HTTPbis Working Group
Internet-Draft
Obsoletes: 2616 (if approved)
Intended status: Standards Track
Expires: January 16, 2014

R. Fielding, Ed.
Adobe
M. Nottingham, Ed.
Akamai
J. Reschke, Ed.
greenbytes
July 15, 2013

Hypertext Transfer Protocol (HTTP/1.1): Caching
draft-ietf-httpbis-p6-cache-23

Abstract

The Hypertext Transfer Protocol (HTTP) is an application-level protocol for distributed, collaborative, hypertext information systems. This document defines requirements on HTTP caches and the associated header fields that control cache behavior or indicate cacheable response messages.

Editorial Note (To be removed by RFC Editor)

Discussion of this draft takes place on the HTTPBIS working group mailing list (ietf-http-wg@w3.org), which is archived at <http://lists.w3.org/Archives/Public/ietf-http-wg/>.

The current issues list is at <http://tools.ietf.org/wg/httpbis/trac/report/3> and related documents (including fancy diffs) can be found at <http://tools.ietf.org/wg/httpbis/>.

The changes in this draft are summarized in Appendix D.4.

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 http://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 January 16, 2014.

Copyright Notice

Table of Contents

1. Introduction

HTTP is typically used for distributed information systems, where performance can be improved by the use of response caches. This document defines aspects of HTTP/1.1 related to caching and reusing response messages.

An HTTP cache is a local store of response messages and the subsystem that controls storage, retrieval, and deletion of messages in it. A cache stores cacheable responses in order to reduce the response time and network bandwidth consumption on future, equivalent requests. Any client or server MAY employ a cache, though a cache cannot be used by a server that is acting as a tunnel.

A shared cache is a cache that stores responses to be reused by more than one user; shared caches are usually (but not always) deployed as a part of an intermediary. A private cache, in contrast, is dedicated to a single user.

The goal of caching in HTTP/1.1 is to significantly improve performance by reusing a prior response message to satisfy a current request. A stored response is considered "fresh", as defined in Section 4.1, if the response can be reused without "validation" (checking with the origin server to see if the cached response remains valid for this request). A fresh response can therefore reduce both latency and network overhead each time it is reused. When a cached response is not fresh, it might still be reusable if it can be freshened by validation (Section 4.2) or if the origin is unavailable (Section 4.1.4).

1.1. Conformance and Error Handling

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Conformance criteria and considerations regarding error handling are defined in Section 2.5 of [Part1].

1.2. Syntax Notation

This specification uses the Augmented Backus-Naur Form (ABNF) notation of [RFC5234] with the list rule extension defined in Section 1.2 of [Part1]. Appendix B describes rules imported from other documents. Appendix C shows the collected ABNF with the list rule expanded.

1.2.1. Delta Seconds

The delta-seconds rule specifies a non-negative integer, representing time in seconds.

```
delta-seconds  = 1*DIGIT
```

If a cache receives a delta-seconds value larger than the largest positive integer it can represent, or if any of its subsequent calculations overflows, it MUST consider the value to be 2147483648 (2^31). Recipients parsing a delta-seconds value MUST use an arithmetic type of at least 31 bits of range, and senders MUST NOT generate delta-seconds with a value greater than 2147483648.

2. Overview of Cache Operation

Proper cache operation preserves the semantics of HTTP transfers ([Part2]) while eliminating the transfer of information already held in the cache. Although caching is an entirely OPTIONAL feature of HTTP, we assume that reusing the cached response is desirable and that such reuse is the default behavior when no requirement or local configuration prevents it. Therefore, HTTP cache requirements are focused on preventing a cache from either storing a non-reusable response or reusing a stored response inappropriately, rather than mandating that caches always store and reuse particular responses.

Each cache entry consists of a cache key and one or more HTTP responses corresponding to prior requests that used the same key. The most common form of cache entry is a successful result of a retrieval request: i.e., a 200 (OK) response to a GET request, which contains a representation of the resource identified by the request target (Section 4.3.1 of [Part2]). However, it is also possible to cache permanent redirects, negative results (e.g., 404 (Not Found)), incomplete results (e.g., 206 (Partial Content)), and responses to methods other than GET if the method's definition allows such caching and defines something suitable for use as a cache key.

The primary cache key consists of the request method and target URI. However, since HTTP caches in common use today are typically limited to caching responses to GET, many caches simply decline other methods and use only the URI as the primary cache key.

If a request target is subject to content negotiation, its cache entry might consist of multiple stored responses, each differentiated by a secondary key for the values of the original request's selecting header fields (Section 4.3).

3. Storing Responses in Caches

A cache MUST NOT store a response to any request, unless:

- The request method is understood by the cache and defined as being cacheable, and
- the response status code is understood by the cache, and
- the "no-store" cache directive (see Section 7.2) does not appear in request or response header fields, and
- the "private" cache response directive (see Section 7.2.2.6) does not appear in the response, if the cache is shared, and
- the Authorization header field (see Section 4.1 of [Part7]) does not appear in the request, if the cache is shared, unless the response explicitly allows it (see Section 3.2), and
- the response either:
  - contains an Expires header field (see Section 7.3), or
  - contains a max-age response cache directive (see Section 7.2.2.8), or
  - contains a s-maxage response cache directive (see Section 7.2.2.9) and the cache is shared, or
  - contains a Cache Control Extension (see Section 7.2.3) that allows it to be cached, or
  - has a status code that is defined as cacheable (see Section 4.1.2), or
  - contains a public response cache directive (see Section 7.2.2.5).

Note that any of the requirements listed above can be overridden by a cache-control extension; see Section 7.2.3.

In this context, a cache has "understood" a request method or a response status code if it recognizes it and implements all specified caching-related behavior.

Note that, in normal operation, some caches will not store a response that has neither a cache validator nor an explicit expiration time, as such responses are not usually useful to store. However, caches

are not prohibited from storing such responses.

3.1. Storing Incomplete Responses

A response message is considered complete when all of the octets indicated by the message framing ([Part1]) are received prior to the connection being closed. If the request is GET, the response status is 200 (OK), and the entire response header block has been received, a cache MAY store an incomplete response message body if the cache entry is recorded as incomplete. Likewise, a 206 (Partial Content) response MAY be stored as if it were an incomplete 200 (OK) cache entry. However, a cache MUST NOT store incomplete or partial content responses if it does not support the Range and Content-Range header fields or if it does not understand the range units used in those fields.

A cache MAY complete a stored incomplete response by making a subsequent range request ([Part5]) and combining the successful response with the stored entry, as defined in Section 3.3. A cache MUST NOT use an incomplete response to answer requests unless the response has been made complete or the request is partial and specifies a range that is wholly within the incomplete response. A cache MUST NOT send a partial response to a client without explicitly marking it as such using the 206 (Partial Content) status code.

