates

- The syntax `Name.Thing` (e.g. `Name.DecodeParms`) is now supported as equivalent to `Name('/Thing')` and is the recommended way to refer names within a PDF
- New API `Pdf.remove_unneeded_resources()` which removes objects from each page’s resource dictionary that are not used in the page. This can be used to create smaller files.

Fixes

- Fixed an error parsing inline images that have masks
- Fixed several instances of catching C++ exceptions by value instead of by reference

v0.3.0

Breaking

- Modified `Object.write` method signature to require `filter` and `decode_parms` as keyword arguments
- Implement automatic type conversion from the PDF Null type to `None`
- Removed `Object.unparse_resolved` in favor of `Object.unparse(resolved=True)`
- libqpdf 8.0.2 is now required at minimum

Updates

- Improved IPython/Jupyter interface to directly export temporary PDFs
- Updated to qpdf 8.1.0 in wheels
- Added Python 3.7 support for Windows
- Added a number of missing options from QPDF to `Pdf.open` and `Pdf.save`
- Added ability to delete a slice of pages
- Began using Jupyter notebooks for documentation

v0.2.2

- Added Python 3.7 support to build and test (not yet available for Windows, due to lack of availability on Appveyor)
- Removed setter API from `PdfImage` because it never worked anyway
- Improved handling of `PdfImage` with trivial palettes

v0.2.1

- `Object.check_owner` renamed to `Object.is_owned_by`
- `Object.objgen` and `Object.get_object_id` are now public functions
• Major internal reorganization with `pikepdf.models` becoming the submodule that holds support code to ease access to PDF objects as opposed to wrapping QPDF.

v0.2.0

• Implemented automatic type conversion for `int`, `bool` and `Decimal`, eliminating the `pikepdf.{Integer,Boolean,Real}` types. Removed a lot of associated numerical code.

Everything before v0.2.0 can be considered too old to document.

1.3.3 Tutorial

This brief tutorial should give you an introduction and orientation to pikepdf’s paradigm and syntax. From there, we refer to you various topics.

Opening and saving PDFs

In contrast to better known PDF libraries, pikepdf uses a single object to represent a PDF, whether reading, writing or merging. We have cleverly named this `pikepdf.Pdf`. In this documentation, a `Pdf` is a class that allows manipulate the PDF, meaning the file.

```python
from pikepdf import Pdf
new_pdf = Pdf.new()
with Pdf.open('sample.pdf') as pdf:
    pdf.save('output.pdf')
```

You may of course use `from pikepdf import Pdf as ...` if the short class name conflicts or `from pikepdf import Pdf as PDF` if you prefer uppercase.

`pikepdf.open()` is a shorthand for `pikepdf.Pdf.open()`.

The PDF class API follows the example of the widely-used Pillow image library. For clarity there is no default constructor since the arguments used for creation and opening are different. `Pdf.open()` also accepts seekable streams as input, and `Pdf.save()` accepts streams as output.

Inspecting pages

Manipulating pages is fundamental to PDFs. pikepdf presents the pages in a PDF through the `pikepdf.Pdf.pages` property, which follows the `list` protocol. As such page numbers begin at 0.

Let’s open a simple PDF that contains four pages.

```python
In [1]: from pikepdf import Pdf
In [2]: pdf = Pdf.open('./tests/resources/fourpages.pdf')
```

How many pages?

```python
In [3]: len(pdf.pages)
Out[3]: 4
```

pikepdf integrates with IPython and Jupyter’s rich object APIs so that you can view PDFs, PDF pages, or images within PDF in a IPython window or Jupyter notebook. This makes it to test visual changes.
You can also examine individual pages, which we’ll explore in the next section. Suffice to say that you can access pages by indexing them and slicing them.

Note: `pikepdf.Pdf.open()` can open almost all types of encrypted PDF! Just provide the `password=` keyword argument.

For more details on document assembly, see *PDF split, merge and document assembly*.

### Pages are dictionaries

In PDFs, the main data structure is the dictionary, a key-value data structure much like a Python `dict` or `attrdict`. The major difference is that the keys can only be names, and can only be PDF types, including other dictionaries.

PDF dictionaries are represented as `pikepdf.Dictionary`, and names are of type `pikepdf.Name`. A page is just a dictionary with a few required files and a reference from the document’s “page tree”. (pikepdf manages the page tree for you.)

```python
In [7]: from pikepdf import Pdf
In [8]: example = Pdf.open('..tests/resources/congress.pdf')
In [9]: page1 = example.pages[0]
```

**repr() output**

Let’s example the page’s `repr()` output:

```python
In [10]: page1
Out[10]:
<pikepdf.Dictionary(type_="/Page")({
    "/Contents": pikepdf.Stream(stream_dict={
        "/Length": 50
    }, data=<...>),
    "/MediaBox": [ 0, 0, 200, 304 ],
    "/Parent": <reference to /Pages>,
    "/Resources": {
        "/XObject": {
            "/Im0": pikepdf.Stream(stream_dict={
                "/BitsPerComponent": 8,
                "/ColorSpace": "/DeviceRGB",
                "/Filter": [ "/DCTDecode" ],
                "/Height": 1520,
                "/Length": 192956,
                "/Subtype": "/Image",
```
The angle brackets in the output indicate that this object cannot be constructed with a Python expression because it contains a reference. When angle brackets are omitted from the \texttt{repr()} of a pikepdf object, then the object can be replicated with a Python expression, such as \texttt{eval(repr(x))} == \texttt{x}. Pages typically concern indirect references to themselves and other pages, so they cannot be represented as an expression.

In Jupyter and IPython, pikepdf will instead attempt to display a preview of the PDF page, assuming a PDF rendering backend is available.

**Item and attribute notation**

Dictionary keys may be looked up using attributes (\texttt{page1.MediaBox}) or keys (\texttt{page1['/MediaBox']}).

```python
In [11]: page1.MediaBox  # preferred notation for required names
Out[11]: pikepdf.Array([ 0, 0, 200, 304 ])

In [12]: page1['/MediaBox']  # also works
Out[12]: pikepdf.Array([ 0, 0, 200, 304 ])
```

By convention, pikepdf uses attribute notation for standard names, and item notation for names that are set by PDF developers. For example, the images belong to a page always appear at \texttt{page.Resources.XObject} but the name of images is set by the PDF creator:

```python
In [13]: page1.Resources.XObject['/Im0']
```

Item notation here would be quite cumbersome: \texttt{['/Resources']['/XObject']['/Im0']} (not recommended).

Attribute notation is convenient, but not robust if elements are missing. For elements that are not always present, you can use \texttt{.get()}, which behaves like \texttt{dict.get()} in core Python. A library such as glom might help when working with complex structured data that is not always present.

(For now, we’ll set aside what a page’s \texttt{MediaBox} and \texttt{Resources.XObject} are for. See \textit{Working with pages} for details.)

**Deleting pages**

Removing pages is easy too.

```python
In [14]: del pdf.pages[1:3]  # Remove pages 2-3 labeled "second page" and "third page"

In [15]: len(pdf.pages)
Out[15]: 2
```
Saving changes

Naturally, you can save your changes with `pikepdf.Pdf.save()`. `filename` can be a `pathlib.Path`, which we accept everywhere. (Saving is commented out to avoid upsetting the documentation generator.)

```python
In [16]: pdf.save('output.pdf')
```

You may save a file multiple times, and you may continue modifying it after saving.

To save an encrypted (password protected) PDF, use a `pikepdf.Encryption` object to specify the encryption settings. By default, pikepdf selects the strongest security handler and algorithm (AES-256), but allows full access to modify file contents. A `pikepdf.Permissions` object can be used to specify restrictions.

```python
In [17]: no_extracting = pikepdf.Permissions(extract=False)
```

```python
In [18]: pdf.save('encrypted.pdf', encryption=pikepdf.Encryption(
   ....:     user="user password", owner="owner password", allow=no_extracting
   ....: ))
```

Next steps

Have a look at pikepdf topics that interest you, or jump to our detailed API reference…

### 1.3.4 PDF split, merge, and document assembly

This section discusses working with PDF pages: splitting, merging, copying, deleting. We’re treating pages as a unit, rather than working with the content of individual pages.

Let’s continue with `fourpages.pdf` from the Tutorial.

```python
In [1]: from pikepdf import Pdf
```

```python
In [2]: pdf = Pdf.open('..tests/resources/fourpages.pdf')
```

#### Split a PDF into one page PDFs

All we need is a new PDF to hold the destination page.

```python
In [3]: pdf = Pdf.open('..tests/resources/fourpages.pdf')
```

```python
In [4]: for n, page in enumerate(pdf.pages): 
   ....:     dst = Pdf.new()
   ....:     dst.pages.append(page)
   ....:     dst.save('{:02d}.pdf'.format(n))
   ....:
```

**Note:** This example will transfer data associated with each page, so that every page stands on its own. It will *not* transfer some metadata associated with the PDF as a whole, such the list of bookmarks.
Merge (concatenate) PDF from several PDFs

We create an empty `Pdf` which will be the container for all the others.

```python
In [5]: from glob import glob
In [6]: pdf = Pdf.new()
In [7]: for file in glob('*.pdf'):
    ...:     src = Pdf.open(file)
    ...:     pdf.pages.extend(src.pages)
In [8]: pdf.save('merged.pdf')
```

This code sample is enough to merge most PDFs, but there are some things it does not do that a more sophisticated function might do. One could call `pikepdf.Pdf.remove_unreferenced_resources()` to remove unref-erenced resources. It may also be necessary to chose the most recent version of all source PDFs. Here is a more sophisticated example:

```python
In [9]: from glob import glob
In [10]: pdf = Pdf.new()
In [11]: version = pdf.pdf_version
In [12]: for file in glob('*.pdf'):
    ...:     src = Pdf.open(file)
    ...:     version = max(version, src.pdf_version)
    ...:     pdf.pages.extend(src.pages)
In [13]: pdf.remove_unreferenced_resources()
In [14]: pdf.save('merged.pdf', min_version=version)
```

This improved example would still leave metadata blank. It’s up to you to decide how to combine metadata from multiple PDFs.

