

# The future content imperative: The role of PDF in content lifecycles

## Table of contents

- 1 Executive summary
- 2 Meeting the format challenges of the modern content lifecycle
- 5 Choosing the right format for the right job
- 8 A format standard for the future
- 9 Conclusion

## Executive summary

Content demands of the twenty-first century require organizations and markets to rethink twentieth-century perceptions of content. E-paper, e-forms, regulatory compliance, collaboration, and new content media formats are all placing new demands on organizations and their content lifecycles.

Today, content lifecycles have grown more complex to include more review cycles across multiple organizations. Documents must be able to present diverse content to multiple audiences, each with different needs and goals; and they must be able to be preserved over a long period of time, as technology evolves. There is no single content file format that is ideal for every purpose or stage of the content lifecycle. The execution of content lifecycles requires different file formats, each optimized for certain lifecycle stages.

The PDF standard (ISO 32000) and the various PDF-based standards are file formats that can serve a variety of roles throughout the content lifecycle, especially during the review, publication, and archival stages. The use of PDF has become ubiquitous, with billions of files in existence. The standardization of PDF helps ensure that it will remain an important standard file format into the indefinite future. Its key strengths are high-fidelity content rendering, multiplatform systems support, robust forms technology, support for rich media and interactivity, strong security features, and digital signatures.

Extensible Markup Language (XML) is not a file format but rather a standard language for representing various styles of structured content. The main benefit is that there are common rules and tools for processing XML files, even if the schema structure is different. Searching can typically be done on XML files without understanding the details of the particular schema being used. PDF supports the use of XML for exchange of information among processes and information systems such as in forms automation. Open Document Format (ODF) and Office Open XML (OOXML) are content authoring formats which employ XML. PDF is complimentary to both ODF and OOXML.

Companies and governments are rejecting proprietary approaches to content technologies as they realize content tools change too often to ensure the longevity of associated proprietary vendor file formats. Electronic records in particular require some reasonable assurance of format longevity. PDF has already withstood the test of time. The first PDF files created 15 years ago are still viable electronic records today. Now, with PDF under the control of standards organizations, independent software vendors, private and public sector organizations can be reassured of their investments in PDF.

### **Meeting the format challenges of the modern content lifecycle**

Content demands of the twenty-first century require organizations and markets to rethink twentieth-century perceptions of content. Today, organizations must meet the following content requirements.

- Content often needs to be preserved at multiple stages within a content lifecycle, not just the final output.
- Paper documents are no longer the only final output.
- Document collaboration outside organizational boundaries is common and as a result content lifecycles frequently span multiple organizations.
- Content automation requires that formats must be process and machine friendly, not just people friendly.
- Everyday documents have evolved to include every conceivable content element, beyond just text, pictures, and artwork, such as multimedia and 3D.
- Industry and regulatory requirements for content lifecycles and documents have emerged which include detailed content format standards and archival requirements.
- The execution of content lifecycles requires different file formats, optimized for different lifecycle stages.

The PDF specification was first published in 1993 with the introduction of the Acrobat® applications of Adobe. and PDF quickly became a de facto document publishing and distribution standard in both public and private sectors. Its ability to exactly reproduce original document fidelity with fully searchable text enabled companies to quickly adopt PDF as their preferred electronic document format for customers, suppliers, partners, and employees. Governments across the globe standardized on PDF as the preferred format for citizen-facing documents and forms. As rich media, new collaboration paradigms, and the Internet drove new document metaphors, Adobe continued to innovate and update the PDF specification while preserving backward compatibility with previous PDF documents.

Adobe led the direction of PDF for 15 years. Now, PDF is an open standard (ISO 32000), helping to ensure it will continue to evolve to serve the global content revolution as a nonproprietary format.

### **Content lifecycle diversity breeds complexity**

Content lifecycles used to be discrete treatments of content within organizations with simple workflows and a single electronic content format such as Word, WordPerfect, Wang, and so on. Today, content lifecycles have grown in complexity to include more review cycles across multiple organizations, electronic workflow automation, different formats tailored to different lifecycle stages (for example, authoring versus electronic publishing), automation of some content creation, an explosion of industry and regulatory workflow and content structural requirements, and the integration of newer forms of content (for example, drawings, audio, video, form fields and data, and 3D objects). Documents must be able to present diverse content to multiple audiences, each with different needs and goals, and they must be able to be preserved over a long period of time, as technology evolves. PDF is an ideal document format versatile enough to satisfy this need.

