GZIP

>> Back

Table of Contents

Format Description Properties

- ID: fdd000599
- Short name: gzip
- Content categories: aggregate
- Format Category: file-format
- Other facets: container-bundle, binary, structured, symbolic, compression
- Last significant FDD update: 2024-01-30
- Draft status: Preliminary

Identification and description

Full name	GZIP
Description	The GZIP file format is a single-file, single-stream, lossless data compression file format which generally has the suffix .gz. GZIP may also refer to the utility which produces the file format. The GZIP format is widely adopted. It provides a container that stores the original file name, timestamp, basic cyclic redundancy checks (CRC), other optional data, and a chosen file compressed by a chosen compression algorithm. **Format vs Tool** GZIP serves a dual role as both a file format and a software application dedicated to file compression and decompression. As a file format, it encapsulates data for efficient compression. As a software application, it provides tools for compressing and decompressing files. The GZIP tool, commonly associated with Unix systems, has an associated manual. Various implementations of the GZIP tool exist, utilizing different compression algorithms. A comprehensive summary of these implementations can be found in the "Implementations" section on the Wikipedia page about GZIP. **Structure**

According to the GZIP File Format Specification, Version 4.3, RFC 1952, a GZIP file is structured as:

- Header (10 Bytes): Contains a magic number, a version number, and a timestamp.
- Optional Extra Headers: May include additional information such as the original file name.
- Body: Contains a DEFLATE-compressed payload, representing the actual compressed data.
- Footer (8 Bytes): Comprises a CRC-32 checksum and the length of the original uncompressed data.

The GZIP file itself consists of a sequence of "members" or compressed data sets, with each member's format specified in the subsequent section. Members are concatenated one after another in the file, without any extra information before, between, or after them. This format is defined in detail by RFC 1952.

GZIP is frequently used in conjunction with the Tape Archive (TAR) file format to create a compressed archive format with the extension .tar.gz. This combination allows for the compression of multiple files and directories into a single compressed archive. See also Tape Archive (tar) File Format Family.

Relationship with DEFLATE

GZIP employs the DEFLATE algorithm, a combination of LZ77 and Huffman coding lossless compression algorithms. DEFLATE was developed as a replacement for Lempel-Ziv-Welch (LZW) lossless compression algorithm and other patented compression algorithms that limited the usability of other popular archivers. This choice aimed to overcome patent restrictions and enhance compression efficiency.

The DEFLATE compressed data format used by GZIP offers improved compression compared to Unix compress, along with fast decompression. Additionally, it includes a CRC-32 as an integrity check for data. The header format in GZIP allows for the storage of more information than the compress format, such as the original file name and file modification time.

The zlib library supports DEFLATE compression and decompression, offering three types of wrapping around deflate streams: raw deflate, zlib wrapping (used in portable network graphic (PNG) format data blocks), and GZIP wrapping. Zlib wrapping is more compact (six bytes) compared to GZIP (a minimum of 18 bytes), and its integrity check (Adler-32) is faster than the CRC-32 used by GZIP. Raw deflate is utilized by programs handling the .zip format, another format that wraps around deflate compressed data.

Relationship with compress and TAR

According to a Stack Overflow answer by the GZIP format creator, the Unix tool `compress` traditionally compressed individual files. The introduction of TAR allowed users to create a comprehensive archive encapsulating files, attributes, and directory structure. It became common to use the `compress` tool on the TAR file, resulting in a .tar.Z file. This dual approach transitioned into the GZIP format.

TAR evolved to support compression simultaneously to GZIP, eliminating the need to use the `compress` tool. With GZIP, the integration of TAR results in a tar.gz file. Tar.gz is widely adopted on Unix systems for its portability.

See Notes for information on the capitalization of GZIP.

Production phase	May be used at any lifecycle phase for bundling/packaging files together for exchange, storage, or distribution.
Relationship to other formats	
May contain	tar, Tape Archive (tar) File Format Family. One common workflow is to use the tar utility to create an archive of files, their attributes, and their directory structure into a single .tar file, and then compress it with compress to make a .tar.Z file.
Contains	DEFLATE. DEFLATE Compressed Data Format. A GZIP file is deflate compression with a checksum and header metadata included. But DEFLATE could be used on its own, outside of GZIP. DEFLATE is also used by ZIP File Format (PKWARE).
Modification of	VRML, Virtual Reality Modeling Language Family

Local use

LC experience or existing holdings	The Library of Congress has many GZIP files, especially in web archiving collections.
LC preference	The Library of Congress Recommended Formats Statement (RFS) lists GZIP as a preferred format for web archives.