3.2. Storing Responses to Authenticated Requests

A shared cache MUST NOT use a cached response to a request with an Authorization header field (Section 4.1 of [Part7]) to satisfy any subsequent request unless a cache directive that allows such responses to be stored is present in the response.

In this specification, the following Cache-Control response directives (Section 7.2.2) have such an effect: must-revalidate, public, s-maxage.

Note that cached responses that contain the "must-revalidate" and/or "s-maxage" response directives are not allowed to be served stale (Section 4.1.4) by shared caches. In particular, a response with either "max-age=0, must-revalidate" or "s-maxage=0" cannot be used to satisfy a subsequent request without revalidating it on the origin server.

3.3. Combining Partial Content

A response might transfer only a partial representation if the connection closed prematurely or if the request used one or more Range specifiers ([Part5]). After several such transfers, a cache

might have received several ranges of the same representation. A cache MAY combine these ranges into a single stored response, and reuse that response to satisfy later requests, if they all share the same strong validator and the cache complies with the client requirements in Section 4.3 of [Part5].

When combining the new response with one or more stored responses, a cache MUST:

- delete any Warning header fields in the stored response with warn-code 1xx (see Section 7.5);
- retain any Warning header fields in the stored response with warn-code 2xx; and,
- use other header fields provided in the new response, aside from Content-Range, to replace all instances of the corresponding header fields in the stored response.

4. Constructing Responses from Caches

When presented with a request, a cache MUST NOT reuse a stored response, unless:

- The presented effective request URI (Section 5.5 of [Part1]) and that of the stored response match, and
- the request method associated with the stored response allows it to be used for the presented request, and
- selecting header fields nominated by the stored response (if any) match those presented (see Section 4.3), and
- the presented request does not contain the no-cache pragma (Section 7.4), nor the no-cache cache directive (Section 7.2.1), unless the stored response is successfully validated (Section 4.2), and
- the stored response does not contain the no-cache cache directive (Section 7.2.2.2), unless it is successfully validated (Section 4.2), and
- the stored response is either:
  - fresh (see Section 4.1), or
  - allowed to be served stale (see Section 4.1.4), or

* successfully validated (see Section 4.2).

Note that any of the requirements listed above can be overridden by a cache-control extension; see Section 7.2.3.

When a stored response is used to satisfy a request without validation, a cache MUST generate an Age header field (Section 7.1), replacing any present in the response with a value equal to the stored response's current_age; see Section 4.1.3.

A cache MUST write through requests with methods that are unsafe (Section 4.2.1 of [Part2]) to the origin server; i.e., a cache is not allowed to generate a reply to such a request before having forwarded the request and having received a corresponding response.

Also, note that unsafe requests might invalidate already stored responses; see Section 6.

When more than one suitable response is stored, a cache MUST use the most recent response (as determined by the Date header field). It can also forward the request with "Cache-Control: max-age=0" or "Cache-Control: no-cache" to disambiguate which response to use.

A cache that does not have a clock available MUST NOT use stored responses without revalidating them upon every use.

4.1. Freshness

A fresh response is one whose age has not yet exceeded its freshness lifetime. Conversely, a stale response is one where it has.

A response's freshness lifetime is the length of time between its generation by the origin server and its expiration time. An explicit expiration time is the time at which the origin server intends that a stored response can no longer be used by a cache without further validation, whereas a heuristic expiration time is assigned by a cache when no explicit expiriation time is available.

A response's age is the time that has passed since it was generated by, or successfully validated with, the origin server.

When a response is "fresh" in the cache, it can be used to satisfy subsequent requests without contacting the origin server, thereby improving efficiency.

The primary mechanism for determining freshness is for an origin server to provide an explicit expiration time in the future, using either the Expires header field (Section 7.3) or the max-age response

cache directive (Section 7.2.2.8). Generally, origin servers will assign future explicit expiration times to responses in the belief that the representation is not likely to change in a semantically significant way before the expiration time is reached.

If an origin server wishes to force a cache to validate every request, it can assign an explicit expiration time in the past to indicate that the response is already stale. Compliant caches will normally validate a stale cached response before reusing it for subsequent requests (see Section 4.1.4).

Since origin servers do not always provide explicit expiration times, caches are also allowed to use a heuristic to determine an expiration time under certain circumstances (see Section 4.1.2).

The calculation to determine if a response is fresh is:

```
response_is_fresh = (freshness_lifetime > current_age)
```

freshness_lifetime is defined in Section 4.1.1; current_age is defined in Section 4.1.3.

Clients can send the max-age or min-fresh cache directives in a request to constrain or relax freshness calculations for the corresponding response (Section 7.2.1).

When calculating freshness, to avoid common problems in date parsing:

- Although all date formats are specified to be case-sensitive, cache recipients SHOULD match day, week and timezone names case-insensitively.

- If a cache recipient's internal implementation of time has less resolution than the value of an HTTP-date, the recipient MUST internally represent a parsed Expires date as the nearest time equal to or earlier than the received value.

- Cache recipients MUST NOT allow local time zones to influence the calculation or comparison of an age or expiration time.

- Cache recipients SHOULD consider a date with a zone abbreviation other than "GMT" to be invalid for calculating expiration.

Note that freshness applies only to cache operation; it cannot be used to force a user agent to refresh its display or reload a resource. See Section 8 for an explanation of the difference between caches and history mechanisms.

4.1.1. Calculating Freshness Lifetime

A cache can calculate the freshness lifetime (denoted as freshness_lifetime) of a response by using the first match of:

- If the cache is shared and the s-maxage response cache directive (Section 7.2.2.9) is present, use its value, or
- If the max-age response cache directive (Section 7.2.2.8) is present, use its value, or
- If the Expires response header field (Section 7.3) is present, use its value minus the value of the Date response header field, or
- Otherwise, no explicit expiration time is present in the response. A heuristic freshness lifetime might be applicable; see Section 4.1.2.

Note that this calculation is not vulnerable to clock skew, since all of the information comes from the origin server.

When there is more than one value present for a given directive (e.g., two Expires header fields, multiple Cache-Control: max-age directives), the directive's value is considered invalid. Caches are encouraged to consider responses that have invalid freshness information to be stale.

4.1.2. Calculating Heuristic Freshness

Since origin servers do not always provide explicit expiration times, a cache MAY assign a heuristic expiration time when an explicit time is not specified, employing algorithms that use other header field values (such as the Last-Modified time) to estimate a plausible expiration time. This specification does not provide specific algorithms, but does impose worst-case constraints on their results.

A cache MUST NOT use heuristics to determine freshness when an explicit expiration time is present in the stored response. Because of the requirements in Section 3, this means that, effectively, heuristics can only be used on responses without explicit freshness whose status codes are defined as cacheable, and responses without explicit freshness that have been marked as explicitly cacheable (e.g., with a "public" response cache directive).

