

Star: A Stupid Archive Format

Carter Brainerd
0xCB@protonmail.com

04/05/19

1 Abstract

Archiving is not something most people think about on a regular basis. The `tar` program has been the de facto standard for archiving files for over 30 years. `Tar` assumes that it is always working in a strictly Unix environment, which is why we do not see `tar` being regularly used on Windows. In short, `tar` is not a cross-platform archiving solution. We introduce a simpler, cross-platform archiving solution called `star`.

2 Introduction

While the idea of `tar` started out as a great idea, it's evolved to include multiple different formats, flags, options, and implementations. It has gotten too complicated. The `star` program gets rid of the extraneous flags and options. It is simple, clean, secure, and reliable. It is coded in the Crystal language, so the program itself is easy to read and understand. Everything about `star` is simple: the program, the file format, and its processes.

3 File Specification

The `.star` file format is a binary file format (called a ‘starfile’). All hexadecimal strings shown in the table are meant to be on one line. They are separated by newlines in the table for spacing purposes.

Total bytes	Byte Value	Meaning
16	\x73\x20\x74\x20\x61\x20 \x72\x20\x31\x00\x00...	s t a r 1\x00*16 (file magic)
8	\xCA\xFE\xCA\xFE \xBA\xBE\xBA\xBE	Beginning of file list
n	filename{*&*}filehash{:\x00:} filename{*&*}filehash{:\x00:}...	File list separated by {:\x00:}
8	\xBA\xBE\xBA\xBE \xCA\xFE\xCA\xFE	End of the file list
24	\x53\x20\x54\x20\x41\x20 \x52\x20\x42\x20\x45\x20 \x47\x20\x49\x20\x4e\x20 \x5c\x20\x5c\x20\x26\x24	Beginning of file (S T A R B E G I N \\\&\$)
n	file contents	The contents of the file
24	\x53\x20\x54\x20\x41\x20 \x52\x20\x45\x20\x4e\x20 \x44\x20\x4f\x20\x46\x20 \x5c\x20\x5c\x20\x26\x24	End of file (S T A R E N D O F \\\&\$)
n	{begin file hex}{file contents} {end of file hex}	Basic layout for file contents. Repeats for every file in the archive.
16	\x65\x20\x73\x20\x74\x20 \x61\x20\x72\x00\x00...	End of star file

4 Simple

For `star`, simplicity is the name of the game. The command-line program uses a task-style syntax (eg. `star extract --verbose myfile.star`) instead of `tar`'s many flags and flag combinations. `star` sticks to the idea of "if there is nothing to say, don't say it." There is never any nonsense clogging up your terminal window (unless you turn on verbose mode).

Besides the command-line program, the file format is dead simple too. There are no headers, file metadata, permissions, or OS-level information stored in the archive. This simplicity is what allows `star` to be a truly cross-platform archiving solution.

5 Compatible

The essence of `star` is cross-platform compatibility. The file format intentionally leaves out OS specific file information to preserve cross-compatibility. The original program is written in the Crystal programming language, so it can be compiled for nearly every operating system.

This file format simplicity is especially important for embedded devices that don't have file systems in place. Since the starfile only contains the file contents, filesystem-less systems are able to easily handle the files. Not only are starfiles easy to handle, they also do not necessarily need to be extracted to read the combined files.

6 Security

Every starfile has basic tamper and data-corruption security built in. Each entry in the file list contains a SHA-256 hash of that file's contents. `Star` goes through every file and checks the hash in the file list with the newly calculated hash of the contents of the file. If the program encounters a mismatch, it stops the extraction and warns the user.

6.1 Security oversights

While the program does check for hash mismatches, an attacker can change the file contents as well as the hash in the file list. This can be avoided by including a `sha256sum` of the starfile and comparing the hash of the intended file and the hash of the downloaded data.

7 Conclusion

We have introduced a new and modern archiving file format. The program does not store any OS-specific metadata or permissions, so they can be transferred across operating systems without any trouble. This format also can easily be used by machines without a dedicated file system because the full contents of each file are included in the archive, so they can be read and parsed without writing anything to the disk.