Sustainability factors

Disclosure	Fully disclosed.
Documentation	GZIP file format specification, version 4.3 (RFC 1952). The DEFLATE Compressed Data Format Specification, Version 1.3 (RFC 1951), describes the DEFLATE format contained within the GZIP wrapper. Both RFC 1952 and RFC 1951 state that they are informational: “this memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited." Documentation for the GZIP data compression program is available here on the GNU’s Not Unix (GNU) website.
Adoption	The official website for the GZIP program states that GZIP is in “wide use”.
Licensing and patents	The GNU Operating System website states that, “GZIP is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 3 of the License, or (at your option) any later version.” The GZIP specification was copyrighted by L. Peter Deutsch in 1996.The Acknowledgements section of the specification notes that Jean-Loup Gailly designed the GZIP format and, along with Mark Adler, developed the related software. Glenn Randers-Pehrson converted the document to RFC and HTML format.
Transparency	Depends upon algorithms and tools to read. Would require sophistication to build tools from scratch.
Self-documentation	A GZIP file's header contains its magic number, a version number and a time stamp. The optional header may include data such as the original filename. **Accessibility Features**

	No specific features in the file format. Features to support accessibility would be found in the bundled and compressed files (such as embedded captions and subtitles in audiovisual content, tagged and structured text in textual documents, and alt text for images). Aggregate files can also contain separate files for transcripts, timed text or captions as part of the bundled package. See Relationships to other formats for details.
External dependencies	None, beyond the availability of software to extract and decompress the files contained in a GZIP file.
Technical protection considerations	This format does compression, it does not do any encryption. Separate tools would have to be used on the source material to obtain those kinds of additional security measures.

Quality and functionality factors

Aggregate	
Compression	The GZIP format typically uses DEFLATE for compression.
Support for Error Dectection	RFC 1952 states that "If FHCRC is set, a CRC16 for the gzip header is present, immediately before the compressed data. The CRC16 consists of the two least significant bytes of the CRC32 for all bytes of the gzip header up to and not including the CRC16."

File type signifiers and format identifiers

Tag	Value	Note
Filename extension	gz z	See PRONOM: https://www.nationalarchives.gov.uk/PRONOM/x-fmt/266. RFC 6713 includes "gz" but not "z"/
Internet Media Type	application/gzip	See RFC 6713: The 'application/zlib' and 'application/gzip' Media Types
Magic numbers	1F8B08	See: PRONOM: https://www.nationalarchives.gov.uk/PRONOM/x-fmt/266 Wikidata: https://www.wikidata.org/wiki/Q10287816 The first two bytes are always going to be 1F and 8B. The third byte is for compression, which is usually 08 for "DEFLATE". See: RFC 1951. "These have the fixed values ID1 = 31 (0x1f, \037), ID2 = 139, (0x8b, \213), to identify the file as being in GZIP format." See: RFC 1952 section 2.3.1. "This identifies the compression method used in the file. CM = 0-7 are reserved. CM = 8 denotes the "deflate" compression method, which is the one customarily used by GZIP and which is documented elsewhere." See: RFC 1952 section 2.3.1. See Ange Albertini's GNU GZip poster for more details.
Uniform Type Identifier (Mac OS)	org.gnu.gnu-zip-archive	See PRONOM: https://www.nationalarchives.gov.uk/PRONOM/x-fmt/266
Pronom PUID	x-fmt/266	See https://www.nationalarchives.gov.uk/PRONOM/x-fmt/266
Wikidata Title ID	Q10287816	See https://www.wikidata.org/wiki/Q10287816
Other	NF00204	NARA File Format Preservation Plan ID. See https://www.archives.gov/files/lod/dpframework/id/NF00204.ttl

