Introduction to File Formats

A General Overview

February 2026

This document provides a general overview of common file formats used on the modern web, including text-based formats, image formats, and data interchange standards.

Table of Contents

1. Text-Based Formats

Text-based formats form the foundation of the web. HTML (HyperText Markup Language) defines the structure and content of web pages using a system of nested tags and attributes. Since its creation by Tim Berners-Lee in 1991, HTML has evolved through several major versions, with HTML5 being the current standard maintained by the WHATWG.

CSS (Cascading Style Sheets) controls the visual presentation of HTML documents. It separates content from design, allowing developers to define typography, colors, layouts, and responsive breakpoints independently of the document structure. Modern CSS supports custom properties, grid and flexbox layouts, animations, and media queries for responsive design.

JavaScript is the programming language of the web. Originally designed for simple client-side interactivity, it has grown into a full-featured language used for both frontend and backend development. JavaScript enables dynamic content updates, form validation, API communication, and complex application logic directly in the browser.

Plain text files (.txt) contain unformatted text without any markup or styling. Despite their simplicity, they remain widely used for documentation, configuration, logs, and data exchange where human readability and universal compatibility are priorities.

2. Image Formats

The web uses several image formats, each optimized for different use cases.

PNG (Portable Network Graphics) uses lossless compression, making it ideal for graphics, icons, and images with sharp edges or transparency. PNG supports alpha channels for variable transparency, 24-bit color, and interlacing for progressive loading.

JPEG (Joint Photographic Experts Group) uses lossy compression optimized for photographic images. It achieves much smaller file sizes than PNG for photos but introduces compression artifacts, especially at lower quality settings. JPEG does not support transparency.

GIF (Graphics Interchange Format) supports animation through multiple frames stored in a single file. Limited to a 256-color palette, GIF is best suited for simple animations, icons, and graphics with flat colors.

SVG (Scalable Vector Graphics) is an XML-based format that describes images using geometric shapes, paths, and text. Because SVG images are resolution-independent, they scale to any size without quality loss, making them ideal for logos, icons, and diagrams.

WebP is a modern format developed by Google that supports both lossy and lossless compression, as well as animation and transparency. WebP typically achieves 25-35% smaller file sizes than JPEG and PNG for equivalent quality, and has broad browser support as of 2024.

3. Data Interchange Formats

Data interchange formats allow structured information to be shared between systems in a standardized way.

JSON (JavaScript Object Notation) has become the dominant data interchange format on the web. Its lightweight syntax of key-value pairs, arrays, and nested objects is both human-readable and easy for machines to parse. JSON is the standard format for REST API responses, configuration files, and data storage in NoSQL databases.

XML (Extensible Markup Language) is a more verbose but highly flexible format that uses hierarchical tags to represent structured data. While less common than JSON for APIs, XML remains important for document formats (XHTML, SVG, RSS, EPUB), configuration files, and enterprise systems that require schema validation.

CSV (Comma-Separated Values) is the simplest tabular data format, storing rows of values separated by commas or other delimiters. CSV files are universally supported by spreadsheet applications, databases, and programming languages, making them a reliable choice for data exchange.

RSS and Atom are XML-based syndication formats that allow websites to publish frequently updated content (blog posts, news articles, podcast episodes) in a machine-readable feed that can be consumed by feed readers and aggregators.

4. Document Formats

Document formats are designed for presenting formatted text, images, and other content in a fixed layout.

PDF (Portable Document Format) was created by Adobe in 1993 to present documents consistently across platforms and devices. PDF files preserve fonts, images, layout, and formatting regardless of the software or hardware used to view them. PDFs can contain text, images, vector graphics, hyperlinks, form fields, digital signatures, and embedded multimedia.

The PDF specification became an open standard (ISO 32000) in 2008. Modern PDFs support accessibility features including tagged content, alternative text for images, and reading order metadata. PDF/A is an archival variant designed for long-term preservation of electronic documents.

Markdown is a lightweight markup language that uses plain-text formatting syntax to create structured documents. Originally created by John Gruber in 2004, Markdown has become the standard for README files, documentation, forum posts, and content management systems. Its simplicity makes it easy to write and read in source form while still supporting headings, lists, links, images, code blocks, and tables.

EPUB (Electronic Publication) is the standard format for e-books and digital publications. Built on web technologies (HTML, CSS, SVG), EPUB files are essentially packaged websites with metadata, supporting reflowable text that adapts to different screen sizes and reader preferences.

5. Compression and Archives

Compression formats reduce file sizes for efficient storage and transfer.

ZIP is the most widely used archive format, combining multiple files and directories into a single compressed file. ZIP supports several compression algorithms (Deflate being the most common) and is natively supported by all major operating systems.

GZIP (GNU Zip) is a single-file compression format widely used for web content delivery. HTTP servers commonly use gzip encoding to compress responses, reducing bandwidth by 60-80% for text-based content.

TAR (Tape Archive) bundles multiple files into a single archive without compression. It is typically combined with gzip (.tar.gz or .tgz) or other compressors for both archival and compression.

6. Configuration Files

Configuration files control application behavior without modifying code.

Common formats include INI files (simple key-value pairs in sections), YAML (human-friendly data serialization using indentation), TOML (a minimal configuration language designed to be unambiguous), and JSON (increasingly used for tool and project configuration).

Environment files (.env) store sensitive configuration like API keys and database credentials separately from source code. Web-specific configuration files include robots.txt (crawler directives), manifest.json (progressive web app metadata), and humans.txt (team credits following the humanstxt.org convention).

7. Media Formats

Audio and video formats enable rich media experiences on the web.

MP3 remains the most widely supported audio format, using lossy compression to achieve small file sizes suitable for streaming and downloads. AAC (Advanced Audio Coding) offers better quality at similar bitrates and is the default format for Apple devices and YouTube.

OGG Vorbis is an open-source alternative to MP3 with generally better quality at equivalent bitrates. Opus, another open format, has become the preferred codec for real-time communication (WebRTC) due to its low latency and adaptability.

For video, MP4 (H.264/AVC) is the dominant format for web delivery, balancing quality and file size with near-universal browser support. WebM (VP8/VP9) is Google's open format alternative, while AV1 is a newer royalty-free codec offering significantly better compression at the cost of slower encoding.

The HTML5 video and audio elements allow native media playback without plugins. Media Source Extensions (MSE) enable adaptive bitrate streaming, allowing players to switch quality levels based on network conditions.

8. Summary

The modern web relies on a diverse ecosystem of file formats, each designed to serve specific needs:

Text-based formats (HTML, CSS, JavaScript, plain text) form the structural and interactive foundation of web content.

Image formats (PNG, JPEG, GIF, SVG, WebP) balance quality, file size, transparency support, and animation capabilities.

Data interchange formats (JSON, XML, CSV, RSS) enable structured communication between systems and applications.

Document formats (PDF, Markdown, EPUB) preserve formatted content for consistent presentation and long-term archival.

Compression formats (ZIP, GZIP, TAR) reduce storage and bandwidth requirements for efficient delivery.

Understanding the characteristics, strengths, and limitations of each format helps developers and content creators choose the right format for each use case, balancing compatibility, performance, quality, and accessibility.

References

[1] MDN Web Docs - Web technology for developers
[2] W3C - World Wide Web Consortium Standards
[3] IETF RFC 2616 - Hypertext Transfer Protocol
[4] ISO 32000-2:2020 - Document management (PDF 2.0)
[5] WHATWG HTML Living Standard
[6] Ecma International - ECMAScript Specification