Reversing the order of pages

Suppose the file was scanned backwards. We can easily reverse it in place - maybe it was scanned backwards, a common problem with automatic document scanners.

```python
In [15]: pdf.pages.reverse()
```

```python
In [16]: pdf
Out[16]: <pikepdf.Pdf description='../../../tests/resources/fourpages.pdf'>
```

Pretty nice, isn’t it? But the pages in this file already were in correct order, so let’s put them back.

```python
In [17]: pdf.pages.reverse()
```
Copying pages from other PDFs

Now, let's add some content from another file. Because pdf.pages behaves like a list, we can use `pages.extend()` on another file's pages.

```python
In [18]: pdf = Pdf.open('..tests/resources/fourpages.pdf')
In [19]: appendix = Pdf.open('..tests/resources/sandwich.pdf')
In [20]: pdf.pages.extend(appendix.pages)
```

We can use `pages.insert()` to insert into one of more pages into a specific position, bumping everything else ahead.

Copying pages between Pdf objects will create a shallow copy of the source page within the target Pdf, rather than the typical Python behavior of creating a reference. As such, modifying pdf.pages[-1] will not affect appendix.pages[0]. (Normally, assigning objects between Python lists creates a reference, so that the two objects are identical, list[0] is list[1].)

```python
In [21]: graph = Pdf.open('..tests/resources/graph.pdf')
In [22]: pdf.pages.insert(1, graph.pages[0])
```

The method above will break any indirect references (such as table of contents entries and hyperlinks) within pdf to pdf.pages[2]. Perhaps that is the behavior you want, if the replacement means those references are no longer valid. This is shown by the change in pikepdf.Object.objgen.

Emplacing pages

To preserve indirect references, use `pikepdf.Object.emplace()`, which will (conceptually) delete all of the content of target and replace it with the content of source, thus preserving indirect references to the page.

```python
In [24]: congress = Pdf.open('..tests/resources/congress.pdf')
In [25]: pdf.pages[2].objgen
Out[25]: (4, 0)
In [26]: pdf.pages[2] = congress.pages[0]
In [27]: pdf.pages[2].objgen
Out[27]: (33, 0)
```
Copying pages within a PDF

As you may have guessed, we can assign pages to copy them within a Pdf:

```
In [32]: pdf = Pdf.open('..tests/resources/fourpages.pdf')
In [33]: pdf.pages[3] = pdf.pages[0]  # The last shall be made first
```

As above, copying a page creates a shallow copy rather than a Python object reference.

Also as above `pikepdf.Object.emplace()` can be used to create a copy that preserves the functionality of indirect references within the PDF.

Using counting numbers

Because PDF pages are usually numbered in counting numbers (1, 2, 3...), pikepdf provides a convenience accessor `.p()` that uses counting numbers:

```
In [34]: pdf.pages.p(1)  # The first page in the document
In [35]: pdf.pages[0]  # Also the first page in the document
In [36]: pdf.pages.remove(p=1)  # Remove first page in the document
```

To avoid confusion, the `.p()` accessor does not accept Python slices, and `.p(0)` raises an exception. It is also not possible to delete using it.

PDFs may define their own numbering scheme or different numberings for different sections, such as using Roman numerals for an introductory section. `.pages` does not look up this information.

Pages information from Root

**Warning:** It’s possible to obtain page information through `pikepdf.Pdf.Root` object but **not recommended**. (In PDF parlance, this is the `/Root` object).

The internal consistency of the various `/Page` and `/Pages` is not guaranteed when accessed in this manner, and in some PDFs the data structure for these is fairly complex. Use the `.pages` interface.

1.3.5 Working with pages

This section details with how to view and edit the contents of a page.

pikepdf is not an ideal tool for producing new PDFs from scratch – and there are many good tools for that, as mentioned elsewhere. pikepdf is better at inspecting, editing and transforming existing PDFs.

Page objects in PDFs are dictionaries.

```
In [1]: from pikepdf import Pdf, Page
In [2]: example = Pdf.open('..tests/resources/congress.pdf')
In [3]: pageobj1 = example.pages[0]
In [4]: pageobj1
```
The page’s /Contents key contains instructions for drawing the page content. This is a content stream, which is a stream object that follows special rules.

Also attached to this page is a /Resources dictionary, which contains a single XObject image. The image is compressed with the /DCTDecode filter, meaning it is encoded with the DCT (discrete cosine transform), so it is a JPEG. pikepdf has special APIs for working with images.

The /MediaBox describes the bounding box of the page in PDF pt units (1/72" or 0.35 mm).

You can access the page dictionary data structure directly, but it’s fairly complicated. There are a number of rules, optional values and implied values. It’s easier to use page helpers, which ensure that the page is modified in a semantically correct manner.

Page helpers

pikepdf provides a helper class, pikepdf.Page, which provides higher-level functions to manipulate pages than the standard page dictionary used in the previous examples.

Currently pikepdf does not automatically return helper classes. You must initialize them. In a future release, it will return them automatically.

One advantage of page helpers is that they resolve implicit information. For example, page.trimbox will return an appropriate trim box for this page, which in this case is equal to the media box.
## 1.3.6 Object model

This section covers the object model pikepdf uses in more detail.

A `pikepdf.Object` is a Python wrapper around a C++ `QPDFObjectHandle` which, as the name suggests, is a handle (or pointer) to a data structure in memory, or possibly a reference to data that exists in a file. Importantly, an object can be a scalar quantity (like a string) or a compound quantity (like a list or dict, that contains other objects). The fact that the C++ class involved here is an object `handle` is an implementation detail; it shouldn’t matter for a pikepdf user.

The simplest types in PDFs are directly represented as Python types: `int`, `bool`, and `None` stand for PDF integers, booleans and the “null”. `Decimal` is used for floating point numbers in PDFs. If a value in a PDF is assigned to a Python `float`, pikepdf will convert it to `Decimal`.

Types that are not directly convertible to Python are represented as `pikepdf.Object`, a compound object that offers a superset of possible methods, some of which only if the underlying type is suitable. Use the EAFP (easier to ask forgiveness than permission) idiom, or `isinstance` to determine the type more precisely. This partly reflects the fact that the PDF specification allows many data fields to be one of several types.

For convenience, the `repr()` of a `pikepdf.Object` will display a Python expression that replicates the existing object (when possible), so it will say:

```bash
>>> catalog_name = pdf.root.Type
pikepdf.Name("/Catalog")
>>> isinstance(catalog_name, pikepdf.Name)
True
>>> isinstance(catalog_name, pikepdf.Object)
True
```

### Making PDF objects

You may construct a new object with one of the classes:

- `pikepdf.Array`
- `pikepdf.Dictionary`
- `pikepdf.Name` - the type used for keys in PDF Dictionary objects
- `pikepdf.String` - a text string (treated as `bytes` and `str` depending on context)

These may be thought of as subclasses of `pikepdf.Object`. (Internally they are `pikepdf.Object`.)

There are a few other classes for special PDF objects that don’t map to Python as neatly.

- `pikepdf.Operator` - a special object involved in processing content streams
- `pikepdf.Stream` - a special object similar to a `Dictionary` with binary data attached
- `pikepdf.InlineImage` - an image that is embedded in content streams

The great news is that it’s often unnecessary to construct `pikepdf.Object` objects when working with pikepdf. Python types are transparently *converted* to the appropriate pikepdf object when passed to pikepdf APIs — when possible. However, pikepdf sends `pikepdf.Object` types back to Python on return calls, in most cases, because pikepdf needs to keep track of objects that came from PDFs originally.

### Object lifecycle and memory management

As mentioned above, a `pikepdf.Object` may reference data that is lazily loaded from its source `pikepdf.Pdf`. Closing the `Pdf` with `pikepdf.Pdf.close()` will invalidate some objects, depending on whether or not the data
was loaded, and other implementation details that may change. Generally speaking, a `pikepdf.Pdf` should be held open until it is no longer needed, and objects that were derived from it may or may not be usable after it is closed.

Simple objects (booleans, integers, decimals, `None`) are copied directly to Python as pure Python objects.

For PDF stream objects, use `pikepdf.Object.read_bytes()` to obtain a copy of the object as pure bytes data, if this information is required after closing a PDF.

When objects are copied from one `pikepdf.Pdf` to another, the underlying data is copied immediately into the target. As such it is possible to merge hundreds of `Pdf` into one, keeping only a single source and the target file open at a time.

### 1.3.7 Stream objects

A `pikepdf.Stream` object works like a PDF dictionary with some encoded bytes attached. The dictionary is metadata that describes how the stream is encoded. PDF can, and regularly does, use a variety of encoding filters. A stream can be encoded with one or more filters. Images are a type of stream object.

Most of the interesting content in a PDF (images and content streams) are inside page objects.

Because the PDF specification unfortunately defines several terms involve the word stream, let’s attempt to clarify:

- **stream object** A PDF object that contains binary data and a metadata dictionary to describes it, represented as `pikepdf.Stream`. In HTML this is equivalent to a `<img>` with inline image data.

- **object stream** A stream object (not a typo, an object stream really is a type of stream object) in a PDF that contains a number of other objects in a PDF, grouped together for better compression. In pikepdf there is an option to save PDFs with this feature enabled to improve compression. Otherwise, this is just a detail about how PDF files are encoded.

- **content stream** A stream object that contains some instructions to draw graphics and text on a page, or inside a Form XObject. In HTML this is equivalent to the HTML file itself. Content streams do not cross pages.

- **Form XObject** A group of images, text and drawing commands that can be rendered elsewhere in a PDF as a group. This is often used when a group of objects are needed at different scales or multiple pages. In HTML this is like an `<svg>`.

**Reading stream objects**

Fortunately, `pikepdf.Stream.read_bytes()` will apply all filters and decode the uncompressed bytes, or throw an error if this is not possible. `pikepdf.Stream.read_raw_bytes()` provides access to the compressed bytes.

Three types of stream object are particularly noteworthy: content streams, which describe the order of drawing operators; images; and XMP metadata. pikepdf provides helper functions for working with these types of streams.

### 1.3.8 Working with content streams

A content stream is a stream object associated with either a page or a Form XObject that describes where and how to draw images, vectors, and text.

Content streams are binary data that can be thought of as a list of operators and zero or more operands. Operands are given first, followed by the operator. It is a stack-based language based loosely on PostScript, but without any programmable features. There are no variables, loops or conditionals.