Notes

General	The capitalization of GZIP is inconsistent across sources, appearing as GZIP, GZip, gzip, and Gzip. The official specification capitalizes the format name as GZIP.
History	The GZIP format and software were created by Jean-loup Gailly and Mark Adler to replace the Unix `compress` utility. This decision stemmed from concerns over patent issues related to the LZW algorithm used by `compress`. The GZIP utility not only avoided patent infringement but also delivered superior compression. Despite its inception back in the early '90s, GZIP remains widely utilized today. The initial public release of GZIP, version 0.1, occurred on October 31, 1992, followed by version 1.0 in February 1993. There have been over 600 releases since 1993 which are indexed in the GZIP git repository. GZIP is now maintained by Jim Meyering and Paul Eggert.

Format specifications

- RFC 1952, "GZIP file format specification version 4.3". (https://tools.ietf.org/html/rfc1952).
- RFC 1951, "DEFLATE Compressed Data Format Specification version 1.3". (https://tools.ietf.org/html/rfc1951). This describes the DEFLATE format contained within the GZIP wrapper.

Useful references

URLs

- GZIP Home Page. GZIP. (http://www.gzip.org/).
- DEFLATE. Wikipedia. (https://en.wikipedia.org/wiki/Deflate).
- GZIP. Wikipedia. (https://en.wikipedia.org/wiki/Gzip).
- GZIP - Implementations. Wikipedia. (https://en.wikipedia.org/wiki/Gzip#Implementations).
- Huffman coding. Wikipedia. (https://en.wikipedia.org/wiki/Huffman_coding).
- Lempel–Ziv–Welch. Wikipedia. (https://en.wikipedia.org/wiki/Lempel%E2%80%93Ziv%E2%80%93Welch).
- LZ77 and LZ78. Wikipedia. (https://en.wikipedia.org/wiki/LZ77_and_LZ78).
- GZIP Format. Kaitai. (https://formats.kaitai.io/gzip/).
- GZIP Git Repository - Revision Log. GNU GZIP. (https://git.savannah.gnu.org/cgit/gzip.git/log/?ofs=650).
- GZip Image. GitHub - corkami/pics. (https://github.com/corkami/pics/blob/master/binary/GZip.png).
- How are zlib, gzip and zip related? What do they have in common and how are they different? Stack Overflow. (https://stackoverflow.com/questions/20762094/how-are-zlib-gzip-and-zip-related-what-do-they-have-in-common-and-how-are-they/20765054#20765054).
- GNU General Public License. GNU. (https://www.gnu.org/licenses/gpl.html).
- GNU GZIP. GNU. (https://www.gnu.org/software/gzip/).
- GNU GZIP Manual. GNU. (https://www.gnu.org/software/gzip/manual/).
- Informational vs Experimental Documents. IETF. (https://www.ietf.org/standards/process/informational-vs-experimental/).
- ZIP File Format (PKWARE). The Library of Congress - Format Description Properties. (https://www.loc.gov/preservation/digital/formats/fdd/fdd000354.shtml).
- Tape Archive (tar) File Format Family. The Library of Congress - Format Description Properties. (https://www.loc.gov/preservation/digital/formats/fdd/fdd000531.shtml).
- mailto:bug-gzip@gnu.org (mailto:bug-gzip@gnu.org). GZIP has a mailing list used to discuss all aspects of GZIP including development, enhancement requests, and bug reports.
- zlib (https://zlib.net/). zlib is a free software library written in C which you can use to read and write GZIP files and memory streams.
- Tiny.gzip sample file. Ange Albertini (https://github.com/corkami/pocs/blob/master/mini/gzip.gz).

- Poster diagram of gzip. Ange Albertini (https://github.com/corkami/pics/blob/master/binary/GZip.png).
- PRONOM entry for fmt/266 (https://www.nationalarchives.gov.uk/PRONOM/x-fmt/266). Information in PRONOM from UK National Archives about GZIP file. PUID: fmt/266
- Wikidata entry for Q10287816 (https://www.wikidata.org/wiki/Q10287816). Information in Wikidata about GZIP. Wikidata Title ID: Q10287816
- NARA File Format Preservation Plan ID for NF00204. (https://www.archives.gov/files/lod/dpframework/id/NF00204.ttl).

Last Updated: 05/14/2024