anima Working Group
Internet-Draft
Intended status: Standards Track
Expires: 25 July 2021

M. Richardson
Sandelman Software Works

21 January 2021

On storing CBOR encoded items on stable storage
draft-richardson-cbor-file-magic-01

Abstract

This document proposes an on-disk format for CBOR objects that is friendly to common on-disk recognition systems like the Unix file(1) command.

This document is being discussed at: https://github.com/mcr/cbor-magic-number

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 25 July 2021.

Copyright Notice

Table of Contents

1. Introduction

Since very early in computing, operating systems have sought ways to mark which files could be processed by which programs.

For instance, the Unix file(1) command, which has existed since 1973 ([file]), has been able to identify many file formats for decades. Many systems (Linux, MacOS, Windows) will select the correct application based upon the file contents, if the system can not determine it by other means: for instsance, MacOS maintains a resource fork that includes MIME information and therefore ideally never needs to know what anything about the file. Other systems do this by file extensions.

While having a MIME type associated with the file is a better solution in general, when files become disconnected from their type information, such as when attempting to do forensics on a damaged system, then being able to identify a file type can become very important.

It is noted that in the MIME type registration, that a magic number is asked for, if available, as is a file extension.

A challenge for the file(1) program is often that it can be confused by the encoding vs the content. For instance, an Android "apk" used to transfer and store an application may be identified as a ZIP file. Both OpenOffice or MSOffice files are XML files, but appear as ZIP, unless they are flat files, in which case they appear to be generic

XML files.

As CBOR becomes a more and more common encoding for a wide variety of artifacts, identifying them as CBOR is probably not useful. This document provides a way to encode a magic number into the beginning of a CBOR format file. Two options are presented, with the intention of standardizing only one.

These proposals are invasive to how CBOR protocols are written to disk, but in both cases, the proposed envelope does not require that the tag be transfered on the wire.

In addition to the on-disk identification aspects, there are some protocols which may benefit from having such a magic on the wire if they presently using a different (legacy) encoding scheme. The presence of the identifiable magic sequence signals that CBOR is being used or a legacy scheme.

2. Requirements for a Magic Number

A magic number is ideally a unique fingerprint, present in the first 4 or 8 bytes of the file, which does not change when the content change, and does not depend upon the length of the file.

Less ideal solutions have a pattern that needs to be matched, but in which some bytes need to be ignored. While the Unix file(1) command can be told to ignore bytes, this can lead to ambiguities.

3. Protocol Proposal

This proposal makes use of CBOR Sequences as described in [RFC8742].

This proposal consists of two tags and a constant string for a total of 12 bytes.

1. The file shall start with the Self-described CBOR tag, 55799, as described in [RFC8949] section 3.4.6.

2. The file shall continue with a CBOR tag, from the First Come First Served space, which uniquely identifies the CBOR Protocol. The use of a four-byte tag is encouraged.

3. The three byte CBOR array containing 0x42_4F_52. When encoded it shows up as "CBOR"

The first part identifies the file as being CBOR, and does so with all the desirable properties explained in Specifically, it does not seem to conflict with any known file types, and it is not valid Unicode.[RFC8949] section 3.4.6.

The second part identifies which CBOR Protocol is used. CBOR Protocol designers should obtain a tag for each major object that they might store on disk. As there are more than 4 million available 4-byte tags, there should be issue in allocating a few to all available CBOR Protocols. The policy is First Come First Served, so all that is required is an email to IANA, having filled in the small template provided in section 9.2 of [RFC8949].

The third part is a constant value 0x43_42_4f_52, "CBOR". This means that should a file be reviewed by a human (directly in an editor, or in a hexdump display), it will include the string "CBOR" prominently. The value is also included because the two tags need to tag something.

4. Security Considerations

This document provides a way to identify CBOR Protocol objects. Clearly identifying CBOR contents on disk may have a variety of impacts.

The most obvious is that it may allow malware to identify interesting objects on disk, and then corrupt them.

5. IANA Considerations

This document makes no new requests to IANA.

6. Acknowledgements

The CBOR WG brainstormed this protocol on January 20, 2021.

7. Changelog

8. References

8.1. Normative References

[BCP14] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8742] Bormann, C., "Concise Binary Object Representation (CBOR) Sequences", RFC 8742, DOI 10.17487/RFC8742, February 2020, <https://www.rfc-editor.org/info/rfc8742>.

[RFC8949] Bormann, C. and P. Hoffman, "Concise Binary Object Representation (CBOR)", STD 94, RFC 8949, DOI 10.17487/RFC8949, December 2020, <https://www.rfc-editor.org/info/rfc8949>.

8.2. Informative References

[file] Wikipedia, "file (command)", 20 January 2021, <https://en.wikipedia.org/wiki/File_%28command%29>.

[ilbm] Wikipedia, "Interleaved BitMap", 20 January 2021, <https://en.wikipedia.org/wiki/ILBM>.

Contributors

Author's Address

Michael Richardson
Sandelman Software Works

Email: mcr+ietf@sandelman.ca