A typical example is as follows (with additional whitespace):
pikepdf provides a C++ optimized content stream parser and a filter. The parser is best used for reading and interpreting content streams; the filter is best used for rewriting them.

```python
In [1]: pdf = pikepdf.open("../tests/resources/congress.pdf")
In [2]: page = pdf.pages[0]
In [3]: for operands, operator in pikepdf.parse_content_stream(page):
   ...:     print("Operands {}, operator {}").format(operands, operator))
```

Operands [], operator q
Operands [Decimal('200.0000'), 0, 0, Decimal('304.0000'), Decimal('0.0000'), Decimal('0.0000')], operator cm
Operands [pikepdf.Name("/Im0")], operator Do
Operands [], operator Q

### Extracting text from PDFs

If you guessed that the content streams were the place to look for text inside a PDF – you’d be correct. Unfortunately, extracting the text is fairly difficult because content stream actually specifies as a font and glyph numbers to use. Sometimes, there is a 1:1 transparent mapping between Unicode numbers and glyph numbers, and dump of the content stream will show the text. In general, you cannot rely on there being a transparent mapping: in fact, it is perfectly legal for a font to specify no Unicode mapping at all, or to use an unconventional mapping (when a PDF contains a subsetted font for example).

We strongly recommend against trying to scrape text from the content stream.

pikepdf does not currently implement text extraction. We recommend pdfminer.six, a read-only text extraction tool. If you wish to write PDFs containing text, consider reportlab.

### 1.3.9 Working with images

PDFs embed images as binary stream objects within the PDF’s data stream. The stream object’s dictionary describes properties of the image such as its dimensions and color space. The same image may be drawn multiple times on multiple pages, at different scales and positions.

In some cases such as JPEG2000, the standard file format of the image is used verbatim, even when the file format contains headers and information that is repeated in the stream dictionary. In other cases such as for PNG-style encoding, the image file format is not used directly.

pikepdf currently has no facility to embed new images into PDFs. We recommend img2pdf instead, because it does the job so well. pikepdf instead allows for image inspection and lossless/transcode free (where possible) “pdf2img”.

### Playing with images

pikepdf provides a helper class `PdfImage` for manipulating images in a PDF. The helper class helps manage the complexity of the image dictionaries.

```python
In [1]: from pikepdf import Pdf, PdfImage, Name
In [2]: example = Pdf.open('..tests/resources/congress.pdf')
In [3]: page1 = example.pages[0]
```
In [4]: list(page1.images.keys())
Out[4]: ['/Im0']

In [5]: rawimage = page1.images['/Im0']  # The raw object/dictionary

In [6]: pdfimage = PdfImage(rawimage)

In [7]: pdfimage
Out[7]: <pikepdf.PdfImage image mode=RGB size=1000x1520 at 0x7f1d8c383048>

In Jupyter (or IPython with a suitable backend) the image will be displayed.

![Image](image.jpg)

You can also inspect the properties of the image. The parameters are similar to Pillow's.

In [8]: pdfimage.colorspace
Out[8]: '/DeviceRGB'

In [9]: pdfimage.width, pdfimage.height
Out[9]: (1000, 1520)

Note: .width and .height are the resolution of the image in pixels, not the size of the image in page coordinates. The size of the image in page coordinates is determined by the content stream.

Extracting images

Extracting images is straightforward. extract_to() will extract images to a specified file prefix. The extension is determined while extracting and appended to the filename. Where possible, extract_to writes compressed data directly to the stream without transcoding.

In [10]: pdfimage.extract_to(fileprefix='image'))
Out[10]: 'image.jpg'

It also possible to extract to a writable Python stream using .extract_to(stream=...`).
You can also retrieve the image as a Pillow image:

```python
In [11]: pdfimage.as_pil_image()
Out[11]: <PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=1000x1520 at 0x7F1D8C33EFD0>
```

Another way to view the image is using Pillow’s `Image.show()` method.

Not all images can be extracted. Also, some PDFs describe an image with a mask, with transparency effects. pikepdf can only extract the images themselves, not rasterize them exactly as they appear in a PDF viewer. In the vast majority of cases, however, the image can be extracted as it appears.

**Note:** This simple example PDF displays a single full page image. Some PDF creators will paint a page using multiple images, and features such as layers, transparency and image masks. Accessing the first image on a page is like an HTML parser that scans for the first `<img src="">` tag it finds. A lot more could be happening. There can be multiple images drawn multiple times on a page, vector art, overwriting, masking, and transparency. A set of resources can be grouped together in a “Form XObject” (not to be confused with a PDF Form), and drawn at all once. Images can be referenced by multiple pages.

### Replacing an image

In this example we extract an image and replace it with a grayscale equivalent.

```python
In [12]: import zlib
In [13]: rawimage = pdfimage.obj
In [14]: pillowimage = pdfimage.as_pil_image()
In [15]: grayscale = pillowimage.convert('L')
In [16]: grayscale = grayscale.resize((32, 32))
In [17]: rawimage.write(zlib.compress(grayscale.tobytes()), filter=Name("/FlateDecode")).
In [18]: rawimage.ColorSpace = Name("/DeviceGray")
In [19]: rawimage.Width, rawimage.Height = 32, 32
```

Notes on this example:

- It is generally possible to use `zlib.compress()` to generate compressed image data, although this is not as efficient as using a program that knows it is preparing a PDF.
- In general we can resize an image to any scale. The PDF content stream specifies where to draw an image and at what scale.
- This example would replace all occurrences of the image if it were used multiple times in a PDF.

### 1.3.10 Character encoding

There are three hard problems in computer science: 1) Converting from PDF, 2) Converting to PDF, and 3) OO
In most circumstances, pikepdf performs appropriate encodings and decodings on its own, or returns `pikepdf.String` if it is not clear whether to present data as a string or binary data.

`str(pikepdf.String)` is performed by inspecting the binary data. If the binary data begins with a UTF-16 byte order mark, then the data is interpreted as UTF-16 and returned as a Python `str`. Otherwise, the data is returned as a Python `str`, if the binary data will be interpreted as PDFDocEncoding and decoded to `str`. Again, in most cases this is correct behavior and will operate transparently.

Some functions are available in circumstances where it is necessary to force a particular conversion.

**PDFDocEncoding**

The PDF specification defines PDFDocEncoding, a character encoding used only in PDFs. It is quite similar to ASCII but not equivalent.

When pikepdf is imported, it automatically registers "pdfdoc" as a codec with the standard library, so that it may be used in string and byte conversions.

```python
"•".encode('pdfdoc') == b'\x81'
```

**Other codecs**

Two other codecs are commonly used in PDFs, but they are already part of the standard library.

**WinAnsiEncoding** is identical Windows Code Page 1252, and may be converted using the "cp1251" codec.

**MacRomanEncoding** may be converted using the "macroman" codec.

### 1.3.11 PDF Metadata

PDF has two different types of metadata: XMP metadata, and DocumentInfo, which is deprecated but still relevant. For backward compatibility, both should contain the same content. pikepdf provides a convenient interface that coordinates edits to both, but is limited to the most common metadata features.

XMP (Extensible Metadata Platform) Metadata is a metadata specification in XML format that is used in many formats other than PDF. For full information on XMP, see Adobe’s XMP Developer Center. The XMP Specification also provides useful information.

pikepdf can read compound metadata quantities may be read, but only scalar quantities can be modified. For more complex changes consider using the `python-xmp-toolkit` library and its libexempi dependency; but note that it is not capable of synchronizing changes to the older DocumentInfo metadata.

**Accessing metadata**

The XMP metadata stream is attached the PDF’s root object, but to simplify management of this, use `pikepdf.Pdf.open_metadata()`. The returned `pikepdf.models.PdfMetadata` object may be used for reading, or entered with a `with` block to modify and commit changes. If you use this interface, pikepdf will synchronize changes to new and old metadata.

A PDF must still be saved after metadata is changed.
In [1]: pdf = pikepdf.open('..tests/resources/sandwich.pdf')
In [2]: meta = pdf.open_metadata()
In [3]: meta['xmp:CreatorTool']
Out[3]: 'ocrmypdf 5.3.3 / Tesseract OCR-PDF 3.05.01'

If no XMP metadata exists, an empty XMP metadata container will be created.

Open metadata in a with block to open it for editing. When the block is exited, changes are committed (updating XMP and the Document Info dictionary) and attached to the PDF object. The PDF must still be saved. If an exception occurs in the block, changes are discarded.

In [4]: with pdf.open_metadata() as meta:
   ...
   ...
   meta['dc:title'] = "Let's change the title"
   ...

The list of available metadata fields may be found in the XMP Specification.

Checking PDF/A conformance

The metadata interface can also test if a file claims to be conformant to the PDF/A specification.

In [5]: pdf = pikepdf.open('..tests/resources/veraPDF test suite 6-2-10-t02-pass-a.pdf')
In [6]: meta = pdf.open_metadata()
In [7]: meta.pdfa_status
Out[7]: '1B'

Note: Note that this property merely tests if the file claims to be conformant to the PDF/A standard. Use a tool such as veraPDF to verify conformance.

Low-level XMP metadata access

You can read the raw XMP metadata if desired. For example, one could extract it and edit it using the full featured python-xmp-toolkit library.

In [8]: xmp = pdf.root.Metadata.read_bytes()
In [9]: type(xmp)
Out[9]: bytes
In [10]: print(xmp.decode())
\\\\\\\<?xpacket begin='' id='W5M0MpCehiHzreSzNTczkc9d'?>
<x:xmpmeta xmlns:x="adobe:ns:meta/">
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/" rdf:about="">
      <dc:creator>
        <rdf:Seq>
        <rdf:li>veraPDF Consortium</rdf:li>
        <rdf:li>veraPDF Consortium</rdf:li>
        </rdf:Seq>
      </dc:creator>
      <rdf:Seq>
      </rdf:Seq>
    </rdf:Description>
  </rdf:RDF>
</x:xmpmeta>
(continues on next page)
Editing XMP with a generic XML library is probably not worth the trouble; the semantics are fairly complex.

**Warning:** Manually changes to XMP stream object will not be synchronized with live PdfMetadata object or the DocumentInfo block.

