



UNITED STATES  
DEPARTMENT OF VETERANS AFFAIRS

Department of Veterans Affairs  
Product Development (PD)

Veterans Benefits Management System (VBMS)

---

PDF Specification Document

Release 14.0  
SPC-1123-vbms



Task Order 0003, Core/Correspondence Consolidation

T4NG Prime Contract No: VA118-16-D-1007

(PWS 5.2.2.d)

Version 6.0

November 2017

Submitted to:

Department of Veterans Affairs  
Ryan VanVickle  
ryan.vanvickle@va.gov  
(202) 280-8409

Submitted by:

Booz Allen Hamilton Inc.  
575 Herndon Parkway  
Herndon, VA 20170  
(571) 346-4000  
(FAX) 346-4010

Booz | Allen | Hamilton

## REVISION HISTORY

Date	Version	Description	Author
10 November 2017	6.0	Release 14.0 – Removed a bullet item from <u>Section 4.1</u> .	Booz Allen Hamilton
18 August 2017	5.0	Release 13.1 – Minor editorial updates throughout, no content updates	Booz Allen Hamilton
26 May 2017	4.0	Release 13.0 – Administrative change to use SPC in the document number, no content updates	Booz Allen Hamilton
04 January 2017	3.0	Removed Mission Statement and updated template	Booz Allen Hamilton
10 July 2013	2.1	Updated Mission Statement, Updated Document Number	Booz Allen Hamilton
25 October 2012	2.0	Release to VA	Booz Allen Hamilton
13 April 2012	1.0	Release to VA	Booz Allen Hamilton
4 April 2012	Draft	Initial Version	Booz Allen Hamilton

## CONTENTS

<b>1</b>	<b>DOCUMENT PURPOSE .....</b>	<b>3</b>
1.1	HISTORY AND GOALS.....	3
	<i>Figure 1: PDF Specification Goals.....</i>	<i>4</i>
<b>2</b>	<b>ISO STANDARDS .....</b>	<b>4</b>
2.1	ISO 19005-1 (PDF/A-1).....	4
2.2	ISO 19005-2 (PDF/A-2).....	6
2.3	ISO 24517-1 (PDF/E) .....	7
2.4	PDF HEALTHCARE (PDF/H).....	7
2.5	ISO 14289 (PDF/UA) .....	8
2.6	ISO 32000-1.....	8
<b>3</b>	<b>METADATA .....</b>	<b>9</b>
3.1	INFO DICTIONARY .....	9
3.2	EXTENSIBLE METADATA PLATFORM.....	9
<b>4</b>	<b>VBMS PDF DOCUMENT SPECIFICATION .....</b>	<b>10</b>
4.1	REQUIREMENTS.....	10
4.2	METADATA.....	12
	<i>Table 1: Recommended PDF Metadata.....</i>	<i>12</i>

# 1 Document Purpose

---

The purpose of this document is to create the specification for PDF document generation within VBMS. The guidance provided in the *VBMS SOA Reference Architecture* document outlines the patterns and practices for documenting a specification.

The keywords “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” in this specification are to be interpreted in this specification as described in IETF RFC 2119. These keywords are capitalized when used to unambiguously specify requirements over protocol, application features, and behaviors that affect the interoperability and security of implementations. When these words are not capitalized, they are meant in their natural-language sense.

The following source provides additional information about IETF RFC 2119:

- <http://www.ietf.org/rfc/rfc2119.txt>

This specification is broken down into the following sections:

- History and goals of the PDF specification
- Relevant ISO standards on which the VBMS PDF Document Specification is built
- Background discussion on relevant PDF document metadata standards
- VBMS PDF Document Specification

## 1.1 History and Goals

The goal of a PDF is to enable users to exchange and view electronic documents easily and reliably, independent of the environment in which they were created, viewed, or printed. At the core of PDF is an advanced imaging model derived from the PostScript® page description language. This PDF Imaging Model enables the description of text and graphics in a device-independent and resolution-independent manner.

A PDF document consists of a collection of objects that together describe the appearance of one or more pages, possibly accompanied by additional interactive elements and higher-level application data. A PDF file contains the objects making up a PDF document along with associated structural information, all represented as a single self-contained sequence of bytes.

A document's pages and other visual elements can contain any combination of text, graphics, and images. A page's appearance is described by a PDF content stream, which contains a sequence of graphics objects to be painted on the page. This appearance is fully specified. All layout and formatting decisions have already been made by the application generating the content stream.