If the response has a Last-Modified header field (Section 2.2 of [Part4]), caches are encouraged to use a heuristic expiration value that is no more than some fraction of the interval since that time. A typical setting of this fraction might be 10%.

When a heuristic is used to calculate freshness lifetime, a cache SHOULD attach a Warning header field with a 113 warn-code to the response if its current_age is more than 24 hours and such a warning is not already present.

Note: Section 13.9 of [RFC2616] prohibited caches from calculating heuristic freshness for URIs with query components (i.e., those containing '?'). In practice, this has not been widely implemented. Therefore, origin servers are encouraged to send explicit directives (e.g., Cache-Control: no-cache) if they wish to preclude caching.

4.1.3. Calculating Age

The Age header field is used to convey an estimated age of the response message when obtained from a cache. The Age field value is the cache's estimate of the number of seconds since the response was generated or validated by the origin server. In essence, the Age value is the sum of the time that the response has been resident in each of the caches along the path from the origin server, plus the amount of time it has been in transit along network paths.

The following data is used for the age calculation:

age_value

The term "age_value" denotes the value of the Age header field (Section 7.1), in a form appropriate for arithmetic operation; or 0, if not available.

date_value

The term "date_value" denotes the value of the Date header field, in a form appropriate for arithmetic operations. See Section 7.1.1.2 of [Part2] for the definition of the Date header field, and for requirements regarding responses without it.

now

The term "now" means "the current value of the clock at the host performing the calculation". A host ought to use NTP ([RFC1305]) or some similar protocol to synchronize its clocks to Coordinated Universal Time.

request_time

The current value of the clock at the host at the time the request resulting in the stored response was made.

response_time

The current value of the clock at the host at the time the response was received.

A response's age can be calculated in two entirely independent ways:

1. the "apparent_age": response_time minus date_value, if the local clock is reasonably well synchronized to the origin server's clock. If the result is negative, the result is replaced by zero.

2. the "corrected_age_value", if all of the caches along the response path implement HTTP/1.1. A cache MUST interpret this value relative to the time the request was initiated, not the time that the response was received.

```
apparent_age = max(0, response_time - date_value);

response_delay = response_time - request_time;
corrected_age_value = age_value + response_delay;
```

These are combined as

```
corrected_initial_age = max(apparent_age, corrected_age_value);
```

unless the cache is confident in the value of the Age header field (e.g., because there are no HTTP/1.0 hops in the Via header field), in which case the corrected_age_value MAY be used as the corrected_initial_age.

The current_age of a stored response can then be calculated by adding the amount of time (in seconds) since the stored response was last validated by the origin server to the corrected_initial_age.

```
resident_time = now - response_time;
current_age = corrected_initial_age + resident_time;
```

4.1.4. Serving Stale Responses

A "stale" response is one that either has explicit expiry information or is allowed to have heuristic expiry calculated, but is not fresh according to the calculations in Section 4.1.

A cache MUST NOT generate a stale response if it is prohibited by an explicit in-protocol directive (e.g., by a "no-store" or "no-cache" cache directive, a "must-revalidate" cache-response-directive, or an

applicable "s-maxage" or "proxy-revalidate" cache-response-directive; see Section 7.2.2).

A cache MUST NOT send stale responses unless it is disconnected (i.e., it cannot contact the origin server or otherwise find a forward path) or doing so is explicitly allowed (e.g., by the max-stale request directive; see Section 7.2.1).

A cache SHOULD append a Warning header field with the 110 warn-code (see Section 7.5) to stale responses. Likewise, a cache SHOULD add the 112 warn-code to stale responses if the cache is disconnected.

Note that if a cache receives a first-hand response (one where the freshness model is not in use; i.e., its age is 0, whether it is an entire response, or a 304 (Not Modified) response) that it would normally forward to the requesting client, and the received response is no longer fresh, the cache MAY forward it to the requesting client without adding a new Warning (but without removing any existing Warning header fields). A cache ought not attempt to validate a response simply because that response became stale in transit.

4.2. Validation

When a cache has one or more stored responses for a requested URI, but cannot serve any of them (e.g., because they are not fresh, or one cannot be selected; see Section 4.3), it can use the conditional request mechanism [Part4] in the forwarded request to give the origin server an opportunity to both select a valid stored response to be used, and to update it. This process is known as "validating" or "revalidating" the stored response.

When sending such a conditional request, a cache adds a validator (or more than one), that is used to find out whether a stored response is an equivalent copy of a current representation of the resource.

One such validator is the If-Modified-Since header field, whose value is that of the Last-Modified header field from the selected (see Section 4.3) stored response, if available.

Another is the If-None-Match header field, whose value is that of the ETag header field(s) from relevant responses stored for the primary cache key, if present. However, if any of the stored responses contains only partial content, the cache ought not include its entity-tag in the If-None-Match header field unless the request is for a range that would be fully satisfied by that stored response.

Cache handling of a response to a conditional request is dependent upon its status code:

- A 304 (Not Modified) response status code indicates that the stored response can be updated and reused; see Section 4.2.1.

- A full response (i.e., one with a payload body) indicates that none of the stored responses nominated in the conditional request is suitable. Instead, the cache can use the full response to satisfy the request and MAY replace the stored response(s).

- However, if a cache receives a 5xx (Server Error) response while attempting to validate a response, it can either forward this response to the requesting client, or act as if the server failed to respond. In the latter case, it can send a previously stored response (see Section 4.1.4).

4.2.1. Freshening Stored Responses upon Validation

When a cache receives a 304 (Not Modified) response and already has one or more stored 200 (OK) responses for the same cache key, the cache needs to identify which of the stored responses are updated by this new response and then update the stored response(s) with the new information provided in the 304 response.

The stored response to update is identified by using the first match (if any) of:

- If the new response contains a strong validator (see Section 2.1 of [Part4]), then that strong validator identifies the selected representation for update. All of the stored responses with the same strong validator are selected. If none of the stored responses contain the same strong validator, then the cache MUST NOT use the new response to update any stored responses.

- If the new response contains a weak validator and that validator corresponds to one of the cache's stored responses, then the most recent of those matching stored responses is selected for update.

- If the new response does not include any form of validator (such as in the case where a client generates an If-Modified-Since request from a source other than the Last-Modified response header field), and there is only one stored response, and that stored response also lacks a validator, then that stored response is selected for update.

If a stored response is selected for update, the cache MUST:

- delete any Warning header fields in the stored response with warn-code 1xx (see Section 7.5);

- retain any Warning header fields in the stored response with warn-code 2xx; and,

- use other header fields provided in the 304 (Not Modified) response to replace all instances of the corresponding header fields in the stored response.