### The Document Info dictionary

The Document Info block is an older, now deprecated object in which metadata may be stored. The Document Info is not attached to the /Root object. It may be accessed using the `.docinfo` property. If no Document Info exists, touching the `.docinfo` will properly initialize an empty one.

Here is an example of a Document Info block.

```python
In [11]: pdf = pikepdf.open('../tests/resources/sandwich.pdf')
In [12]: pdf.docinfo
Out[12]:
pikepdf.Dictionary({
    '/CreationDate': 'D:20170911132748-07'00''',
    '/Creator': 'ocrmypdf 5.3.3 / Tesseract OCR-PDF 3.05.01',
    '/ModDate': 'D:20170911132748-07'00''',
    '/Producer': 'GPL Ghostscript 9.21'
})
```

It is permitted in pikepdf to directly interact with Document Info as with other PDF dictionaries. However, it is better to use `.open_metadata()` because that interface will apply changes to both XMP and Document Info in a consistent manner.

You may copy from data from a Document Info object in the current PDF or another PDF into XMP metadata using `load_from_docinfo()`.

### 1.3.12 Main objects

**class** pikepdf.Pdf

In-memory representation of a PDF

**Root**

The `/Root` object of the PDF.
add_blank_page(*, page_size=(612, 792))

Add a blank page to this PD. If pages already exist, the page will be added to the end. Pages may be reordered using Pdf.pages.

The caller may add content to the page by modifying its objects after creating it.

Parameters page_size (tuple) – The size of the page in PDF units (1/72 inch or 0.35mm).

Default size is set to a US Letter 8.5” x 11” page.

allow

Report permissions associated with this PDF.

By default these permissions will be replicated when the PDF is saved. Permissions may also only be changed when a PDF is being saved, and are only available for encrypted PDFs. If a PDF is not encrypted, all operations are reported as allowed.

pikepdf has no way of enforcing permissions.

Returns  pikepdf.models.Permisons

check_linearization (self: pikepdf.Pdf, stream: object = sys.stderr) → None

Reports information on the PDF’s linearization

Parameters stream – A stream to write this information too; must implement .write() and .flush() method. Defaults to sys.stderr.

close()

Close a Pdf object and release resources acquired by pikepdf.

If pikepdf opened the file handle it will close it (e.g. when opened with a file path). If the caller opened the file for pikepdf, the caller close the file.

pikepdf lazily loads data from PDFs, so some pikepdf.Object may implicitly depend on the pikepdf.Pdf being open. This is always the case for pikepdf.Stream but can be true for any object. Do not close the Pdf object if you might still be accessing content from it.

When an Object is copied from one Pdf to another, the Object is copied into the destination Pdf immediately, so after accessing all desired information from the source Pdf it may be closed.

Caution: Closing the Pdf is currently implemented by resetting it to an empty sentinel. It is currently possible to edit the sentinel as if it were a live object. This behavior should not be relied on and is subject to change.


Copy object from foreign PDF to this one.

docinfo

Access the (deprecated) document information dictionary.

The document information dictionary is a brief metadata record that can store some information about the origin of a PDF. It is deprecated and removed in the PDF 2.0 specification. Use the .open_metadata() API instead, which will edit the modern (and unfortunately, more complicated) XMP metadata object and synchronize changes to the document information dictionary.

encryption

Report encryption information for this PDF.

Encryption settings may only be changed when a PDF is saved.

Returns: pikepdf.models.EncryptionInfo
filename
   The source filename of an existing PDF, when available.

get_object (*args, **kwargs)
   Overloaded function.
   1. get_object(self: pikepdf.Pdf, objgen: Tuple[int, int]) -> pikepdf.Object
      Look up an object by ID and generation number
      Return type: pikepdf.Object
   2. get_object(self: pikepdf.Pdf, objid: int, gen: int) -> pikepdf.Object
      Look up an object by ID and generation number
      Return type: pikepdf.Object

get_warnings (self: pikepdf.Pdf) -> list

is_linearized
   Returns True if the PDF is linearized.
   Specifically returns True iff the file starts with a linearization parameter dictionary. Does no additional validation.

make_indirect (*args, **kwargs)
   Overloaded function.
      Attach an object to the Pdf as an indirect object
      Direct objects appear inline in the binary encoding of the PDF. Indirect objects appear inline as references (in English, “look up object 4 generation 0”) and then read from another location in the file. The PDF specification requires that certain objects are indirect - consult the PDF specification to confirm.
      Generally a resource that is shared should be attached as an indirect object. pikepdf.Stream objects are always indirect, and creating them will automatically attach it to the Pdf.
      See Also: pikepdf.Object.is_indirect()
      Return type: pikepdf.Object
   2. make_indirect(self: pikepdf.Pdf, obj: object) -> pikepdf.Object
      Encode a Python object and attach to this Pdf as an indirect object
      Return type: pikepdf.Object

make_stream (data)
   Create a new pikepdf.Stream object that is attached to this PDF.

   Parameters data (bytes) – Binary data for the stream object

static new () -> pikepdf.Pdf
   Create a new empty PDF from stratch.

   Open an existing file at filename_or_stream.
If `filename_or_stream` is path-like, the file will be opened for reading. The file should not be modified by another process while it is open in pikepdf. The file will not be altered when opened in this way. Any changes to the file must be persisted by using `.save()`.

If `filename_or_stream` has `.read()` and `.seek()` methods, the file will be accessed as a readable binary stream. pikepdf will read the entire stream into a private buffer.

`.open()` may be used in a `with`-block, `.close()` will be called when the block exists.

**Examples**

```python
>>> with Pdf.open("test.pdf") as pdf:
    ...

>>> pdf = Pdf.open("test.pdf", password="rosebud")
```

**Parameters**

- `filename_or_stream (os.PathLike)` – Filename of PDF to open
- `password (str or bytes)` – User or owner password to open an encrypted PDF. If a str is given it will be converted to UTF-8.
- `hex_password (bool)` – If True, interpret the password as a hex-encoded version of the exact encryption key to use, without performing the normal key computation. Useful in forensics.
- `ignore_xref_streams (bool)` – If True, ignore cross-reference streams. See qpdf documentation.
- `suppress_warnings (bool)` – If True (default), warnings are not printed to stderr. Use `pikepdf.Pdf.get_warnings()` to retrieve warnings.
- `attempt_recovery (bool)` – If True (default), attempt to recover from PDF parsing errors.
- `inherit_page_attributes (bool)` – If True (default), push attributes set on a group of pages to individual pages

**Raises**

- `pikepdf.PasswordError` – If the password failed to open the file.
- `pikepdf.PdfError` – If for other reasons we could not open the file.
- `TypeError` – If the type of `filename_or_stream` is not usable.
- `FileNotFoundError` – If the file was not found.

**open_metadata (set_pikepdf_as_editor=True, update_docinfo=True)**

Open the PDF’s XMP metadata for editing

Recommend for use in a `with` block. Changes are committed to the PDF when the block exits. (The `Pdf` must still be opened.)

**Example**

1.3. In use
>>> with pdf.open_metadata() as meta:
    meta['dc:title'] = 'Set the Dublin Core Title'
    meta['dc:description'] = 'Put the Abstract here'

Parameters

- **set_pikepdf_as_editor** *(bool)* – Update the metadata to show that this version of pikepdf is the most software to modify the metadata. Recommended, except for testing.

- **update_docinfo** *(bool)* – Update the deprecated PDF DocumentInfo block to be consistent with XMP.

Returns  pikepdf.models.PdfMetadata

**pages**

Returns the list of pages.

Return type:  pikepdf._qpdf.PageList

**pdf_version**

The version of the PDF specification used for this file, such as ‘1.7’.

**remove_unreferenced_resources** *(self: pikepdf.Pdf) → None*

Remove from /Resources of each page any object not referenced in page’s contents

PDF pages may share resource dictionaries with other pages. If pikepdf is used for page splitting, pages may reference resources in their /Resources dictionary that are not actually required. This purges all unnecessary resource entries.

Suggested before saving.

**root**

Alias for .Root, the /Root object of the PDF.


Save all modifications to this pikepdf.Pdf.

Parameters

- **filename** *(str or stream)* – Where to write the output. If a file exists in this location it will be overwritten. The file should not be the same as the input file, because data from the input file may be lazily loaded; as such overwriting in place will null-out objects.

- **static_id** *(bool)* – Indicates that the /ID metadata, normally calculated as a hash of certain PDF contents and metadata including the current time, should instead be generated deterministically. Normally for debugging.

- **preserve_pdfa** *(bool)* – Ensures that the file is generated in a manner compliant with PDF/A and other stricter variants. This should be True, the default, in most cases.

- **min_version** *(str or tuple)* – Sets the minimum version of PDF specification that should be required. If left alone QPDF will decide. If a tuple, the second element is an integer, the extension level. If the version number is not a valid format, QPDF will decide what to do.
• **force_version** *(str or tuple)* – Override the version recommend by QPDF, potentially creating an invalid file that does not display in old versions. See QPDF manual for details. If a tuple, the second element is an integer, the extension level.

• **fix_metadata_version** *(bool)* – If True (default) and the XMP metadata contains the optional PDF version field, ensure the version in metadata is correct. If the XMP metadata does not contain a PDF version field, none will be added. To ensure that the field is added, edit the metadata and insert a placeholder value in `pdf:PDFVersion`.

• **object_stream_mode** *(pikepdf.ObjectStreamMode)* – `disable` prevents the use of object streams. `preserve` keeps object streams from the input file. `generate` uses object streams wherever possible, creating the smallest files but requiring PDF 1.5+.

• **compress_streams** *(bool)* – Enables or disables the compression of stream objects in the PDF. Metadata is never compressed. By default this is set to `True`, and should be except for debugging.

• **stream_decode_level** *(pikepdf.StreamDecodeLevel)* – Specifies how to encode stream objects. See documentation for `StreamDecodeLevel`.

• **normalize_content** *(bool)* – Enables parsing and reformatting the content stream within PDFs. This may debugging PDFs easier.

• **linearize** *(bool)* – Enables creating linear or “fast web view”, where the file’s contents are organized sequentially so that a viewer can begin rendering before it has the whole file. As a drawback, it tends to make files larger.