In addition to describing the static appearance of pages, a PDF document can contain interactive elements that are possible only in an electronic representation. PDF supports annotations of

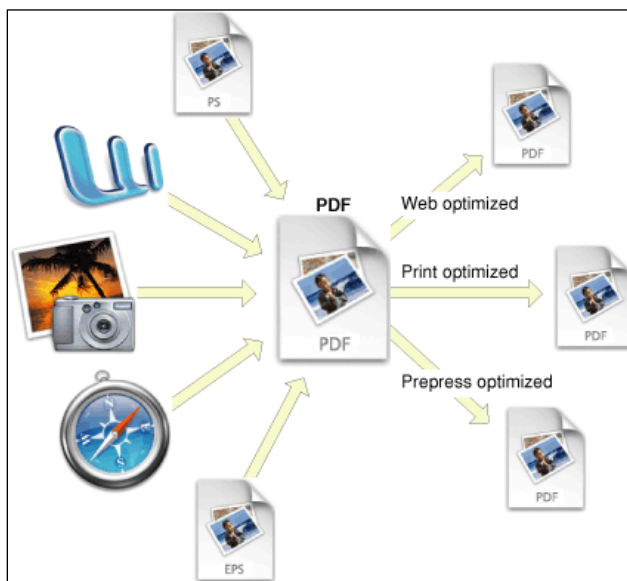
many kinds for such things as text notes, hypertext links, markup, file attachments, sounds, and movies. A document can define its own user interface. Keyboard and mouse inputs can trigger actions that are specified by PDF objects.

The document can contain interactive form fields to be filled in by the user and can export the values of these fields to or import them from other applications.

Finally, a PDF document can contain higher-level information that is useful for interchange of content among applications. In addition to specifying appearance, a document's content can include identification and logical structure information that allows it to be searched, edited, or extracted for reuse elsewhere.

Figure 1 depicts the goal of the VBMS PDF specification, which is to convert all relevant media types into a file format that is optimized for current and future usages, in expected scanning and rendering scenarios.

**Figure 1: PDF Specification Goals**



The following source provides additional information about PDF:

- PDF Reference, sixth edition, Adobe® Portable Document Format, Version 1.7, November 2006 [http://www.adobe.com/devnet/acrobat/pdfs/pdf\\_reference\\_1-7.pdf](http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference_1-7.pdf)

## 2 ISO Standards

### 2.1 ISO 19005-1 (PDF/A-1)

The ISO 19005-1 standard for PDF/A-1 was published in 2005 and last reviewed and confirmed in 2015. It is geared towards long-term preservation. It provides a mechanism for representing electronic documents that preserves their visual appearance over time, independent of the tools and systems used for creating, storing, and rendering the files.

It aims to support future access and future migration needs by providing frameworks for the following:

- Embedding metadata about electronic documents
- Defining the logical structure and semantic properties of electronic documents

The PDF/A-1 specifications are based on the specifications of PDF 1.4 and describes which aspects of a PDF are compulsory, optional, or prohibited in a PDF/A-1 compliant file.

PDF/A-1 defines two conformance levels:

- A “Level A” conformant file (PDF/A-1a) SHALL adhere to all requirements of the ISO standard.
- A “Level B” conformant file (PDF/A-1b) does not have to adhere to two requirements: the use of tagging to specify the document’s logical structure and reading order and the use of Unicode character maps that map character codes to Unicode values.

As preserving the natural reading order and using Unicode are crucial for the preservation of the document as it was intended by the creator, the VBMS PDF specification will be a superset of PDF/A-1a conformance.

PDF/A-1 files MUST include:

- Embedded fonts
- Device-independent color
- XMP metadata (where applicable)

PDF/A-1 files MAY NOT include:

- Encryption
- LZW compression
- Embedded files
- External content references
- PDF transparency
- Multi-media
- JavaScript

The following source provides additional information about ISO:

- 19005-1. PDF ISO 19005-1, December 2004 <https://www.iso.org/standard/38920.html>

## 2.2 ISO 19005-2 (PDF/A-2)

In July 2011, the ISO released the PDF/A-2 standard. Where PDF/A-1 is based on PDF version 1.4, PDF/A-2 takes advantage of features that only became available in later versions of PDF, up to and including PDF version 1.7. But most importantly, PDF/A-2 is no longer based on a particular Adobe PDF version, but instead, is now based on an ISO standard 32000-1.