4.3. Calculating Secondary Keys with Vary

When a cache receives a request that can be satisfied by a stored response that has a Vary header field (Section 7.1.4 of [Part2]), it MUST NOT use that response unless all of the selecting header fields nominated by the Vary header field match in both the original request (i.e., that associated with the stored response), and the presented request.

The selecting header fields from two requests are defined to match if and only if those in the first request can be transformed to those in the second request by applying any of the following:

- adding or removing whitespace, where allowed in the header field's syntax

- combining multiple header fields with the same field name (see Section 3.2 of [Part1])

- normalizing both header field values in a way that is known to have identical semantics, according to the header field's specification (e.g., re-ordering field values when order is not significant; case-normalization, where values are defined to be case-insensitive)

If (after any normalization that might take place) a header field is absent from a request, it can only match another request if it is also absent there.

A Vary header field-value of "*" always fails to match.

The stored response with matching selecting header fields is known as the selected response.

If multiple selected responses are available (potentially including responses without a Vary header field), the cache will need to choose one to use. When a selecting header field has a known mechanism for doing so (e.g., qvalues on Accept and similar request header fields), that mechanism MAY be used to select preferred responses; of the remainder, the most recent response (as determined by the Date header field) is used, as per Section 4.

If no selected response is available, the cache cannot satisfy the presented request. Typically, it is forwarded to the origin server in a (possibly conditional; see Section 4.2) request.

5. Updating Caches with HEAD Responses

A response to the HEAD method is identical to what an equivalent request made with a GET would have been, except it lacks a body. This property of HEAD responses is used to both invalidate and update cached GET responses.

If one or more stored GET responses can be selected (as per Section 4.3) for a HEAD request, and the Content-Length, ETag or Last-Modified value of a HEAD response differs from that in a selected GET response, the cache MUST consider that selected response to be stale.

If the Content-Length, ETag and Last-Modified values of a HEAD response (when present) are the same as that in a selected GET response (as per Section 4.3), the cache SHOULD update the remaining header fields in the stored response using the following rules:

- delete any Warning header fields in the stored response with warn-code 1xx (see Section 7.5);
- retain any Warning header fields in the stored response with warn-code 2xx; and,
- use other header fields provided in the response to replace all instances of the corresponding header fields in the stored response.

6. Request Methods that Invalidate

Because unsafe request methods (Section 4.2.1 of [Part2]) such as PUT, POST or DELETE have the potential for changing state on the origin server, intervening caches can use them to keep their contents up-to-date.

A cache MUST invalidate the effective Request URI (Section 5.5 of [Part1]) as well as the URI(s) in the Location and Content-Location response header fields (if present) when a non-error response to a request with an unsafe method is received.

However, a cache MUST NOT invalidate a URI from a Location or Content-Location response header field if the host part of that URI differs from the host part in the effective request URI (Section 5.5 of [Part1]). This helps prevent denial of service attacks.

A cache MUST invalidate the effective request URI (Section 5.5 of [Part1]) when it receives a non-error response to a request with a method whose safety is unknown.

Here, a "non-error response" is one with a 2xx (Successful) or 3xx (Redirection) status code. "Invalidate" means that the cache will either remove all stored responses related to the effective request URI, or will mark these as "invalid" and in need of a mandatory validation before they can be sent in response to a subsequent request.

Note that this does not guarantee that all appropriate responses are invalidated. For example, a state-changing request might invalidate responses in the caches it travels through, but relevant responses still might be stored in other caches that it has not.

7. Header Field Definitions

This section defines the syntax and semantics of HTTP/1.1 header fields related to caching.

7.1. Age

The "Age" header field conveys the sender's estimate of the amount of time since the response was generated or successfully validated at the origin server. Age values are calculated as specified in Section 4.1.3.

```
Age = delta-seconds
```

Age field-values are non-negative integers, representing time in seconds (see Section 1.2.1).

The presence of an Age header field in a response implies that a response is not first-hand. However, the converse is not true, since HTTP/1.0 caches might not implement the Age header field.

7.2. Cache-Control

The "Cache-Control" header field is used to specify directives for caches along the request/response chain. Such cache directives are unidirectional in that the presence of a directive in a request does not imply that the same directive is to be given in the response.

A cache MUST obey the requirements of the Cache-Control directives defined in this section. See Section 7.2.3 for information about how Cache-Control directives defined elsewhere are handled.

Note: Some HTTP/1.0 caches might not implement Cache-Control.

A proxy, whether or not it implements a cache, MUST pass cache directives through in forwarded messages, regardless of their significance to that application, since the directives might be applicable to all recipients along the request/response chain. It is not possible to target a directive to a specific cache.

Cache directives are identified by a token, to be compared case-insensitively, and have an optional argument, that can use both token and quoted-string syntax. For the directives defined below that define arguments, recipients ought to accept both forms, even if one is documented to be preferred. For any directive not defined by this specification, recipients MUST accept both forms.

```
Cache-Control   = 1#cache-directive

cache-directive = token [ "=" ( token / quoted-string ) ]
```

For the cache directives defined below, no argument is defined (nor allowed) unless stated otherwise.

7.2.1. Request Cache-Control Directives

7.2.1.1. max-age

Argument syntax:

delta-seconds (see Section 1.2.1)

The "max-age" request directive indicates that the client is unwilling to accept a response whose age is greater than the specified number of seconds. Unless the max-stale request directive is also present, the client is not willing to accept a stale response.

Note: This directive uses the token form of the argument syntax; e.g., 'max-age=5', not 'max-age="5"'. Senders SHOULD NOT use the quoted-string form.

7.2.1.2. max-stale

Argument syntax:

delta-seconds (see Section 1.2.1)

The "max-stale" request directive indicates that the client is willing to accept a response that has exceeded its freshness

lifetime. If max-stale is assigned a value, then the client is willing to accept a response that has exceeded its freshness lifetime by no more than the specified number of seconds. If no value is assigned to max-stale, then the client is willing to accept a stale response of any age.

Note: This directive uses the token form of the argument syntax; e.g., 'max-stale=10', not 'max-stale="10"'. Senders SHOULD NOT use the quoted-string form.

7.2.1.3. min-fresh

Argument syntax:

delta-seconds (see Section 1.2.1)

The "min-fresh" request directive indicates that the client is willing to accept a response whose freshness lifetime is no less than its current age plus the specified time in seconds. That is, the client wants a response that will still be fresh for at least the specified number of seconds.

Note: This directive uses the token form of the argument syntax; e.g., 'min-fresh=20', not 'min-fresh="20"'. Senders SHOULD NOT use the quoted-string form.

7.2.1.4. no-cache

The "no-cache" request directive indicates that a cache MUST NOT use a stored response to satisfy the request without successful validation on the origin server.