While each industry can define a unique content lifecycle, a basic content lifecycle (as shown in figure 1) can be found within every type of organization consisting of these stages: creation, review, publication, utilization, and archival. Let's explore the challenges of each generic content lifecycle stage and see where PDF adds value throughout this lifecycle.

## Content lifecycle

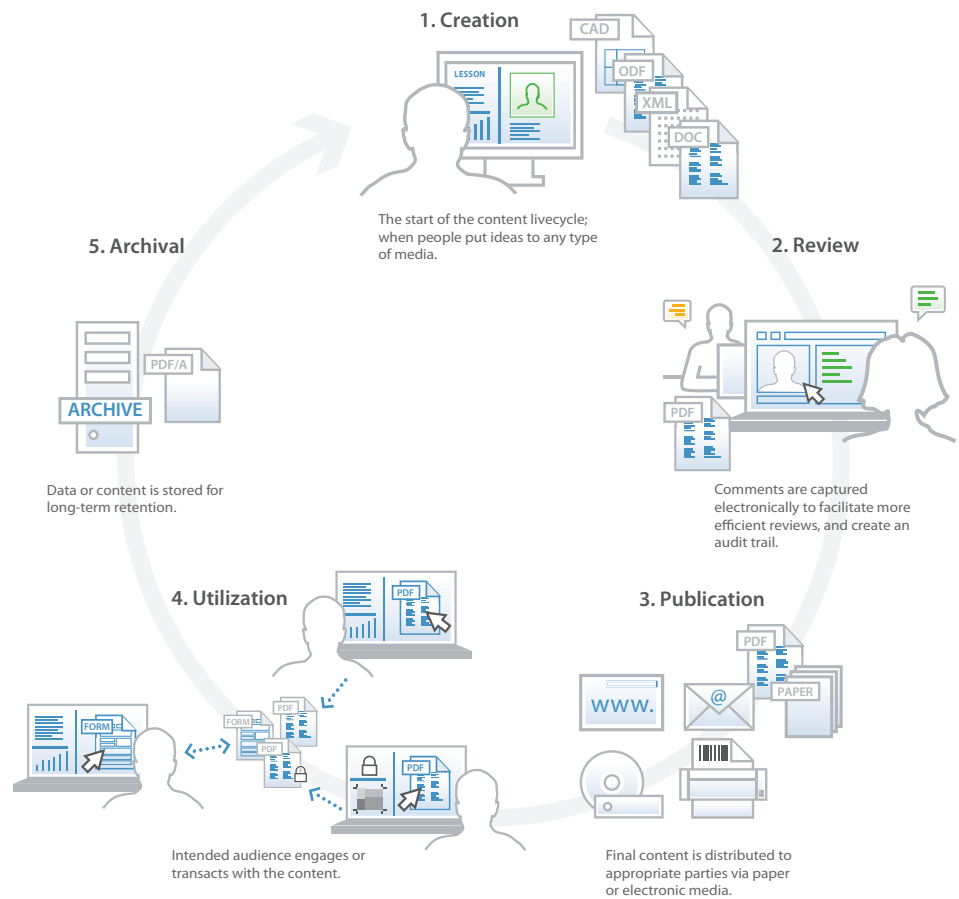


Figure 1. Content lifecycle, with typical document formats

### Creation

Every piece of content initially starts its life when people put ideas to media, when information flows from creative mind to computer screen. No matter what happens next, it has to start with someone creating content.

In the simplest example, a word processor is used as the only creation tool for a traditional business document. As business documents became more complex, new content objects were added such as images, drawings, tables, form fields, and charts. Word processor manufacturers tried to evolve their products by adding embedded editing tools for objects other than text. Today, the modern business document is created using numerous authoring tools, each specializing in specific content types. When electronic distribution of documents became popular, even more types of content objects were added to documents. The modern interactive electronic document includes: audio, video, animation, interactive forms and drawings, and direct integration with multiple business processes and data types.

