
Chapter 6. QPDF JSON

6.1. Overview

Beginning with qpdf version 8.3.0, the **qpdf** command-line program can produce a JSON representation of the non-content data in a PDF file. It includes a dump in JSON format of all objects in the PDF file excluding the content of streams. This JSON representation makes it very easy to look in detail at the structure of a given PDF file, and it also provides a great way to work with PDF files programmatically from the command-line in languages that can't call or link with the qpdf library directly. Note that stream data can be extracted from PDF files using other qpdf command-line options.

6.2. JSON Guarantees

The qpdf JSON representation includes a JSON serialization of the raw objects in the PDF file as well as some computed information in a more easily extracted format. QPDF provides some guarantees about its JSON format. These guarantees are designed to simplify the experience of a developer working with the JSON format.

Compatibility

The top-level JSON object output is a dictionary. The JSON output contains various nested dictionaries and arrays. With the exception of dictionaries that are populated by the fields of objects from the file, all instances of a dictionary are guaranteed to have exactly the same keys. Future versions of qpdf are free to add additional keys but not to remove keys or change the type of object that a key points to. The qpdf program validates this guarantee, and in the unlikely event that a bug in qpdf should cause it to generate data that doesn't conform to this rule, it will ask you to file a bug report.

The top-level JSON structure contains a “version” key whose value is simple integer. The value of the `version` key will be incremented if a non-compatible change is made. A non-compatible change would be any change that involves removal of a key, a change to the format of data pointed to by a key, or a semantic change that requires a different interpretation of a previously existing key. A strong effort will be made to avoid breaking compatibility.

Documentation

The **qpdf** command can be invoked with the **--json-help** option. This will output a JSON structure that has the same structure as the JSON output that qpdf generates, except that each field in the help output is a description of the corresponding field in the JSON output. The specific guarantees are as follows:

- A dictionary in the help output means that the corresponding location in the actual JSON output is also a dictionary with exactly the same keys; that is, no keys present in help are absent in the real output, and no keys will be present in the real output that are not in help.
- A string in the help output is a description of the item that appears in the corresponding location of the actual output. The corresponding output can have any format.
- An array in the help output always contains a single element. It indicates that the corresponding location in the actual output is also an array, and that each element of the array has whatever format is implied by the single element of the help output's array.

For example, the help output indicates includes a “pagelabels” key whose value is an array of one element. That element is a dictionary with keys “index” and “label”. In addition to describing the meaning of those keys, this tells you that the actual JSON output will contain a `pagelabels` array, each of whose elements is a dictionary that contains an `index` key, a `label` key, and no other keys.