7.2.1.5. no-store

The "no-store" request directive indicates that a cache MUST NOT store any part of either this request or any response to it. This directive applies to both private and shared caches. "MUST NOT store" in this context means that the cache MUST NOT intentionally store the information in non-volatile storage, and MUST make a best-effort attempt to remove the information from volatile storage as promptly as possible after forwarding it.

This directive is NOT a reliable or sufficient mechanism for ensuring privacy. In particular, malicious or compromised caches might not recognize or obey this directive, and communications networks might be vulnerable to eavesdropping.

Note that if a request containing this directive is satisfied from a

cache, the no-store request directive does not apply to the already stored response.

7.2.1.6. no-transform

The "no-transform" request directive indicates that an intermediary (whether or not it implements a cache) MUST NOT transform the payload, as defined in Section 5.7.2 of [Part1].

7.2.1.7. only-if-cached

The "only-if-cached" request directive indicates that the client only wishes to obtain a stored response. If it receives this directive, a cache SHOULD either respond using a stored response that is consistent with the other constraints of the request, or respond with a 504 (Gateway Timeout) status code. If a group of caches is being operated as a unified system with good internal connectivity, a member cache MAY forward such a request within that group of caches.

7.2.2. Response Cache-Control Directives

7.2.2.1. must-revalidate

The "must-revalidate" response directive indicates that once it has become stale, a cache MUST NOT use the response to satisfy subsequent requests without successful validation on the origin server.

The must-revalidate directive is necessary to support reliable operation for certain protocol features. In all circumstances a cache MUST obey the must-revalidate directive; in particular, if a cache cannot reach the origin server for any reason, it MUST generate a 504 (Gateway Timeout) response.

The must-revalidate directive ought to be used by servers if and only if failure to validate a request on the representation could result in incorrect operation, such as a silently unexecuted financial transaction.

7.2.2.2. no-cache

Argument syntax:

```
#field-name
```

The "no-cache" response directive indicates that the response MUST NOT be used to satisfy a subsequent request without successful validation on the origin server. This allows an origin server to prevent a cache from using it to satisfy a request without contacting

it, even by caches that have been configured to send stale responses.

If the no-cache response directive specifies one or more field-names, then a cache MAY use the response to satisfy a subsequent request, subject to any other restrictions on caching. However, any header fields in the response that have the field-name(s) listed MUST NOT be sent in the response to a subsequent request without successful revalidation with the origin server. This allows an origin server to prevent the re-use of certain header fields in a response, while still allowing caching of the rest of the response.

The field-names given are not limited to the set of header fields defined by this specification. Field names are case-insensitive.

Note: Although it has been back-ported to many implementations, some HTTP/1.0 caches will not recognize or obey this directive. Also, no-cache response directives with field-names are often handled by caches as if an unqualified no-cache directive was received; i.e., the special handling for the qualified form is not widely implemented.

Note: This directive uses the quoted-string form of the argument syntax. Senders SHOULD NOT use the token form (even if quoting appears not to be needed for single-entry lists).

7.2.2.3. no-store

The "no-store" response directive indicates that a cache MUST NOT store any part of either the immediate request or response. This directive applies to both private and shared caches. "MUST NOT store" in this context means that the cache MUST NOT intentionally store the information in non-volatile storage, and MUST make a best-effort attempt to remove the information from volatile storage as promptly as possible after forwarding it.

This directive is NOT a reliable or sufficient mechanism for ensuring privacy. In particular, malicious or compromised caches might not recognize or obey this directive, and communications networks might be vulnerable to eavesdropping.

7.2.2.4. no-transform

The "no-transform" response directive indicates that an intermediary (regardless of whether it implements a cache) MUST NOT transform the payload, as defined in Section 5.7.2 of [Part1].

7.2.2.5. public

The "public" response directive indicates that any cache MAY store the response, even if the response would normally be non-cacheable or cacheable only within a non-shared cache. (See Section 3.2 for additional details related to the use of public in response to a request containing Authorization, and Section 3 for details of how public affects responses that would normally not be stored, due to their status codes not being defined as cacheable.)

7.2.2.6. private

Argument syntax:

```
#field-name
```

The "private" response directive indicates that the response message is intended for a single user and MUST NOT be stored by a shared cache. A private cache MAY store the response and reuse it for later requests, even if the response would normally be non-cacheable.

If the private response directive specifies one or more field-names, this requirement is limited to the field-values associated with the listed response header fields. That is, a shared cache MUST NOT store the specified field-names(s), whereas it MAY store the remainder of the response message.

The field-names given are not limited to the set of header fields defined by this specification. Field names are case-insensitive.

Note: This usage of the word "private" only controls where the response can be stored; it cannot ensure the privacy of the message content. Also, private response directives with field-names are often handled by caches as if an unqualified private directive was received; i.e., the special handling for the qualified form is not widely implemented.

Note: This directive uses the quoted-string form of the argument syntax. Senders SHOULD NOT use the token form (even if quoting appears not to be needed for single-entry lists).

7.2.2.7. proxy-revalidate

The "proxy-revalidate" response directive has the same meaning as the must-revalidate response directive, except that it does not apply to private caches.

7.2.2.8. max-age

Argument syntax:

delta-seconds (see Section 1.2.1)

The "max-age" response directive indicates that the response is to be considered stale after its age is greater than the specified number of seconds.

Note: This directive uses the token form of the argument syntax; e.g., 'max-age=5', not 'max-age="5"'. Senders SHOULD NOT use the quoted-string form.

7.2.2.9. s-maxage

Argument syntax:

delta-seconds (see Section 1.2.1)

The "s-maxage" response directive indicates that, in shared caches, the maximum age specified by this directive overrides the maximum age specified by either the max-age directive or the Expires header field. The s-maxage directive also implies the semantics of the proxy-revalidate response directive.

Note: This directive uses the token form of the argument syntax; e.g., 's-maxage=10', not 's-maxage="10"'. Senders SHOULD NOT use the quoted-string form.

7.2.3. Cache Control Extensions

The Cache-Control header field can be extended through the use of one or more cache-extension tokens, each with an optional value.

Informational extensions (those that do not require a change in cache behavior) can be added without changing the semantics of other directives. Behavioral extensions are designed to work by acting as modifiers to the existing base of cache directives.

Both the new directive and the standard directive are supplied, such that applications that do not understand the new directive will default to the behavior specified by the standard directive, and those that understand the new directive will recognize it as modifying the requirements associated with the standard directive. In this way, extensions to the cache-control directives can be made without requiring changes to the base protocol.

This extension mechanism depends on an HTTP cache obeying all of the cache-control directives defined for its native HTTP-version, obeying certain extensions, and ignoring all directives that it does not understand.

For example, consider a hypothetical new response directive called "community" that acts as a modifier to the private directive. We define this new directive to mean that, in addition to any private cache, any cache that is shared only by members of the community named within its value is allowed to cache the response. An origin server wishing to allow the UCI community to use an otherwise private response in their shared cache(s) could do so by including

```
Cache-Control: private, community="UCI"
```

A cache seeing this header field will act correctly even if the cache does not understand the community cache-extension, since it will also see and understand the private directive and thus default to the safe behavior.