Assembly of the rich diversity of modern content types into the single, secure and interactive container is required and PDF satisfies these requirements by design. Tools such as PDF Makers from Adobe enable creation of PDF pages directly within creation tools, such as Microsoft Office or Autodesk® AutoCAD®, leveraging the structure and attributes of each unique content type into the PDF pages and capturing more of the rich content available from the particular applications like document structure, animations and interactions.

For example, titles and section headings in a word processor application can be automatically referenced to interactive PDF bookmarks. Drawing layers can be automatically converted into layers in a PDF that can be made visible or invisible, just as in the CAD program itself. Three-dimensional objects can be integrated into a PDF so that manipulation of the object, perspective and explosion of object parts can be performed in the PDF. The versatility of the PDF standard supports this diverse array of content types and objects. As an ISO standard file format (not software application), the specifications needed by future generations to interpret and render PDF files will be available even after the tools that originally created the PDFs no longer exist or are not compatible with future computing platforms.

### **Review**

Most business documents undergo some level of review, by superiors, peers, partners, and even the public. Paper-based documents are cumbersome and time consuming, especially given the manual effort required to consolidate comments and suggestions. When electronic routing of documents became an essential part of review cycles, comments started being embedded inside the electronic file. This worked well as long as the reviewers had access to the same creation tools, (for example, employees of a company using the same word processor application and version).

Today's content review cycles typically reach beyond organizational boundaries. Sending proprietary file types to reviewers in different organizations provides no guarantee that they can even open them, let alone mark them up with comments and suggestions. Comments pertaining to content style and layout are impossible with these tools as well, since simple differences between computer systems such as installed fonts, printer drivers, color palettes, and computing platforms (for example, Windows® versus Mac OS) alter the look and feel of a document. These factors often lead to a disconnect in review cycles that requires physical printing of documents, thereby diminishing much of the value produced by electronic review cycles. Such disconnects can increase cycle times, errors, and the overall cost of review cycles.

The PDF standard is well suited as a universal review-cycle format. The exact look and feel of the document is preserved, independent of the individual user's computing environment. Commenting and markup is directly supported in the PDF standard so that PDF review tools made by different software vendors can markup the same PDF and all comments can be compiled into a single list. The resulting markups can all be displayed in the author's version of the PDF file. The PDF standard also supports the use of digital signatures, so that individual comments can be signed by the reviewer, to satisfy regulatory review requirements and establish a content review audit trail.

### **Publication**

While the use of paper by businesses is still required, today's business documents must be distributed to a wide variety of audiences in different organizations, with different needs and format preferences (for example, paper or electronic) and different distribution channels (for example, hard copy, e-mail, CD ROM/DVD, web or mobile device download). Environmental concerns are placing additional pressure on organizations to reduce unnecessary use of paper, further favoring the use of electronic document formats.

The PDF standard ensures that documents are presented as the creator intended, regardless of the recipient's computing environment, the channel used to distribute the PDF file, or the storage medium. The PDF standard also includes support for strong encryption features and methods to enable content management applications to control PDF files even after they have left an organization's sphere of control. These features enable access control for all recipients of the PDF file and even more granular security controls for individual users, user groups, and user process roles. Vendors of content security and digital rights management infrastructures can exploit these components of PDF to enable virtually complete control of a PDF file after it has been distributed, including revocation of any or all users' rights to open it.

### Utilization

A stage of the content lifecycle that is often omitted in discussions of content file formats and content tools is when a document actually gets used by someone. Published content is the manna that feeds most every major business or governmental process. Enabling users to use a document to execute a business process, for example a maintenance activity, is as important as the content lifecycle stages that created the document and the later stages that archive the document.

The PDF standard supports process execution primarily through two metaphors: 1) as a window through which users interface with the process, and 2) as a container for the original content and information collected as part of using the document to execute processes.

For example, interactive forms are a key component of the PDF standard. Form data can be saved with a PDF file for later recall and printing and the form data can be submitted by itself for processing in either real time or via e-mail.

A PDF file can also be used as a binder, more securely collecting information along the way as a business process is executed. This information can include form data, workflow data, unstructured content (for example, text and images), and it can even include audio and video information collected during process execution.