PDF/A-2 introduced a number of features:

- **JPEG2000 Compression** – The JPEG2000 compression was introduced with the PDF 1.5 specification, which was past the release time of the PDF/A-1 standard. Adding the JPEG2000 compression benefits particularly scanned documents such as maps, books, and documents with color content such as checks or passports.
- **Embedded PDF/A Files via Collections** – Acrobat allows users to create collections (sometimes also referred to as “portfolios”) where multiple PDF/A documents are combined into one “container PDF” document. A possible use of a PDF/A collection is for instance the archival of emails where email attachments can be converted to PDF/A and stored as “collections” inside a converted PDF/A email text body. PDF/A collections can also benefit security applications where a signature can be applied to individual single pages. The PDF/A collection then combines the signed single page. Individual pages can subsequently be removed without affecting the validity of the signatures of the remaining pages.
- **Transparency** – Although transparency is part of PDF 1.4, at the time of the PDF/A-1 standard release, it was not defined well enough to be included in the PDF/A-1 standard. The specification has substantially matured since then, and transparency has become a common characteristic of PDF documents. Transparency is often found in the form of drop shadows, cross fades, and highlight mark-ups.
- **Optional Content (Layers)** – Optional content, sometimes also referred to as layers, is useful for mapping applications or engineering drawings where individual layers can show or hide according to the information requirements of the viewing person. Another area of use is in user manuals of products that are sold internationally – where different languages can be implemented on different layers.
- **New Conformance Level PDF/A-2u** – “u” for Unicode – PDF/A-1b and PDF/A-2b concentrate on visual integrity, where “b” stands for “basic”. PDF/A-1a and PDF/A-2a concentrate on accessibility – hence the “a” notation. New to PDF/A-2 is the conformance level.
- **PDF/A-2u (“u” for “Unicode”)** – It simplifies the text searching and copying of Unicode text for digitally created PDF documents and PDF documents that were scanned with subsequent optical character recognition (OCR).

- Object Level XMP Metadata – PDF/A-2 specifies the requirements for custom XMP metadata.
- Comment Types and Annotations – Some of the newer comment types were added to the list of prohibited annotation types, and at the same time, some of the newer comment types such as text editing comments are now acceptable to the PDF/A-2 standard.
- Digital Signatures – While PDF/A-1 already allows for digital signatures, PDF/A-2 defines the rules that need to be applied to guarantee interoperability.

The following source provides additional information about ISO 19005-2:

- PDF ISO 19005-2, July 2011 <https://www.iso.org/standard/50655.html>

## 2.3 ISO 24517-1 (PDF/E)

ISO 24517-1 was published in 2008 and defines a format for the creation of documents used in engineering workflows and is based on the PDF version 1.6. It leverages U3D, another open standard, for the representation of 3D content.

## 2.4 PDF Healthcare (PDF/H)

PDF Healthcare is a Best Practices Guide (BPG) that is supplemented by an Implementation Guide (IG). The PDF Healthcare BPG and the IG are based on open, published specifications with direction specific to the healthcare industry. The BPG and IG describe attributes to facilitate the capture, exchange, preservation, and protection of healthcare information. Such attributes include the ability for health care providers and consumers to develop a secure, electronic container that stores and transmits relevant healthcare information, important for maintaining and improving health. The healthcare information can include but is not limited to personal, handwritten documents, (structured or unstructured) clinical notes, (structured) laboratory test result reports, (unstructured) word-processed/text summary reports, electronic forms, scanned document images, digital diagnostic images, photographs, and signal tracings (e.g., electrocardiograms [ECGs]).

PDF Healthcare is NOT a proposed standard. The PDF Healthcare BPG and IG are intended to be used to guide the generation and consumption of secure and portable containers of personal health information and electronic health record information rather than replacing existing standards or adding new standards for health care information interoperability.

The following source provides additional information about PDF/H:

- [http://www.aiim.org/Resources/Standards/AIIM\\_BP\\_Healthcare](http://www.aiim.org/Resources/Standards/AIIM_BP_Healthcare)



## 2.5 ISO 14289 (PDF/UA)

PDF/Universal Accessibility is the International Standard for accessible PDF documents. The mission of PDF/UA is to develop technical and other standards for the authoring, remediation, and validation of PDF content to ensure accessibility for people that use assistive technology such as screen readers for users who are blind.

PDF documents must comply with the checkpoints specified in § 1194.22 (Web-based Intranet and Internet Information and Applications).

Constraints mandated by Section 508 are as follows:

- All images **MUST** have alternate text.
- Pages designed to convey information via color **MUST** be adequately tagged.
- All low-contrast (shaded) text **MUST** be rendered as image (with tags) or removed.
- Row and column headers **MUST** be included for data tables. Multi-level headings for row/columns or spanning cells **SHOULD** be avoided.