A cache MUST ignore unrecognized cache directives; it is assumed that any cache directive likely to be unrecognized by an HTTP/1.1 cache will be combined with standard directives (or the response's default cacheability) such that the cache behavior will remain minimally correct even if the cache does not understand the extension(s).

7.3. Expires

The "Expires" header field gives the date/time after which the response is considered stale. See Section 4.1 for further discussion of the freshness model.

The presence of an Expires field does not imply that the original resource will change or cease to exist at, before, or after that time.

The Expires value is an HTTP-date timestamp, as defined in Section 7.1.1.1 of [Part2].

```
Expires = HTTP-date
```

For example

```
Expires: Thu, 01 Dec 1994 16:00:00 GMT
```

A cache recipient MUST interpret invalid date formats, especially the value "0", as representing a time in the past (i.e., "already expired").

If a response includes a Cache-Control field with the max-age directive (Section 7.2.2.8), a recipient MUST ignore the Expires field. Likewise, if a response includes the s-maxage directive (Section 7.2.2.9), a shared cache recipient MUST ignore the Expires field. In both these cases, the value in Expires is only intended for recipients that have not yet implemented the Cache-Control field.

An origin server without a clock MUST NOT generate an Expires field unless its value represents a fixed time in the past (always expired) or its value has been associated with the resource by a system or user with a reliable clock.

Historically, HTTP required the Expires field-value to be no more than a year in the future. While longer freshness lifetimes are no longer prohibited, extremely large values have been demonstrated to cause problems (e.g., clock overflows due to use of 32-bit integers for time values), and many caches will evict a response far sooner than that.

7.4. Pragma

The "Pragma" header field allows backwards compatibility with HTTP/1.0 caches, so that clients can specify a "no-cache" request that they will understand (as Cache-Control was not defined until HTTP/1.1). When the Cache-Control header field is also present and understood in a request, Pragma is ignored.

In HTTP/1.0, Pragma was defined as an extensible field for implementation-specified directives for recipients. This specification deprecates such extensions to improve interoperability.

```
Pragma           = 1#pragma-directive
pragma-directive = "no-cache" / extension-pragma
extension-pragma = token [ "=" ( token / quoted-string ) ]
```

When the Cache-Control header field is not present in a request, caches MUST consider the no-cache request pragma-directive as having the same effect as if "Cache-Control: no-cache" were present (see Section 7.2.1).

When sending a no-cache request, a client ought to include both the pragma and cache-control directives, unless Cache-Control: no-cache is purposefully omitted to target other Cache-Control response directives at HTTP/1.1 caches. For example:

```
GET / HTTP/1.1
Host: www.example.com
Cache-Control: max-age=30
```

```
Pragma: no-cache
```

will constrain HTTP/1.1 caches to serve a response no older than 30 seconds, while precluding implementations that do not understand Cache-Control from serving a cached response.

> Note: Because the meaning of "Pragma: no-cache" in responses is not specified, it does not provide a reliable replacement for "Cache-Control: no-cache" in them.

7.5. Warning

The "Warning" header field is used to carry additional information about the status or transformation of a message that might not be reflected in the message. This information is typically used to warn about possible incorrectness introduced by caching operations or transformations applied to the payload of the message.

Warnings can be used for other purposes, both cache-related and otherwise. The use of a warning, rather than an error status code, distinguishes these responses from true failures.

Warning header fields can in general be applied to any message, however some warn-codes are specific to caches and can only be applied to response messages.

```
Warning       = 1#warning-value

warning-value = warn-code SP warn-agent SP warn-text
                                        [SP warn-date]

warn-code  = 3DIGIT
warn-agent = ( uri-host [ ":" port ] ) / pseudonym
                ; the name or pseudonym of the server adding
                ; the Warning header field, for use in debugging
warn-text  = quoted-string
warn-date  = DQUOTE HTTP-date DQUOTE
```

Multiple warnings can be attached to a response (either by the origin server or by a cache), including multiple warnings with the same code number, only differing in warn-text.

When this occurs, the user agent SHOULD inform the user of as many of them as possible, in the order that they appear in the response.

Systems that generate multiple Warning header fields are encouraged to order them with this user agent behavior in mind. New Warning

header fields are added after any existing Warning header fields.

Warnings are assigned three digit warn-codes. The first digit indicates whether the Warning is required to be deleted from a stored response after validation:

- 1xx Warnings describe the freshness or validation status of the response, and so MUST be deleted by a cache after validation. They can only be generated by a cache when validating a cached entry, and MUST NOT be generated in any other situation.

- 2xx Warnings describe some aspect of the representation that is not rectified by a validation (for example, a lossy compression of the representation) and MUST NOT be deleted by a cache after validation, unless a full response is sent, in which case they MUST be.

If an implementation sends a message with one or more Warning header fields to a receiver whose version is HTTP/1.0 or lower, then the sender MUST include in each warning-value a warn-date that matches the Date header field in the message.

If a system receives a message with a warning-value that includes a warn-date, and that warn-date is different from the Date value in the response, then that warning-value MUST be deleted from the message before storing, forwarding, or using it. (preventing the consequences of naive caching of Warning header fields.) If all of the warning-values are deleted for this reason, the Warning header field MUST be deleted as well.

The following warn-codes are defined by this specification, each with a recommended warn-text in English, and a description of its meaning.

7.5.1. 110 Response is Stale

A cache SHOULD generate this whenever the sent response is stale.

7.5.2. 111 Revalidation Failed

A cache SHOULD generate this when sending a stale response because an attempt to validate the response failed, due to an inability to reach the server.

7.5.3. 112 Disconnected Operation

A cache SHOULD generate this if it is intentionally disconnected from the rest of the network for a period of time.

7.5.4. 113 Heuristic Expiration

A cache SHOULD generate this if it heuristically chose a freshness lifetime greater than 24 hours and the response's age is greater than 24 hours.

7.5.5. 199 Miscellaneous Warning

The warning text can include arbitrary information to be presented to a human user, or logged. A system receiving this warning MUST NOT take any automated action, besides presenting the warning to the user.

7.5.6. 214 Transformation Applied

MUST be added by a proxy if it applies any transformation to the representation, such as changing the content-coding, media-type, or modifying the representation data, unless this Warning code already appears in the response.

7.5.7. 299 Miscellaneous Persistent Warning

The warning text can include arbitrary information to be presented to a human user, or logged. A system receiving this warning MUST NOT take any automated action.

7.5.8. Warn Code Extensions

Extension warn codes can be defined; see Section 9.2.1 for details.