### Archival

The PDF standard is well suited as a document retention format since it is governed by an independent standards organization. The additional ability of PDF files to function as a binder of different content objects and discrete files is especially useful for retaining all materials from a single project. To satisfy regulatory and industry records retention and archival requirements a specialized subset of PDF has been created, PDF/A (archival). As a format standard, PDF/A is not controlled by a single corporation or government and is therefore an ideal format to stand the test of time required by records managers and archivists. Being a PDF subset, it can be read by any software that reads PDF files.

## Choosing the right format for the right job

The diversification of content lifecycles and the evolution of content tools to be more specialized have unfortunately produced confusion concerning the optimal roles for the different content formats available to organizations today. Microsoft Office "binaries," XML "formats," "Open Document Format" and "Office Open XML" are just a sampling of the content file formats confronting organizations today.

### The single file format myth

Organizations must understand that there is no single content file format that is ideal for every purpose or stage of the content lifecycle. The execution of content lifecycles requires different file formats, each optimized for certain lifecycle stages.

PDF files can be derived from any content authoring format used in content lifecycles, making it one of the most ubiquitous and flexible content formats available.

### The PDF family is diverse and growing

The ISO 32000 PDF standard has numerous PDF "relatives" designed for specific industries and functions: PDF/X (a family of graphics exchange formats), PDF/A (a family of formats optimized for archival and records management), PDF/E (a format optimized for architectural and engineering drawings), and additional standards proposals such as PDF/UA (a proposed standard for accessible PDF files for assistive technologies). There is even a PDF Healthcare best practices industry guide for using PDF to support healthcare records. Although these specifications are for specialized uses, all files are still PDF files and can be viewed with any standard PDF viewer.

### **Authoring formats**

Authoring formats, like Microsoft Office (for example, DOC, XLS) and AutoCAD (for example, DWG), were designed to support proprietary authoring tools created by software vendors. Their primary function is to serve the authoring application. They are less suitable for other lifecycle stages or facilitating the exchange of information during the execution of business processes.

As vendors of content authoring tools differentiated their products from competitors by adding more features, their file formats became more proprietary, complex, and cryptic. The file formats themselves were “the product” as much as the authoring tools themselves. Vendors leveraged these proprietary formats for two decades as exit barriers to prevent customers from switching to a competitor’s product. Few of these proprietary formats were published to enable third party developers to build tools to easily interoperate with the content locked in these formats.

During the creation stage of the content lifecycle, organizations typically assume that authors have access to the same tools and skills needed to create the specialized content produced by each authoring tool (for example, a document, a spreadsheet, vector graphics, or an engineering drawing). Authoring file formats are best suited to support specialized content creation, just as the specialized tools they are paired with.

### **Portable Document Format**

The PDF standard and its subsets are file formats that can serve a variety of roles throughout the content lifecycle, especially during the review, publication, and archival stages. Key strengths are high fidelity content rendering, multi-platform support, robust form support, rich media support, support for interactivity, strong security, digital signatures, and ability to function as a binder for the diverse information generated and collected during the execution of business processes.

The content within a PDF file can be processed by any application that understands how to read the format and it is not irrevocably tied with any particular authoring tool, including the Adobe Acrobat line of products. While PDF files and their content are dependent on content authoring tools as their progenitors, they are not authoring formats or tied to any vendor’s proprietary applications.

The use of PDF has become ubiquitous, with billions of files in existence. This is in large part because Adobe has provided the specification of the file format to the public from the start and recently the control of the PDF standard has been turned over to the International Standards Organization (ISO) to become ISO 32000. This adds additional confidence to developers and users of PDF that it will remain an important standard file format into the indefinite future.

### **XML**

Extensible Markup Language is not a file format at all. XML is a standard language for representing content, according to user-defined schemas. The main benefit is that information exchange can be facilitated through a common mark-up notation, provided one knows the schema used to represent the information.

It is ironic that for the past ten years XML has been portrayed by many as a technology that can be used to standardize content formats. In fact, XML literally facilitates infinite customization of content formats across any industry, through custom schemas. There is no such thing as a standard XML document. However, XML for specific purposes can produce content formats that can be used to easily exchange information between different organizations, processes, or information systems.

In recent years, a great amount of attention has been paid to replacing proprietary binary authoring formats with file formats based on XML. There are benefits and drawbacks to this approach. Unfortunately, a rational discussion of this issue has been largely suppressed by those advocating that a single XML-based document format should be adopted by all authoring tool vendors. This is the very antithesis of what XML was created for.