The following source provides additional information about PDF and 508 Compliance:

- <http://www.access-board.gov/sec508/guide/1194.22.htm>

## 2.6 ISO 32000-1

ISO 32000 specifies a digital form for representing PDF documents. PDF was developed and specified by Adobe Systems Incorporated beginning in 1993 and continuing until 2007 when the ISO standard was prepared. The Adobe Systems version PDF 1.7 is the basis for the ISO 32000 edition. The specifications for PDF are backward inclusive, meaning that PDF 1.7 includes all of the functionality previously documented in the Adobe PDF Specifications for versions 1.0 through 1.6.

PDF, together with software for creating, viewing, printing, and processing PDF files in a variety of ways, fulfills a set of requirements for electronic documents including:

- Preservation of document fidelity independent of the device, platform, and software
- Merging of content from diverse sources, such as web sites, word processing and spreadsheet programs, scanned documents, photos, and graphics, into one self-contained document while maintaining the integrity of all original source documents
- Collaborative editing of documents from multiple locations or platforms
- Digital signatures to certify authenticity
- Security and permissions to allow the creator to retain control of the document and associated rights
- Accessibility of content to those with disabilities



- Extraction and reuse of content for use with other file formats and applications
- Electronic forms to gather data and integrate it with business systems

This standard does not specify:

- Specific processes for converting paper or electronic documents to the PDF format
- Specific technical design, user interface or implementation or operational details of rendering
- Specific physical methods of storing these documents such as media and storage conditions
- Methods for validating the conformance of PDF files or readers

The following source provides additional information about ISO 32000:

- PDF ISO 32000, January 2008,  
[http://www.adobe.com/devnet/acrobat/pdfs/PDF32000\\_2008.pdf](http://www.adobe.com/devnet/acrobat/pdfs/PDF32000_2008.pdf)

## 3 Metadata

---

### 3.1 Info Dictionary

The Info dictionary has been included in PDF since version 1.0. The optional Info entry in the trailer of the file can hold a document information dictionary containing metadata for the document. Any entry whose value is not known should be omitted from the dictionary rather than included with an empty string as its value. It contains a set of document info entries, which are simple pairs of data that consist of a key and a matching value. Applications can add their own sets of data to the info dictionary.

### 3.2 Extensible Metadata Platform

XMP (Extensible Metadata Platform) is an Adobe technology for embedding metadata into files. It can be used with a wide variety of data files. With Acrobat 5 and PDF 1.4 (2001) this mechanism was also made available for PDF files. XMP is more powerful than the info dictionary, which is why it is used in a number of PDF-based metadata standards.

XMP metadata travel with files and can be embedded in many common file formats including PDF, TIFF, and JPEG. Metadata properties are grouped in XML schemas. Each schema is identified by a unique namespace URI and holds an arbitrary number of properties. The XMP specification includes more than a dozen predefined schemas with hundreds of properties for common document and image characteristics. The most widely used predefined XMP schema is called the Dublin Core, or dc. It includes general properties such as Title, Creator, Subject, and Description. In addition to predefined schemas, custom schemas can be defined to cover specific metadata requirements.

The following sources provide additional information about XMP and Dublin Core:

- XMP Specification Part 1, Data Model, Serialization, And Core Properties, July 2010
- XMP Specification Part 2, Additional Properties, July 2010
- XMP Specification Part 3, Storage In Files, July 2010
- ISO 16684-1, February 2012
- <http://dublincore.org/>

Many libraries/tools exist to facilitate the setting/getting of XMP metadata content. The following source provides additional information about XMP toolkits:

- [http://en.wikipedia.org/wiki/Extensible\\_Metadata\\_Platform](http://en.wikipedia.org/wiki/Extensible_Metadata_Platform)

## 4 VBMS PDF Document Specification

---

This section describes the specifications each PDF placed in the VBMS system must adhere to in order to be in compliance. The source of the requirement is listed in parentheses after the requirement.