8. History Lists

User agents often have history mechanisms, such as "Back" buttons and history lists, that can be used to redisplay a representation retrieved earlier in a session.

The freshness model (Section 4.1) does not necessarily apply to history mechanisms. I.e., a history mechanism can display a previous representation even if it has expired.

This does not prohibit the history mechanism from telling the user that a view might be stale, or from honoring cache directives (e.g., Cache-Control: no-store).

9. IANA Considerations

9.1. Cache Directive Registry

The HTTP Cache Directive Registry defines the name space for the cache directives. It will be created and maintained at <http://www.iana.org/assignments/http-cache-directives>.

9.1.1. Procedure

A registration MUST include the following fields:

- Cache Directive Name
- Pointer to specification text

Values to be added to this name space require IETF Review (see [RFC5226], Section 4.1).

9.1.2. Considerations for New Cache Control Directives

New extension directives ought to consider defining:

- What it means for a directive to be specified multiple times,
- When the directive does not take an argument, what it means when an argument is present,
- When the directive requires an argument, what it means when it is missing,
- Whether the directive is specific to requests, responses, or able to be used in either.

See also Section 7.2.3.

9.1.3. Registrations

The HTTP Cache Directive Registry shall be populated with the registrations below:

Cache Directive	Reference
max-age	Section 7.2.1.1, Section 7.2.2.8
max-stale	Section 7.2.1.2
min-fresh	Section 7.2.1.3
must-revalidate	Section 7.2.2.1
no-cache	Section 7.2.1.4, Section 7.2.2.2
no-store	Section 7.2.1.5, Section 7.2.2.3
no-transform	Section 7.2.1.6, Section 7.2.2.4
only-if-cached	Section 7.2.1.7
private	Section 7.2.2.6
proxy-revalidate	Section 7.2.2.7
public	Section 7.2.2.5
s-maxage	Section 7.2.2.9
stale-if-error	[RFC5861], Section 4
stale-while-revalidate	[RFC5861], Section 3

9.2. Warn Code Registry

The HTTP Warn Code Registry defines the name space for warn codes. It will be created and maintained at <http://www.iana.org/assignments/http-warn-codes>.

9.2.1. Procedure

A registration MUST include the following fields:

- Warn Code (3 digits)
- Short Description
- Pointer to specification text

Values to be added to this name space require IETF Review (see [RFC5226], Section 4.1).

9.2.2. Registrations

The HTTP Warn Code Registry shall be populated with the registrations below:

Warn Code	Short Description	Reference
110	Response is Stale	Section 7.5.1
111	Revalidation Failed	Section 7.5.2
112	Disconnected Operation	Section 7.5.3
113	Heuristic Expiration	Section 7.5.4
199	Miscellaneous Warning	Section 7.5.5
214	Transformation Applied	Section 7.5.6
299	Miscellaneous Persistent Warning	Section 7.5.7

9.3. Header Field Registration

HTTP header fields are registered within the Message Header Field Registry maintained at <http://www.iana.org/assignments/message-headers/message-header-index.html>.

This document defines the following HTTP header fields, so their associated registry entries shall be updated according to the permanent registrations below (see [BCP90]):

Header Field Name	Protocol	Status	Reference
Age	http	standard	Section 7.1
Cache-Control	http	standard	Section 7.2
Expires	http	standard	Section 7.3
Pragma	http	standard	Section 7.4
Warning	http	standard	Section 7.5

The change controller is: "IETF (iesg@ietf.org) - Internet Engineering Task Force".

10. Security Considerations

This section is meant to inform developers, information providers, and users of known security concerns specific to HTTP/1.1 caching. More general security considerations are addressed in HTTP messaging [Part1] and semantics [Part2].

Caches expose additional potential vulnerabilities, since the contents of the cache represent an attractive target for malicious exploitation. Because cache contents persist after an HTTP request is complete, an attack on the cache can reveal information long after a user believes that the information has been removed from the network. Therefore, cache contents need to be protected as sensitive

information.

Furthermore, the very use of a cache can bring about privacy concerns. For example, if two users share a cache, and the first one browses to a site, the second may be able to detect that the other has been to that site, because the resources from it load more quickly, thanks to the cache.

Implementation flaws might allow attackers to insert content into a cache ("cache poisoning"), leading to compromise of clients that trust that content. Because of their nature, these attacks are difficult to mitigate.

Likewise, implementation flaws (as well as misunderstanding of cache operation) might lead to caching of sensitive information (e.g., authentication credentials) that is thought to be private, exposing it to unauthorized parties.

Note that the Set-Cookie response header field [RFC6265] does not inhibit caching; a cacheable response with a Set-Cookie header field can be (and often is) used to satisfy subsequent requests to caches. Servers who wish to control caching of these responses are encouraged to emit appropriate Cache-Control response header fields.

11. Acknowledgments

See Section 9 of [Part1].

12. References

12.1. Normative References

[Part1] Fielding, R., Ed. and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing", draft-ietf-httpbis-p1-messaging-23 (work in progress), July 2013.

[Part2] Fielding, R., Ed. and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content", draft-ietf-httpbis-p2-semantics-23 (work in progress), July 2013.

[Part4] Fielding, R., Ed. and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Conditional Requests", draft-ietf-httpbis-p4-conditional-23 (work in progress), July 2013.

[Part5] Fielding, R., Ed., Lafon, Y., Ed., and J. Reschke, Ed.,

"Hypertext Transfer Protocol (HTTP/1.1): Range Requests", draft-ietf-httpbis-p5-range-23 (work in progress), July 2013.

[Part7] Fielding, R., Ed. and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Authentication", draft-ietf-httpbis-p7-auth-23 (work in progress), July 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5234] Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

12.2. Informative References

[BCP90] Klyne, G., Nottingham, M., and J. Mogul, "Registration Procedures for Message Header Fields", BCP 90, RFC 3864, September 2004.

[RFC1305] Mills, D., "Network Time Protocol (Version 3) Specification, Implementation", RFC 1305, March 1992.

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5861] Nottingham, M., "HTTP Cache-Control Extensions for Stale Content", RFC 5861, April 2010.

[RFC6265] Barth, A., "HTTP State Management Mechanism", RFC 6265, April 2011.

Appendix A. Changes from RFC 2616

Caching-related text has been substantially rewritten for clarity.