The primary benefit of employing XML in file formats is to allow enforcement of common tools and rules as custom markup languages are used for information exchange with other information systems and processes, including those outside of an organization's boundaries. Each authoring tool still comprises different features that need to be supported in unique XML-based file formats. As a result, it is arguable whether or not we need two similar office document authoring standards (that is, OOXML and ODF). However, regardless of how many competing office document authoring format standards exist, their key purpose is for authoring applications. PDF is also an office document format; however its primary purpose is for capturing finished documents for review, publication, and archiving.

The main drawback of "porting" every file format to an XML-based file format is that pure XML files can be extremely inefficient at storing information. XML files are long streams of sequential text that must be read from start to end. Older, proprietary authoring formats, such as DOC, support random access to information without having to process the entire file. The PDF standard is an object-oriented file that supports random access to its contents, yet it can be optimized for streaming over the web.

In fact, both ODF and OOXML files are not actually one single XML file. They both employ Zip archives to store collections of component files that contribute to the final compound document. Many of those components are XML files that are reduced in size considerably using Zip compression technology. Other files such as JPEG images can maintain their more suitable binary representations in the Zip archive with no need to make them XML.

There is also a significant lack of market support for these formats compared to PDF. Although it was only recently that PDF became an ISO standard, Adobe has made the specification freely available since PDF 1.0. This led to 15 years of third-party development around PDF before it became an open ISO standard, thus ensuring the format's longevity and support. PDF files created 15 years ago can still be successfully processed by a multitude of mainstream applications made by numerous vendors.

One of the most significant ways PDF and XML complement each other and interoperate is through form data (see figure 2). PDF form data can be saved as an XML data file, using a schema of the form designer's choice, for easy exchange among enterprise systems. PDF forms nicely separate the data that is to be exclusively read by humans from that data in an XML language that is tailored for easy processing by data-processing software. These capabilities are why PDF is preferred by so many organizations as a format for electronic business forms.

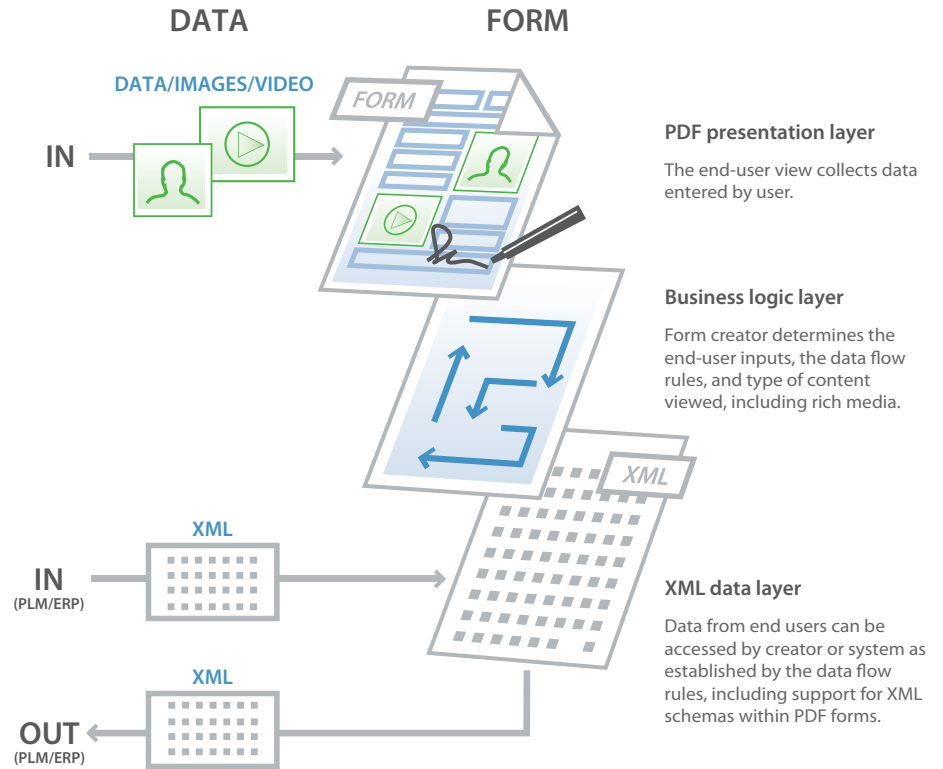


Figure 2. Example PDF and XML interoperability with a PDF form

## A format standard for the future

The ratification of PDF as an open standard has occurred at a tipping point for content lifecycles. Content management and the role of content and documents in the execution of business processes are changing rapidly.