### 4.1 Requirements

To comply with the VBMS PDF specification, each PDF must:

- Be PDF 1.4 or later in PDF Normal or PDF Searchable Image formats only. PDF Image Only formats are not acceptable. (ISO 32000)
- Be scanned at a resolution of 300 dots per inch (dpi) to ensure that the pages of the document are legible both on the computer screen and when printed, while at the same time minimizing the file size. (VBMS Constraint)
- Not be down-sampled, which is a process of decreasing the number of pixels in the image. Down-sampling is an option in PDF optimizing but can lead to poor quality images. (VBMS Constraint)
- Be de-speckled, by removing isolated “dots” within the image which can cause recognition problems, making the result image cleaner. (VBMS Constraint)
- Be de-skewed, improving OCR results by straightening crooked pages. (VBMS Constraint)
- Have all fonts used in the document embedded in the document. Doing so improves text searching in the PDF. Some TrueType fonts have a setting added by the font designer that prevents the font from being embedded and should be avoided. (ISO 19005)
- Be Section 508 compliant PDFs. Specific examples of 508 compliance are as follows:
  - A tagged PDF. Tags describe logical structural aspects (paragraphs, lists, tables, links, illustrations, etc.) and are used by rendering applications and screen-reader devices. (ISO 19005 and ISO 14289)

- Alternative text descriptions for all non-text document elements (e.g., images that are part of the document's content). This does not apply to images that provide decorative elements to the document. (ISO 14289)
  - For full Section 508 information and constraints please see:  
<https://www.section508.gov/>.
- Contain no security restrictions (password, certificate, etc.). These restrictions could hamper future use or transformations of the documents, and can interfere with screen readers. (ISO 19005)
- Not contain any bookmarks or links. (ISO 19005)
- Not contain crop marks, registration marks, date stamp, time stamp, or any other mark that does not appear in the original document. (ISO 19005)
- Specify color spaces in a device-independent manner. The ISO 19005 specification details which device and color spaces can be used in a PDF/A document subject to the choice of output intent. A PDF/A document may contain up to one unique PDF/A output intent which must describe either a Grayscale, RGB, or CMYK color space. The easiest approach which will work in many situations is to use an RGB output intent ICC profile (sRGB is RECOMMENDED), since most color spaces except DeviceCMYK can be used in the document. If PDF annotations specify colors in a C or IC (Interior Color) entry (e.g., border color for a link annotation), the colors should be specified in the DeviceRGB color space; PDF 1.4 does not support device-independent color specifications for annotations. If the document contains colorized annotations, an RGB output intent is required. The v4 (newer) version of the sRGB ICC profile SHOULD be used over the v2 profile. (ISO 19005)
- Be created from source documents using the “Optimize the PDF for fast web view” option to reduce file sizes and file opening times. (ISO 19005)
- Only use image compression algorithms supported by the VBMS document viewer. RECOMMENDED algorithms are as follows:
  - CCITT Group 4 for bitonal (black and white) images. (ISO 19005)
  - JPEG2000 for color images. The advantage of this compression algorithm is that it supports both lossy and lossless data compression. Lossless is required for text-based content where no loss is acceptable. Lossy is acceptable for image content. (ISO 19005)
- Use lower case characters and avoid using special characters except hyphens and underscores in file names. Special characters to avoid include punctuation, spaces, or other non-alphanumeric symbols (e.g., \ / : \* ? < > | “ % # +) . (VBMS Constraint)

- Name the file in the following format: <Scanning\_Vendor\_Name><GUID>.pdf to prevent versioning collisions in VBMS. The file name MUST not exceed 255 characters. (VBMS Constraint)

To comply with the VBMS PDF specification, each PDF SHOULD:

- Use grayscale and color sparingly, as it significantly increases the file size. Grayscale and color should be used only when these features improve the reviewability of the material. (VBMS Constraint)

## 4.2 Metadata

The following table lists the metadata recommended for all PDFs.

Table 1: Recommended PDF Metadata		
Element	Value	Comments
Contributor (dc)	Persons, Organizations, or Services	The entity responsible for making contributions to the content of the resource
Creator (dc)	Person, Organization, or Service	The entity primarily responsible for making the content of the resource
Date (dc)	String	The date(s) associated with an event in the lifecycle of the resource. Typically, Date will be associated with the creation or availability of the resource.
Description (dc)	String	Textual description of the content of the resource
Format (dc)	Number of documents/pages within the file; Size of the file; Dimensions of the resource	Designates the physical or digital manifestation of the resource
Identifier (dc)	String or number conforming to a formal identification system	Unambiguous reference to the resource within a given context
Publisher (dc)	Person, Organization, or Service	An entity responsible for making the resource available
Subject (dc)	String	Describes the content of the resource
Title (dc)	String	A name given to the resource
Type (dc)	Text, StillImage, or Collection	To describe the file format, physical medium, or dimensions of the resource, use the Format element.