The algorithm for calculating age is now less conservative. (Section 4.1.3)

Caches are now required to handle dates with timezones as if they're invalid, because it's not possible to accurately guess. (Section 4.1.3)

The Content-Location response header field is no longer used to determine the appropriate response to use when validating. (Section 4.2)

The algorithm for selecting a cached negotiated response to use has been clarified in several ways. In particular, it now explicitly allows header-specific canonicalization when processing selecting header fields. (Section 4.3)

Requirements regarding denial of service attack avoidance when performing invalidation have been clarified. (Section 6)

Cache invalidation only occurs when a successful response is received. (Section 6)

The conditions under which an authenticated response can be cached have been clarified. (Section 3.2)

The one-year limit on Expires header field values has been removed; instead, the reasoning for using a sensible value is given. (Section 7.3)

The Pragma header field is now only defined for backwards compatibility; future pragmas are deprecated. (Section 7.4)

Cache directives are explicitly defined to be case-insensitive. (Section 7.2)

Handling of multiple instances of cache directives when only one is expected is now defined. (Section 7.2)

The qualified forms of the private and no-cache cache directives are noted to not be widely implemented; e.g., "private=foo" is interpreted by many caches as simply "private". Additionally, the meaning of the qualified form of no-cache has been clarified. (Section 7.2.2)

The "no-store" cache request directive doesn't apply to responses; i.e., a cache can satisfy a request with no-store on it, and does not invalidate it. (Section 7.2.1.5)

The "no-cache" response cache directive's meaning has been clarified. (Section 7.2.2.2)

New status codes can now define that caches are allowed to use heuristic freshness with them. (Section 4.1.2)

Caches are now allow to calculate heuristic freshness for URLs with

query components. (Section 4.1.2)

Some requirements regarding production of the Warning header fields have been relaxed, as it is not widely implemented. Furthermore, the Warning header field no longer uses RFC 2047 encoding, nor allows multiple languages, as these aspects were not implemented. (Section 7.5)

This specification introduces the Cache Directive and Warn Code Registries, and defines considerations for new cache directives. (Section 7.2.3 and Section 7.5.8)

Appendix B. Imported ABNF

The following core rules are included by reference, as defined in Appendix B.1 of [RFC5234]: ALPHA (letters), CR (carriage return), CRLF (CR LF), CTL (controls), DIGIT (decimal 0-9), DQUOTE (double quote), HEXDIG (hexadecimal 0-9/A-F/a-f), LF (line feed), OCTET (any 8-bit sequence of data), SP (space), and VCHAR (any visible US-ASCII character).

The rules below are defined in [Part1]:

```
  OWS           = <OWS, defined in [Part1], Section 3.2.3>
  field-name    = <field-name, defined in [Part1], Section 3.2>
  quoted-string = <quoted-string, defined in [Part1], Section 3.2.6>
  token         = <token, defined in [Part1], Section 3.2.6>

  port          = <port, defined in [Part1], Section 2.7>
  pseudonym     = <pseudonym, defined in [Part1], Section 5.7.1>
  uri-host      = <uri-host, defined in [Part1], Section 2.7>
```

The rules below are defined in other parts:

```
  HTTP-date     = <HTTP-date, defined in [Part2], Section 7.1.1.1>
```

Appendix C. Collected ABNF

In the collected ABNF below, list rules are expanded as per Section 1.2 of [Part1].

```
   Age = delta-seconds

   Cache-Control = *( "," OWS ) cache-directive *( OWS "," [ OWS
    cache-directive ] )

   Expires = HTTP-date

   HTTP-date = <HTTP-date, defined in [Part2], Section 7.1.1.1>

   OWS = <OWS, defined in [Part1], Section 3.2.3>

   Pragma = *( "," OWS ) pragma-directive *( OWS "," [ OWS
    pragma-directive ] )

   Warning = *( "," OWS ) warning-value *( OWS "," [ OWS warning-value ]
    )

   cache-directive = token [ "=" ( token / quoted-string ) ]

   delta-seconds = 1*DIGIT

   extension-pragma = token [ "=" ( token / quoted-string ) ]

   field-name = <field-name, defined in [Part1], Section 3.2>

   port = <port, defined in [Part1], Section 2.7>
   pragma-directive = "no-cache" / extension-pragma
   pseudonym = <pseudonym, defined in [Part1], Section 5.7.1>

   quoted-string = <quoted-string, defined in [Part1], Section 3.2.6>

   token = <token, defined in [Part1], Section 3.2.6>

   uri-host = <uri-host, defined in [Part1], Section 2.7>

   warn-agent = ( uri-host [ ":" port ] ) / pseudonym
   warn-code = 3DIGIT
   warn-date = DQUOTE HTTP-date DQUOTE
   warn-text = quoted-string
   warning-value = warn-code SP warn-agent SP warn-text [ SP warn-date
    ]
```

Appendix D. Change Log (to be removed by RFC Editor before publication)

Changes up to the first Working Group Last Call draft are summarized in <http://trac.tools.ietf.org/html/draft-ietf-httpbis-p6-cache-19#appendix-C>.

D.1. Since draft-ietf-httpbis-p6-cache-19

Closed issues:

- <http://tools.ietf.org/wg/httpbis/trac/ticket/307>: "untangle Cache-Control ABNF"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/353>: "Multiple values in Cache-Control header fields"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/355>: "Case sensitivity of header fields in CC values"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/356>: "Spurious 'MAYs'"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/360>: "enhance considerations for new cache control directives"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/361>: "ABNF requirements for recipients"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/368>: "note introduction of new IANA registries as normative changes"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/373>: "broken prose in description of 'Vary'"

D.2. Since draft-ietf-httpbis-p6-cache-20

Closed issues:

- <http://tools.ietf.org/wg/httpbis/trac/ticket/375>: "'Most Conservative'"

Other changes:

- Conformance criteria and considerations regarding error handling are now defined in Part 1.
- Move definition of "Vary" header field into Part 2.
- Add security considerations with respect to cache poisoning and the "Set-Cookie" header field.

D.3. Since draft-ietf-httpbis-p6-cache-21

Closed issues:

- <http://tools.ietf.org/wg/httpbis/trac/ticket/223>: "Allowing heuristic caching for new status codes"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/406>: "304 without validator"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/418>: "No-Transform"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/430>: "Revert prior change to the meaning of the public cache response directive.

D.4. Since draft-ietf-httpbis-p6-cache-22

Closed issues:

- <http://tools.ietf.org/wg/httpbis/trac/ticket/436>: "explain list expansion in ABNF appendices"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/453>: "Returning the freshest response"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/464>: "placement of extension point considerations"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/469>: "Editorial notes for p6"
- <http://tools.ietf.org/wg/httpbis/trac/ticket/471>: "Vary and future requests"

Index

Authors' Addresses

Roy T. Fielding (editor)
Adobe Systems Incorporated
345 Park Ave
San Jose, CA 95110
USA

EMail: fielding@gbiv.com
URI: http://roy.gbiv.com/

Mark Nottingham (editor)
Akamai

EMail: mnot@mnot.net
URI: http://www.mnot.net/

Julian F. Reschke (editor)
greenbytes GmbH
Hafenweg 16
Muenster, NW 48155
Germany

EMail: julian.reschke@greenbytes.de
URI: http://greenbytes.de/tech/webdav/