### Content automation

Increasing global competition, rising compliance requirements, and a desire to improve productivity are pressuring organizations to automate more parts of the content lifecycle, especially publication and utilization stages. Organizations must be able to respond to opportunities faster, more securely closing the loop between publication and utilization activities, and ensure compliance with regulations, especially long-term retention requirements. Organizations will seek to remove manual processing from content lifecycles and replace them with automated assembly, distribution, and data processing (for example, forms) in response.

### Global information exchange

The exchange of content across organizational boundaries is a standard component of both private and public sector process execution. Technology barriers to information exchange such as proprietary formats; incompatible authoring tools, hardware platforms, and operating systems cannot keep pace with the expanding global information exchange. Content formats must be vendor neutral, platform independent, and support a maximum diversity of content types, regardless of where they fit in the content lifecycle spectrum.

## **Authentication and identity**

The demand for digital authentication of people who are performing activities throughout the content lifecycle is growing, especially due to regulations such as corporate governance and data privacy laws. Activities such as content review, form-fill, secure publishing, and tracking content utilization require authentication of an individual's identity. Market demand is also growing for the use of digital signatures in electronic documents and for the certification of electronic documents.

## **The content format standardization movement**

Companies and governments are rejecting proprietary approaches to content technologies as they realize content tools change too often to ensure the longevity of associated proprietary vendor file formats. They know adopting electronic records requires some reasonable assurance of longevity.

## **The Safety of the PDF standard**

The PDF standard is the most versatile content format available to address the most pressing challenges facing organizations today. PDF has already withstood the test of time. The first PDF files created 15 years ago are still viable electronic records today—an eternity in the information technology marketplace. This is despite the fact that none of the original content tools that created these PDF files and their contents, or the computing platforms they ran on, are in use today.

Now, with PDF under the control of independent standards organizations, independent software vendors, private and public sector organizations can be assured that their investments in PDF were the right choice.

## **Conclusion**

There is no single content file format that is ideal for every stage of the content lifecycle.

1. Content lifecycles may require different formats at different lifecycle stages. Authoring formats are essential for content creation, but are subservient to the vendor tools that create them. PDF is a content format standard that is particularly suited to enable content review, publication, and archival stages. It can also be used to package unstructured content, structured data, and media files at any content lifecycle stage.
2. ODF and OOXML are replacements for older authoring formats. They are not appropriate for use across the entire content lifecycle. ODF and OOXML are not substitutes for PDF, and PDF is not a substitute for ODF or OOXML. They are complimentary. The primary authoring tools that employ ODF and OOXML, OpenOffice and Microsoft Office, support creation of PDF files from these formats for content utilization at later stages of a content lifecycle.
3. XML is not a document format. XML is a standard way of representing information and allows for creation of custom schemas to fit any industry or process need. PDF can employ XML to facilitate the exchange of information between human readable formats and business processes.
4. The approval of the PDF standard as ISO 32000 helps ensure its longevity and protects it from exploitation by any single company or government. As long as vendors support the ISO standard, organizations can employ a variety of tools to build content architectures around PDF.



**Adobe**

**Adobe Systems Incorporated**  
345 Park Avenue  
San Jose, CA 95110-2704  
USA  
[www.adobe.com](http://www.adobe.com)

Adobe, the Adobe logo, and Acrobat are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Windows is either a registered trademark or trademark of Microsoft Corporation in the United States and/or other countries. Autodesk and AutoCAD are either registered trademarks or trademarks of Autodesk, Inc., in the USA and/or other countries. Mac OS is a trademark of Apple Inc., registered in the U.S. and other countries. All other trademarks are the property of their respective owners.

© 2008 Adobe Systems Incorporated. All rights reserved. Printed in the USA.  
95011455 8/08