• **qdf** *(bool)* – Save output QDF mode. QDF mode is a special output mode in QPDF to allow editing of PDFs in a text editor. Use the program `fix-qdf` to fix convert back to a standard PDF.

• **progress** *(callable)* – Specify a callback function that is called as the PDF is written. The function will be called with an integer between 0-100 as the sole parameter, the progress percentage. This function may not access or modify the PDF while it is being written, or data corruption will almost certainly occur.

• **encryption** *(pikepdf.models.Encryption or bool)* – If False or omitted, existing encryption will be removed. If True encryption settings are copied from the originating PDF. Alternately, an Encryption object may be provided that sets the parameters for new encryption.

You may call `.save()` multiple times with different parameters to generate different versions of a file, and you may continue to modify the file after saving it. `.save()` does not modify the Pdf object in memory, except possibly by updating the XMP metadata version with `fix_metadata_version`.

---

**Note:** `pikepdf.Pdf.remove_unreferenced_resources()` before saving may eliminate unnecessary resources from the output file, so calling this method before saving is recommended. This is not done automatically because `.save()` is intended to be idempotent.

---

**Note:** pikepdf can read PDFs will incremental updates, but always any coalesces incremental updates into a single non-incremental PDF file when saving.

---

**show_xref_table** *(self: pikepdf.Pdf) → None*

Pretty-print the Pdf’s xref (cross-reference table)
**trailer**

Provides access to the PDF trailer object.

See section 7.5.5 of the PDF reference manual. Generally speaking, the trailer should not be modified with pikepdf, and modifying it may not work. Some of the values in the trailer are automatically changed when a file is saved.

```python
pikepdf.open(*args, **kwargs)
```

Alias for `pikepdf.Pdf.open()`. Open a PDF.

```python
pikepdf.new(*args, **kwargs)
```

Alias for `pikepdf.Pdf.new()`. Create a new empty PDF.

**class pikepdf.ObjectStreamMode**

Options for saving streams within PDFs, which are more a compact way of saving certains types of data that was added in PDF 1.5. All modern PDF viewers support object streams, but some third party tools and libraries cannot read them.

- **disable**
  - Disable the use of object streams. If any object streams exist in the file, remove them when the file is saved.

- **preserve**
  - Preserve any existing object streams in the original file. This is the default behavior.

- **generate**
  - Generate object streams.

**class pikepdf.StreamDecodeLevel**

Options for decoding streams within PDFs.

- **none**
  - Do not attempt to apply any filters. Streams remain as they appear in the original file. Note that uncompressed streams may still be compressed on output. You can disable that by saving with `.save(..., compress_streams=False)`.

- **generalized**
  - This is the default. libqpdf will apply LZWDecode, ASCII85Decode, ASCIIHexDecode, and FlateDecode filters on the input. When saved with `compress_streams=True`, the default, the effect of this is that streams filtered with these older and less efficient filters will be recompressed with the Flate filter. As a special case, if a stream is already compressed with FlateDecode and `compress_streams=True`, the original compressed data will be preserved.

- **specialized**
  - In addition to uncompressing the generalized compression formats, supported non-lossy compression will also be be decoded. At present, this includes the RunLengthDecode filter.

- **all**
  - In addition to generalized and non-lossy specialized filters, supported lossy compression filters will be applied. At present, this includes DCTDecode (JPEG) compression. Note that compressing the resulting data with DCTDecode again will accumulate loss, so avoid multiple compression and decompression cycles. This is mostly useful for (low-level) retrieving image data; see `pikepdf.PdfImage` for the preferred method.

**class pikepdf.Encryption(*, owner, user, R=6, allow=Permissions(__class__=<class 'pikepdf.models.encryption.Permissions'>), accessibility=True, extract=True, modify_annotation=True, modify_assembly=False, modify_form=True, modify_other=True, print_highres=True, print_lowres=True), aes=True, metadata=True)**

Specify the encryption settings to apply when a PDF is saved.
Parameters

- **owner** *(str)* – The owner password to use. This allows full control of the file. If blank, the PDF will be encrypted and present as “(SECURED)” in PDF viewers. If the owner password is blank, the user password should be as well.

- **user** *(str)* – The user password to use. With this password, some restrictions will be imposed by a typical PDF reader. If blank, the PDF can be opened by anyone, but only modified as allowed by the permissions in allow.

- **R** *(int)* – Select the security handler algorithm to use. Choose from: 2, 3, 4 or 6. By default, the highest version of is selected (6). 5 is a deprecated algorithm that should not be used.

- **allow** *(pikepdf.Permissions)* – The permissions to set. If omitted, all permissions are granted to the user.

- **aes** *(bool)* – If True, request the AES algorithm. If False, use RC4. If omitted, AES is selected whenever possible (R >= 4).

- **metadata** *(bool)* – If True, also encrypt the PDF metadata. If False, metadata is not encrypted. Reading document metadata without decryption may be desirable in some cases. Requires aes=True. If omitted, metadata is encrypted whenever possible.

exception pikepdf.PdfError
exception pikepdf.PasswordError

Object construction

class pikepdf.Object

  as_dict *(self: pikepdf.Object) → pikepdf._qpdf._ObjectMapping*

  as_list *(self: pikepdf.Object) → pikepdf._qpdf._ObjectList*

  emplace *(other)*
  Copy all items from other without making a new object.

  Particularly when working with pages, it may be desirable to remove all of the existing page’s contents and emplace (insert) a new page on top of it, in a way that preserves all links and references to the original page. (Or similarly, for other Dictionary objects in a PDF.)

  When a page is assigned (pdf.pages[0] = new_page), only the application knows if references to the original the original page are still valid. For example, a PDF optimizer might restructure a page object into another visually similar one, and references would be valid; but for a program that reorganizes page contents such as a N-up compositor, references may not be valid anymore.

  This method takes precautions to ensure that child objects in common with self and other are not inadvertently deleted.

Example

```python
>>> pdf.pages[0].objgen
(16, 0)
>>> pdf.pages[0].emplace(pdf.pages[1])
>>> pdf.pages[0].objgen
(16, 0) # Same object
```
get (*args, **kwargs)
Overloaded function.
   1. get(self: pikepdf.Object, key: str, default: object = None) -> object
      For pikepdf.Dictionary objects, behave as dict.get(key, default=None)
   2. get(self: pikepdf.Object, key: pikepdf.Object, default: object = None) -> object
      For pikepdf.Dictionary objects, behave as dict.get(key, default=None)

Return a buffer protocol buffer describing the raw, encoded stream

get_stream_buffer (self: pikepdf.Object) → pikepdf._qpdf.Buffer
Return a buffer protocol buffer describing the decoded stream

is_owned_by (self: pikepdf.Object, possible_owner: pikepdf.Pdf) → bool
Test if this object is owned by the indicated possible_owner.

is_rectangle
Returns True if the object is a rectangle (an array of 4 numbers)

items (self: pikepdf.Object) → iterable

keys (self: pikepdf.Object) → Set[str]

objgen
Return the object-generation number pair for this object
   If this is a direct object, then the returned value is (0, 0). By definition, if this is an indirect object, it has a “objgen”, and can be looked up using this in the cross-reference (xref) table. Direct objects cannot necessarily be looked up.

   The generation number is usually 0, except for PDFs that have been incrementally updated.

Append or prepend to an existing page’s content stream.

page_contents_coalesce (self: pikepdf.Object) → None
Coalesce an array of page content streams into a single content stream.
   The PDF specification allows the /Contents object to contain either an array of content streams or a single content stream. However, it simplifies parsing and editing if there is only a single content stream. This function merges all content streams.

static parse (stream: str, description: str = "") → pikepdf.Object
Parse PDF binary representation into PDF objects.

read_bytes (self: pikepdf.Object) → bytes
Decode and read the content stream associated with this object

read_raw_bytes (self: pikepdf.Object) → bytes
Read the content stream associated with this object without decoding

to_json (self: pikepdf.Object, dereference: bool = False) → bytes
Convert to a QPDF JSON representation of the object.

   Not necessarily compatible with other PDF-JSOn representations that exist in the wild.
      • Names are encoded as UTF-8 strings
• Indirect references are encoded as strings containing obj gen R
• Strings are encoded as UTF-8 strings with unrepresentable binary characters encoded as \uHHHH
• Encoding streams just encodes the stream’s dictionary; the stream data is not represented
• Object types that are only valid in content streams (inline image, operator) as well as “reserved” objects are not representable and will be serialized as null.

Parameters dereference (bool) – If True, dereference the object is this is an indirect object.

Returns JSON bytestring of object. The object is UTF-8 encoded and may be decoded to a Python str that represents the binary values \x00-\xFF as U+0000 to U+00FF; that is, it may contain mojibake.

Return type bytes

unparse (self: pikepdf.Object, resolved: bool = False) → bytes
Convert PDF objects into their binary representation, optionally resolving indirect objects.

wrap_in_array (self: pikepdf.Object) → pikepdf.Object
Return the object wrapped in an array if not already an array.

write (data, *, filter=None, decode_parms=None, type_check=True)
Replace stream object’s data with new (possibly compressed) data.

filter and decode_parms specify that compression that is present on the input data.

When writing the PDF in pikepdf.Pdf.save(), pikepdf may change the compression or apply compression to data that was not compressed, depending on the parameters given to that function. It will never change lossless to lossy encoding.

PNG and TIFF images, even if compressed, cannot be directly inserted into a PDF and displayed as images.

Parameters

• data (bytes) – the new data to use for replacement
• filter (pikepdf.Name or pikepdf.Array) – The filter(s) with which the data is (already) encoded
• decode_parms (pikepdf.Dictionary or pikepdf.Array) – Parameters for the filters with which the object is encode
• type_check (bool) – Check arguments; use False only if you want to intentionally create malformed PDFs.

If only one filter is specified, it may be a name such as Name("/FlateDecode"). If there are multiple filters, then array of names should be given.

