Sustainability of Digital Formats: Planning for Library of Congress Collections

bzip2

>> Back

Table of Contents

Format Description Properties

- ID: fdd000600
- Short name: bzip2
- Content categories: aggregate
- Format Category: file-format
- Other facets: container-bundle, binary, structured, symbolic, compression
- Last significant FDD update: 2023-01-26
- Draft status: Preliminary

Identification and description

Full name	bzip2
Description	The bzip2 file format is a freely available, patent-free data compression program created by Julian Seward. It is both the name of the format and the program used to create it. The program is designed for compressing single files only. It was created as a successor to its predecessor, bzip, to avoid potential patent issues. See: History for more information. Different versions of bzip2 maintain file format compatibility. Newer versions can work with files created by older versions, ensuring a level of stability. However, the format creator acknowledges limitations in the compressed file format and provides source code for decompressing older files created by bzip-0.21. A bz2 stream consists of a 4-byte header, followed by zero or more compressed blocks. An end-of-stream marker contains a 32-bit CRC for the plaintext whole stream processed. The compressed blocks are bit-aligned, and no padding occurs.
Production phase	May be used at any lifecycle phase for bundling/packaging files together for exchange, storage, or distribution.

Local use

LC experience or existing holdings	The Library of Congress has a small number of bzip files across its varied collections.
LC preference	Bzip is not includes in the Library of Congress Recommended Formats Statement.

Sustainability factors

Disclosure	No formal specification for the bzip2 file format exists. Comments welcome.
Documentation	Two unofficial documentation resources are commonly cited. • " Not an Official Specification" by Joe Tsai. • Repository for the bzip2 program on GitLab maintained by Micah Snyder.
Adoption	Widely adopted. The bzip2 file format "ships standard on many Unix/Linux systems." Often compared to gzip and ZIP File Format (PKWARE).
Licensing and patents	The bzip2 homepage states that the license is a GNU's Not Unix (GNU) General Public License (GPL). It is unclear which version of GNU GPL would apply. Other sources state conflicting information about bzip2's license, stating it is a Berkeley Software Distribution (BSD) style license. Comments welcome.
Transparency	Depends upon algorithms and tools to read. Would require sophistication to build tools from scratch.
Self-documentation	Identifies self as a bzip2-compressed file with magic numbers (see magic numbers section). There is no specific language for the inclusion of other metadata. However, documentation is sparse. Comments welcome. **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 bzip2 file.
Technical protection considerations	Does not support encryption.

Quality and functionality factors

Aggregate	
Compression	According to the bzip2 software official manual, bzip2 files are compressed using the Burrows-Wheeler block-sorting text compression algorithm, and Huffman coding.
Support for Error Dectection	Unknown. Comments welcome.

File type signifiers and format identifiers

Tag	Value	Note
Filename extension	bz2	Used for bzip2. See Wikidata: https://www.wikidata.org/wiki/Q27866052
Internet Media Type	application/x-bzip2	See the Mozilla list of common MIME types. Not listed in IANA.
Magic numbers	Hex: 42 5a 68 ASCII: BZh	For more details see: • " Not an Official Specification," by Joe Tsai. • “ BZip 2,” a poster by Ange Albertini. • Gary Kessler's File Signature Table. Note this header, when converted from Hexadecimal to ASCII, is "BZh". “BZ” stands for “bzip”, and the "h" is for "Huffman coding," the compression algorithm used with bzip2. Some sources, such as Wikipedia, will cite the magic numbers as “BZh” instead of the hexadecimal.
Pronom PUID	x-fmt/268	See https://www.nationalarchives.gov.uk/PRONOM/x-fmt/268
Wikidata Title ID	Q27866052	See https://www.wikidata.org/wiki/Q27866052

Notes

General	The bzip2 program, and by extension the bzip2 file format, is based on its predecessor bzip. Despite similarities in appearance and name to bzip, bzip2 is rewritten and re-engineered. It was developed to address potential patent issues with bzip. The format created by bzip2 is not compatible with bzip, and efforts to make them compatible were avoided to maintain the purpose of patent avoidance. Seward expressed commitment to backwards compatibility for future changes. The predecessor program bzip is no longer available.
History	Julian Seward released bzip2, version 0.15, in July 1996. The compressor’s popularity grew over the next several years due to its stability. Julian Seward released version 1.0 in late 2000. In June 2019 Federico Mena became the new maintainer of bzip2. In 2019, Mark Wielaard began maintaining a bzip2 stable repository at Sourceware. In June 2021 Micah Snyder became the new maintainer of the Sourceware repository.

Format specifications

- "Not an Official Specification" by Joe Tsai. (https://github.com/dsnet/compress/blob/master/doc/bzip2-format.pdf).
- Repository for the bzip2 program on GitLab maintained by Micah Snyder. (https://gitlab.com/bzip2/bzip2/).

Useful references

URLs

- “Bzip2 Downloads.” Sourceware. (http://sourceware.org/bzip2/downloads.html).
- ETHW – “History of Lossless Data Compression Algorithms”. Engineering and Technology History Wiki (ETHW). (https://ethw.org/History_of_Lossless_Data_Compression_Algorithms#Legal_Issues).

- "BZip2". Poster by Ange Albertini. (https://github.com/corkami/pics/blob/master/binary/BZ2.png).
- bzip2 Sample File. Ange Albertini. (https://github.com/corkami/pocs/blob/master/mini/bzip2.bz2).
- bzip2 Manual. Sourceware. (https://sourceware.org/bzip2/manual/manual.html).
- bzip2 Git Summary. Sourceware. (https://sourceware.org/git/?p=bzip2.git;a=summary).
- "bzip2 Changing Maintainership". Viruta. (https://viruta.org/bzip2-changing-maintainership.html).
- "Maintaining bzip2". Viruta. (https://viruta.org/maintaining-bzip2.html).
- "The bzip2 home page". Muraroa. (https://web.archive.org/web/19980704181204/http:/www.muraroa.demon.co.uk/). Accessed via the Internet Archive.
- "A Block-Sorting, Lossless Compressor". Muraroa. (https://web.archive.org/web/19980704181204/http:/www.muraroa.demon.co.uk/bunzip021.c). Accessed via the Internet Archive.
- "What is bzip2?" Bzip.org. (https://web.archive.org/web/20050207055319/http:/www.bzip.org/). Accessed via the Internet Archive.
- File Signatures. Gary Kessler. (https://www.garykessler.net/library/file_sigs.html).
- "GPL 1.0 License". GNU. (https://www.gnu.org/licenses/old-licenses/gpl-1.0.html).
- "A Block-Sorting Lossless Data Compression Algorithm". Michael Burrows and D. J. Wheeler. May 10, 1994. Digital SRC Research Report 124. (https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.pdf).
- "Efficient Decoding of Prefixe Codes". Daniel S. Hirschberg and Debra A. LeLewer. Communications of the ACM, Vol 33, Number 4. April 1990. (https://dl.acm.org/doi/10.1145/77556.77566).
- "Fast Algorithms for Sorting and Searching Strings". Jon L. Bentley and Robert Sedgewick. (https://sedgewick.io/wp-content/themes/sedgewick/papers/1997StringsSODA.pdf).
- PRONOM entry for fmt/268 (https://www.nationalarchives.gov.uk/PRONOM/x-fmt/268). Information in PRONOM from UK National Archives about bzip2. PUID: fmt/268
- Wikidata entry for Q27866052 (https://www.wikidata.org/wiki/Q27866052). Information in Wikidata about bzip2. WIkidata Title ID: Q27866052

Last Updated: 04/30/2024

Digital Preservation Home | Digital Formats Home