If there is only one filter, decode_parms is a Dictionary of parameters for that filter. If there are multiple filters, then decode_parms is an Array of Dictionary, where each array index is corresponds to the filter.

class pikepdf.Name
Constructs a PDF Name object

Names can be constructed with two notations:

1. Name.Resources
2. Name('/Resources')

The two are semantically equivalent. The former is preferred for names that are normally expected to be in a PDF. The latter is preferred for dynamic names and attributes.
static __new__(cls, name)
Create and return a new object. See help(type) for accurate signature.

class pikepdf.String
Constructs a PDF String object
static __new__(cls, s)
    Parameters s (str or bytes) – The string to use. String will be encoded for PDF, bytes
will be constructed without encoding.
    Returns pikepdf.Object

class pikepdf.Array
Constructs a PDF Array object
static __new__(cls, a=None)
    Parameters a (iterable) – A list of objects. All objects must be either pikepdf.Object or
convertible to pikepdf.Object.
    Returns pikepdf.Object

class pikepdf.Dictionary
Constructs a PDF Dictionary object
static __new__(cls, d=None, **kwargs)
    Constructs a PDF Dictionary from either a Python dict or keyword arguments.
    These two examples are equivalent:

    >>> pikepdf.Dictionary({'/NameOne': 1, '/NameTwo': 'Two'})
    >>> pikepdf.Dictionary(NameOne=1, NameTwo='Two')

    In either case, the keys must be strings, and the strings correspond to the desired Names in the PDF
Dictionary. The values must all be convertible to pikepdf.Object.
    Returns pikepdf.Object

class pikepdf.Stream
Constructs a PDF Stream object
static __new__(cls, owner, obj)
    Parameters
        - **owner (pikepdf.Pdf)** – The Pdf to which this stream shall be attached.
        - **obj (bytes or list)** – If bytes, the data bytes for the stream. If list,
a list of (operands, operator) tuples such as returned by pikepdf.
parse_content_stream().
    Returns pikepdf.Object

class pikepdf.Operator (op: str) → pikepdf.Object
Construct a PDF Operator object for use in content streams.

Internal objects

These objects are returned by other pikepdf objects. They are part of the API, but not intended to be created explicitly.
class pikepdf._qpdf.PageList
A list-like object enumerating all pages in a pikepdf.Pdf.

Add another page to the end.

extend (*args, **kwargs)
Overloaded function.
Extend the Pdf by adding pages from another Pdf.pages.
2. extend(self: pikepdf._qpdf.PageList, iterable: iterable) -> None
Extend the Pdf by adding pages from an iterable of pages.

insert (self: pikepdf._qpdf.PageList, index: int, obj: object) → None
Insert a page at the specified location.

Parameters
• index (int) – location at which to insert page, 0-based indexing
• obj (pikepdf.Object) – page object to insert

Convenience - look up page number in ordinal numbering, .p(1) is first page

remove (self: pikepdf._qpdf.PageList, **kwargs) → None
Remove a page (using 1-based numbering)

Parameters p (int) – 1-based page number

reverse (self: pikepdf._qpdf.PageList) → None
Reverse the order of pages.

1.3.13 Support models

Support models are abstracts over “raw” objects within a Pdf. For example, a page in a PDF is a Dictionary with set to /Type of /Page. The Dictionary in that case is the “raw” object. Upon establishing what type of object it is, we can wrap it with a support model that adds features to ensure consistency with the PDF specification.

pikepdf does not currently apply support models to “raw” objects automatically, but might do so in a future release (this would break backward compatibility).

For example, to initialize a Page support model:

```python
from pikepdf import Pdf, Page
Pdf = open(...)  
page_support_model = Page(pdf.pages[0])
```

class pikepdf.Page

add_content_token_filter (self: pikepdf.Page, tf: pikepdf.Object::TokenFilter) → None
Attach a pikepdf.TokenFilter to a page’s content stream.

This function applies token filters lazily, if/when the page’s content stream is read for any reason, such as when the PDF is saved. If never access, the token filter is not applied.

Multiple token filters may be added to a page/content stream.

1.3. In use
If the page’s contents is an array of streams, it is coalesced.

```python
as_form_xobject (self: pikepdf.Page, handle_transformations: bool = True) -> pikepdf.Object
```
Return a form XObject that draws this page.

This is useful for n-up operations, underlay, overlay, thumbnail generation, or any other case in which it is useful to replicate the contents of a page in some other context. The dictionaries are shallow copies of the original page dictionary, and the contents are coalesced from the page’s contents. The resulting object handle is not referenced anywhere.

**Parameters**

- **handle_transformations (bool)** – If True, the resulting form XObject’s /Matrix will be set to replicate rotation (/Rotate) and scaling (/UserUnit) in the page’s dictionary. In this way, the page’s transformations will be preserved when placing this object on another page.

```python
contents_coalesce (self: pikepdf.Page) -> None
```
Coalesce a page’s content streams.

A page’s content may be a stream or an array of streams. If this page’s content is an array, concatenate the streams into a single stream. This can be useful when working with files that split content streams in arbitrary spots, such as in the middle of a token, as that can confuse some software.

```python
externalize_inline_images (self: pikepdf.Page, min_size: int = 0) -> None
```
Convert inlines image to normal (external) images.

**Parameters**

- **min_size (int)** – minimum size in bytes

```python
get_filtered_contents (self: pikepdf.Page, if: TokenFilter) -> bytes
```
Apply a pikepdf.TokenFilter to a content stream, without modifying it.

This may be used when the results of a token filter do not need to be applied, such as when filtering is being used to retrieve information rather than edit the content stream.

Note that it is possible to create a subclassed TokenFilter that saves information of interest to its object attributes; it is not necessary to return data in the content stream.

To modify the content stream, use pikepdf.Page.add_content_token_filter().

**Returns**

- the modified content stream

**Return type**

- bytes

```python
obj
```
Get the underlying pikepdf.Object.

```python
remove_unreferenced_resources (self: pikepdf.Page) -> None
```
Removes from the resources dictionary any object not referenced in the content stream.

A page’s resources dictionary maps names to objects elsewhere in the file. This method walks through a page’s contents and keeps tracks of which resources are referenced somewhere in the contents. Then it removes from the resources dictionary any object that is not referenced in the contents. This method is used by page splitting code to avoid copying unused objects in files that used shared resource dictionaries across multiple pages.

```python
rotate (self: pikepdf.Page, angle: int, relative: bool) -> None
```
Rotate a page.

If relative is False, set the rotation of the page to angle. Otherwise, add angle to the rotation of the page. angle must be a multiple of 90. Adding 90 to the rotation rotates clockwise by 90 degrees.

```python
class pikepdf.PdfMatrix (*args)
```
Support class for PDF content stream matrices
PDF content stream matrices are 3x3 matrices summarized by a shorthand \((a, b, c, d, e, f)\) which correspond to the first two column vectors. The final column vector is always \((0, 0, 1)\) since this is using homogenous coordinates.

PDF uses row vectors. That is, \(vr @ A'\) gives the effect of transforming a row vector \(vr=(x, y, 1)\) by the matrix \(A'\). Most textbook treatments use \(A @ vc\) where the column vector \(vc=(x, y, 1)'\).

(@ is the Python matrix multiplication operator added in Python 3.5.)

Addition and other operations are not implemented because they’re not that meaningful in a PDF context (they can be defined and are mathematically meaningful in general).

PdfMatrix objects are immutable. All transformations on them produce a new matrix.

\[
\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}
\]

- \(a\)
- \(b\)
- \(c\)
- \(d\)
- \(e\)
- \(f\)

Return one of the six “active values” of the matrix.

- **encode()**
  - Encode this matrix in binary suitable for including in a PDF

- **static identity()**
  - Constructs and returns an identity matrix

- **rotated(angle_degrees_ccw)**
  - Concatenates a rotation matrix on this matrix

- **scaled(x, y)**
  - Concatenates a scaling matrix on this matrix

- **shorthand**
  - Return the 6-tuple \((a,b,c,d,e,f)\) that describes this matrix

- **translated(x, y)**
  - Translates this matrix

**class pikepdf.PdfImage(obj)**

Support class to provide a consistent API for manipulating PDF images

The data structure for images inside PDFs is irregular and flexible, making it difficult to work with without introducing errors for less typical cases. This class addresses these difficulties by providing a regular, Pythonic API similar in spirit (and convertible to) the Python Pillow imaging library.

- **as_pil_image()**
  - Extract the image as a Pillow Image, using decompression as necessary
    - **Returns** PIL.Image.Image

- **extract_to(**
  - Attempt to extract the image directly to a usable image file
    - If possible, the compressed data is extracted and inserted into a compressed image file format without transcoding the compressed content. If this is not possible, the data will be decompressed and extracted to an appropriate format.

\[1.3. \text{In use}\]
Because it is not known until attempted what image format will be extracted, users should not assume what format they are getting back. When saving the image to a file, use a temporary filename, and then rename the file to its final name based on the returned file extension.

**Examples**

```python
>>> im.extract_to(stream=bytes_io)
'.png'
```  
```python
>>> im.extract_to(fileprefix='/tmp/image00')
'/tmp/image00.jpg'
```

**Parameters**

- `stream` – Writable stream to write data to.
- `fileprefix` *(str or Path)* – The path to write the extracted image to, without the file extension.

**Returns** If `fileprefix` was provided, then the fileprefix with the appropriate extension. If no `fileprefix`, then an extension indicating the file type.

**Return type:** str

**get_stream_buffer()**

Access this image with the buffer protocol

**icc**

If an ICC profile is attached, return a Pillow object that describe it.

Most of the information may be found in `icc.profile`.

**Returns** PIL.ImageCms.ImageCmsProfile

**is_inline**

False for image XObject

**read_bytes()**

Decompress this image and return it as unencoded bytes

**show()**

Show the image however PIL wants to

**class pikepdf.PdfInlineImage(*, image_data, image_object: tuple)**

Support class for PDF inline images

**class pikepdf.models.PdfMetadata(pdf, pikepdf_mark=True, sync_docinfo=True)**

Read and edit the metadata associated with a PDF

The PDF specification contain two types of metadata, the newer XMP (Extensible Metadata Platform, XML-based) and older DocumentInformation dictionary. The PDF 2.0 specification removes the DocumentInformation dictionary.

This primarily works with XMP metadata, but includes methods to generate XMP from DocumentInformation and will also coordinate updates to DocumentInformation so that the two are kept consistent.

XMP metadata fields may be accessed using the full XML namespace URI or the short name. For example `metadata['dc:description']` and `metadata['http://purl.org/dc/elements/1.1/']description']` both refer to the same field. Several common XML namespaces are registered automatically.
See the XMP specification for details of allowable fields.
To update metadata, use a with block.

**Example**

```python
>>> with pdf.open_metadata() as records:
    records['dc:title'] = 'New Title'
```

See also:

pikepdf.Pdf.open_metadata()

load_from_docinfo (docinfo, delete_missing=False, raise_failure=False)
Populate the XMP metadata object with DocumentInfo

Parameters

- **docinfo** – a DocumentInfo, e.g. pdf.docinfo
- **delete_missing** – if the entry is not DocumentInfo, delete the equivalent from XMP
- **raise_failure** – if True, raise any failure to convert docinfo; otherwise warn and continue

A few entries in the deprecated DocumentInfo dictionary are considered approximately equivalent to certain XMP records. This method copies those entries into the XMP metadata.

pdfa_status
Returns the PDF/A conformance level claimed by this PDF, or False
A PDF may claim to PDF/A compliant without this being true. Use an independent verifier such as veraPDF to test if a PDF is truly conformant.

Returns The conformance level of the PDF/A, or an empty string if the PDF does not claim PDF/A conformance. Possible valid values are: 1A, 1B, 2A, 2B, 2U, 3A, 3B, 3U.

Return type str

pdfx_status
Returns the PDF/X conformance level claimed by this PDF, or False
A PDF may claim to PDF/X compliant without this being true. Use an independent verifier such as veraPDF to test if a PDF is truly conformant.

Returns The conformance level of the PDF/X, or an empty string if the PDF does not claim PDF/X conformance.

Return type str

**class** pikepdf.models.Encryption (*, owner, user, R=6, allow=Permissions(__class__=<class 'pikepdf.modelsencryption.Permissions'>), accessibility=True, extract=True, modify_annotation=True, modify_assembly=False, modify_form=True, modify_other=True, print_highres=True, print_lowres=True), aes=True, metadata=True)

Specify the encryption settings to apply when a PDF is saved.

Parameters

- **owner** (str) – The owner password to use. This allows full control of the file. If blank, the PDF will be encrypted and present as “(SECURED)” in PDF viewers. If the owner password is blank, the user password should be as well.
• **user** *(str)* – The user password to use. With this password, some restrictions will be imposed by a typical PDF reader. If blank, the PDF can be opened by anyone, but only modified as allowed by the permissions in *allow*.

• **R** *(int)* – Select the security handler algorithm to use. Choose from: 2, 3, 4 or 6. By default, the highest version of is selected (6). 5 is a deprecated algorithm that should not be used.

• **allow**(pikepdf.Permissions) – The permissions to set. If omitted, all permissions are granted to the user.

• **aes** *(bool)* – If True, request the AES algorithm. If False, use RC4. If omitted, AES is selected whenever possible (R >= 4).

• **metadata** *(bool)* – If True, also encrypt the PDF metadata. If False, metadata is not encrypted. Reading document metadata without decryption may be desirable in some cases. Requires aes=True. If omitted, metadata is encrypted whenever possible.

class pikepdf.Permissions(accessibility=True, extract=True, modify_annotation=True, modify_assembly=False, modify_form=True, modify_other=True, print_lowres=True, print_highres=True)

Stores the permissions for an encrypted PDF.

Unencrypted PDFs implicitly have all permissions allowed. pikepdf does not enforce the restrictions in any way. Permissions can only be changed when a PDF is saved.

  accessibility
  The owner of the PDF permission for screen readers and accessibility tools to access the PDF.

  extract
  The owner of the PDF permission for software to extract content from a PDF.

  modify_annotation

  modify_assembly

  modify_form

  modify_other
  The owner of the PDF permission to modify various parts of a PDF.

  print_lowres

  print_highres
  The owner of the PDF permission to print at low or high resolution.

class pikepdf.models.EncryptionMethod

Describes which encryption method was used on a particular part of a PDF. These values are returned by pikepdf.EncryptionInfo but are not currently used to specify how encryption is requested.

  none
  Data was not encrypted.

  unknown
  An unknown algorithm was used.

  rc4
  The RC4 encryption algorithm was used (obsolete).

  aes
  The AES-based algorithm was used as described in the PDF 1.7 reference manual.

  aesv3
  An improved version of the AES-based algorithm was used as described in the Adobe Supplement to the...
ISO 32000, requiring PDF 1.7 extension level 3. This algorithm still uses AES, but allows both AES-128 and AES-256, and improves how the key is derived from the password.

```python
class pikepdf.models.EncryptionInfo(enodict)
```

Reports encryption information for an encrypted PDF.

This information may not be changed, except when a PDF is saved. This object is not used to specify the encryption settings to save a PDF, due to non-overlapping information requirements.

- **P**: Encoded permission bits.
  - See `Pdf.allow()` instead.
- **R**: Revision number of the security handler.
- **V**: Version of PDF password algorithm.

- **bits**: The number of encryption bits.
- **encryption_key**: The RC4 or AES encryption key used for this file.
- **file_method**: Encryption method used to encode the whole file.
- **stream_method**: Encryption method used to encode streams.
- **string_method**: Encryption method used to encode strings.
- **user_password**: If possible, return the user password.
  - The user password can only be retrieved when a PDF is opened with the owner password and when older versions of the encryption algorithm are used.
  - The password is always returned as `bytes` even if it has a clear Unicode representation.

### 1.3.14 Content streams

In PDF, drawing operations are all performed in content streams that describe the positioning and drawing order of all graphics (including text, images and vector drawing).

pikepdf (and libqpdf) provide two tools for interpreting content streams: a parser and filter. The parser returns higher level information, conveniently grouping all commands with their operands. The parser is useful when one wants to retrieve information from a content stream, such as determine the position of an element. The parser should not be used to edit or reconstruct the content stream because some subtleties are lost in parsing.

The token filter works at a lower level, considering each token including comments, and distinguishing different types of spaces. This allows modifying content streams. A TokenFilter must be subclassed; the specialized version describes how it should transform the stream of tokens.

```python
pikepdf.parse_content_stream(page_or_stream, operators=````)
```

Parse a PDF content stream into a sequence of instructions.

A PDF content stream is list of instructions that describe where to render the text and graphics in a PDF. This is the starting point for analyzing PDFs.
If the input is a page and page.Contents is an array, then the content stream is automatically treated as one coalesced stream.

Each instruction contains at least one operator and zero or more operands.

Parameters

- **page_or_stream**: A page object, or the content stream attached to another object such as a Form XObject.
- **operators**: A space-separated string of operators to whitelist. For example ‘q Q cm Do’ will return only operators that pertain to drawing images. Use ‘BI ID EI’ for inline images. All other operators and associated tokens are ignored. If blank, all tokens are accepted.

Returns

List of (operands, command) tuples where command is an operator (str) and operands is a tuple of str; the PDF drawing command and the command’s operands, respectively.

Return type: list

Example

```python
>>> pdf = pikepdf.Pdf.open(input_pdf)
>>> page = pdf.pages[0]
>>> for operands, command in parse_content_stream(page):
...     print(command)
```

```python
class pikepdf.Token
```

- **raw_value**: The binary representation of a token.

  Return type: bytes

- **type_**: Returns the type of token.

  Return type: pikepdf.TokenType

- **value**: Interprets the token as a string.

  Return type: str or bytes

```python
class pikepdf.TokenType
```

When filtering content streams, each token is labeled according to the role in plays.

Standard tokens

- **array_open**
- **array_close**
- **brace_open**
- **brace_close**
- **dict_open**
dict_close
These tokens mark the start and end of an array, text string, and dictionary, respectively.

integer
real
null
bool
The token data represents an integer, real number, null or boolean, respectively.

Name
The token is the name of an object. In practice, these are among the most interesting tokens.

inline_image
An inline image in the content stream. The whole inline image is represented by the single token.

Lexical tokens

comment
Signifies a comment that appears in the content stream.

word
Otherwise uncategorized bytes are returned as word tokens. PDF operators are words.

bad
An invalid token.

space
Whitespace within the content stream.

eof
Denotes the end of the tokens in this content stream.

class pikepdf.TokenFilter

Handle a pikepdf.Token.

This is an abstract method that must be defined in a subclass of TokenFilter. The method will be
called for each token. The implementation may return either None to discard the token, the original token
to include it, a new token, or an iterable containing zero or more tokens. An implementation may also
buffer tokens and release them in groups (for example, it could collect an entire PDF command with all of
its operands, and then return all of it).

The final token will always be a token of type TokenType.eof, (unless an exception is raised).

If this method raises an exception, the exception will be caught by C++, consumed, and repalced with a
less informative exception. Use pikepdf.Pdf.get_warnings() to view the original.

Return type: None or list or pikepdf.Token

1.3.15 Architecture

pikepdf uses pybind11 to bind the C++ interface of QPDF. pybind11 was selected after evaluating Cython, CFFI and
SWIG as possible binding solutions.

In addition to bindings pikepdf includes support code written in a mix of C++ and Python, mainly to present a clean
Pythonic interface to C++ and implement higher level functionality.
Internals

Internally the package presents a module named `pikepdf` from which objects can be imported. The C++ extension module is currently named `pikepdf._qpdf`. Users of `pikepdf` should not directly access `_qpdf` since it is an internal interface.

In general, modules or objects behind an underscore are private (although they may be returned in some situations).

Thread safety

Because of the global interpreter lock (GIL), it is safe to read pikepdf objects across Python threads. Also because of the GIL, there may not be much performance gain from doing so.

If one or more threads will be modifying pikepdf objects, you will have to coordinate read and write access with a `threading.Lock`.

It is not currently possible to pickle pikepdf objects or marshall them across process boundaries (as would be required to use pikepdf in `multiprocessing`). If this were implemented, it would not be much more efficient than saving a full PDF and sending it to another process.

File handles

Because of technical limitations in underlying libraries, pikepdf keeps the source PDF file open when a content is copied from it to another PDF, even when all Python variables pointing to the source are removed. If a PDF is being assembled from many sources, then all of those sources are held open in memory.

1.3.16 Resources

- QPDF Manual
- PDF 1.7 ISO Specification PDF 32000-1:2008
- Adobe Supplement to ISO 32000 BaseVersion 1.7 ExtensionLevel 3, Adobe Acrobat 9.0, June 2008, for AESv3
- Other Adobe extensions to the PDF specification

References
### Index

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>new</strong> (pikepdf.Array static method), 40</td>
<td>contents_coalesce () (pikepdf.Page method), 42</td>
</tr>
<tr>
<td><strong>new</strong> (pikepdf.Dictionary static method), 40</td>
<td>copy_foreign () (pikepdf.Pdf method), 31</td>
</tr>
<tr>
<td><strong>new</strong> (pikepdf.Name static method), 39</td>
<td>D</td>
</tr>
<tr>
<td><strong>new</strong> (pikepdf.Stream static method), 40</td>
<td>d (pikepdf.PdfMatrix attribute), 43</td>
</tr>
<tr>
<td><strong>new</strong> (pikepdf.String static method), 40</td>
<td>dict_close (pikepdf.TokenType attribute), 48</td>
</tr>
<tr>
<td>A</td>
<td>dict_open (pikepdf.TokenType attribute), 48</td>
</tr>
<tr>
<td>a (pikepdf.PdfMatrix attribute), 43</td>
<td>Dictionary (class in pikepdf), 40</td>
</tr>
<tr>
<td>accessibility (pikepdf.Permissions attribute), 46</td>
<td>disable (pikepdf.ObjectStreamMode attribute), 36</td>
</tr>
<tr>
<td>add_blank_page() (pikepdf.Pdf method), 30</td>
<td>docinfo (pikepdf.Pdf attribute), 31</td>
</tr>
<tr>
<td>add_content_token_filter() (pikepdf.Page method), 41</td>
<td>E</td>
</tr>
<tr>
<td>Array (class in pikepdf), 40</td>
<td>e (pikepdf.PdfMatrix attribute), 43</td>
</tr>
<tr>
<td>array_close (pikepdf.TokenType attribute), 48</td>
<td>emplace () (pikepdf.Object method), 37</td>
</tr>
<tr>
<td>array_open (pikepdf.TokenType attribute), 48</td>
<td>encode () (pikepdf.PdfMatrix method), 43</td>
</tr>
<tr>
<td>as_dict () (pikepdf.Object method), 37</td>
<td>Encryption (class in pikepdf), 36</td>
</tr>
<tr>
<td>as_list () (pikepdf.Object method), 37</td>
<td>encryption (pikepdf.Pdf attribute), 31</td>
</tr>
<tr>
<td>as_pil_image () (pikepdf.PdfImage method), 43</td>
<td>encryption_key (pikepdf.models.EncryptionInfo attribute), 47</td>
</tr>
<tr>
<td>B</td>
<td>EncryptionInfo (class in pikepdf.models), 47</td>
</tr>
<tr>
<td>b (pikepdf.PdfMatrix attribute), 43</td>
<td>eof (pikepdf.TokenType attribute), 49</td>
</tr>
<tr>
<td>bad (pikepdf.TokenType attribute), 49</td>
<td>extend () (pikepdf._qpdf.PageList method), 41</td>
</tr>
<tr>
<td>bits (pikepdf.models.EncryptionInfo attribute), 47</td>
<td>externalize_inline_images () (pikepdf.Page method), 42</td>
</tr>
<tr>
<td>bool (pikepdf.TokenType attribute), 49</td>
<td>extract () (pikepdf.Permissions attribute), 46</td>
</tr>
<tr>
<td>brace_close (pikepdf.TokenType attribute), 48</td>
<td>extract_to () (pikepdf.PdfImage method), 43</td>
</tr>
<tr>
<td>brace_open (pikepdf.TokenType attribute), 48</td>
<td>F</td>
</tr>
<tr>
<td>C</td>
<td>f (pikepdf.PdfMatrix attribute), 43</td>
</tr>
<tr>
<td>c (pikepdf.PdfMatrix attribute), 43</td>
<td>file_method (pikepdf.models.EncryptionInfo attribute), 47</td>
</tr>
<tr>
<td>check_linearization () (pikepdf.Pdf method), 31</td>
<td>filename (pikepdf.Pdf attribute), 31</td>
</tr>
<tr>
<td>close () (pikepdf.Pdf method), 31</td>
<td>G</td>
</tr>
<tr>
<td>comment (pikepdf.TokenType attribute), 49</td>
<td>generalized (pikepdf.StreamDecodeLevel attribute), 36</td>
</tr>
<tr>
<td>contents_coalesce () (pikepdf.Page method), 42</td>
<td>generate (pikepdf.ObjectStreamMode attribute), 36</td>
</tr>
<tr>
<td>copy_foreign () (pikepdf.Pdf method), 31</td>
<td>get () (pikepdf.Object method), 37</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>array_close</td>
<td>pikepdf.TokenType attribute, 49</td>
</tr>
<tr>
<td>array_open</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>as_dict</td>
<td>pikepdf.Object method, 37</td>
</tr>
<tr>
<td>as_list</td>
<td>pikepdf.Object method, 37</td>
</tr>
<tr>
<td>as_pil_image</td>
<td>pikepdf.PdfImage method, 43</td>
</tr>
<tr>
<td>B</td>
<td>pikepdf.PdfMatrix attribute, 43</td>
</tr>
<tr>
<td>b</td>
<td>pikepdf.TokenType attribute, 49</td>
</tr>
<tr>
<td>bits</td>
<td>pikepdf.models.EncryptionInfo attribute, 47</td>
</tr>
<tr>
<td>bool</td>
<td>pikepdf.TokenType attribute, 49</td>
</tr>
<tr>
<td>brace_close</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>brace_open</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>c</td>
<td>pikepdf.PdfMatrix attribute, 43</td>
</tr>
<tr>
<td>check_linearization</td>
<td>pikepdf.Pdf method, 31</td>
</tr>
<tr>
<td>close</td>
<td>pikepdf.Pdf method, 31</td>
</tr>
<tr>
<td>comment</td>
<td>pikepdf.TokenType attribute, 49</td>
</tr>
<tr>
<td>D</td>
<td>pikepdf.PdfMatrix attribute, 43</td>
</tr>
<tr>
<td>d</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>dict_close</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>dict_open</td>
<td>pikepdf.TokenType attribute, 48</td>
</tr>
<tr>
<td>Dictionary</td>
<td>pikepdf.Pdf attribute, 31</td>
</tr>
<tr>
<td>disable</td>
<td>pikepdf.ObjectStreamMode attribute, 36</td>
</tr>
<tr>
<td>docinfo</td>
<td>pikepdf.Pdf attribute, 31</td>
</tr>
<tr>
<td>E</td>
<td>pikepdf.PdfMatrix attribute, 43</td>
</tr>
<tr>
<td>e</td>
<td>pikepdf.Object method, 37</td>
</tr>
<tr>
<td>encode</td>
<td>pikepdf.PdfMatrix method, 43</td>
</tr>
<tr>
<td>Encryption</td>
<td>pikepdf.Pdf attribute, 31</td>
</tr>
<tr>
<td>encryption</td>
<td>pikepdf.models.EncryptionInfo attribute, 47</td>
</tr>
<tr>
<td>encryption_key</td>
<td>pikepdf.models.EncryptionInfo attribute, 47</td>
</tr>
<tr>
<td>EncryptionInfo</td>
<td>pikepdf.models.EncryptionInfo attribute, 47</td>
</tr>
<tr>
<td>eof</td>
<td>pikepdf.TokenType attribute, 49</td>
</tr>
<tr>
<td>extend</td>
<td>pikepdf._qpdf.PageList method, 41</td>
</tr>
<tr>
<td>externalize_inline_images</td>
<td>pikepdf.Page method, 42</td>
</tr>
<tr>
<td>extract</td>
<td>pikepdf.Permissions attribute, 46</td>
</tr>
<tr>
<td>extract_to</td>
<td>pikepdf.PdfImage method, 43</td>
</tr>
<tr>
<td>F</td>
<td>pikepdf.PdfMatrix attribute, 43</td>
</tr>
<tr>
<td>f</td>
<td>pikepdf.ObjectStreamMode attribute, 36</td>
</tr>
<tr>
<td>file_method</td>
<td>pikepdf.models.EncryptionInfo attribute, 47</td>
</tr>
<tr>
<td>filename</td>
<td>pikepdf.Pdf attribute, 31</td>
</tr>
<tr>
<td>G</td>
<td>pikepdf.StreamDecodeLevel attribute, 36</td>
</tr>
<tr>
<td>generate</td>
<td>pikepdf.ObjectStreamMode attribute, 36</td>
</tr>
<tr>
<td>get</td>
<td>pikepdf.Object method, 37</td>
</tr>
</tbody>
</table>
root (pikepdf.Pdf attribute), 34
rotate() (pikepdf.Page method), 42
rotated() (pikepdf.PdfMatrix method), 43

S
save() (pikepdf.Pdf method), 34
scaled() (pikepdf.PdfMatrix method), 43
shorthand (pikepdf.PdfMatrix attribute), 43
show() (pikepdf.PdfImage method), 44
show_xref_table() (pikepdf.Pdf method), 35
space (pikepdf.TokenType attribute), 49
specialized (pikepdf.StreamDecodeLevel attribute), 36
Stream (class in pikepdf), 40
stream_method (pikepdf.models.EncryptionInfo attribute), 47
String (class in pikepdf), 40
string_method (pikepdf.models.EncryptionInfo attribute), 47

T
to_json() (pikepdf.Object method), 38
Token (class in pikepdf), 48
TokenFilter (class in pikepdf), 49
trailer (pikepdf.Pdf attribute), 35
translated() (pikepdf.PdfMatrix method), 43
type_ (pikepdf.Token attribute), 48

U
unknown (pikepdf.models.EncryptionMethod attribute), 46
unparse() (pikepdf.Object method), 39
user_password (pikepdf.models.EncryptionInfo attribute), 47

V
V (pikepdf.models.EncryptionInfo attribute), 47
value (pikepdf.Token attribute), 48

W
word (pikepdf.TokenType attribute), 49
wrap_in_array() (pikepdf.Object method), 39
write() (pikepdf.Object method), 39