Non-compliant and Proud: A Case Study of HTTP Compliance

Paul Adamczyk, Munawar Hafiz and Ralph E. Johnson
Department of Computer Science
University of Illinois at Urbana-Champaign
{padamczy, mhafiz, rjohnson}@uiuc.edu

Abstract

We studied the most popular websites in the US and around the world and discovered that few of them implement the HTTP standard completely. However, the servers are capable of implementing HTTP correctly; it is the configurations that are non-compliant. It is not hard to configure servers correctly, so these websites are non-compliant out of choice, not necessity.

1. Introduction

HTTP is the cornerstone of the World Wide Web. Understanding HTTP is important, because new Web technologies are built on top of it. Such technologies use HTTP as the basis and extend it in many ways to create more complex and powerful protocols. Without understanding HTTP, one runs the risk of reinventing capabilities that are already present in HTTP or completely misusing the protocol. Alternatively, understanding HTTP just by reading the standard can lead to problems, because not all of its features are widely used. This paper examines current status of HTTP compliance.

HTTP defines eight methods for interacting with resources on the Web. Most Web users are not aware of them at all. Typical Web programmers know only of two HTTP methods: GET and POST. These methods have visual representations in Web browsers and they are the only two methods supported by HTML forms. GET is invoked when a user elects to go to a URI in the address bar, when a hyperlink is selected, or in response to pressing a button in a form. POST is invoked by pressing a button. Other HTTP methods do not have any representation in the browser.

This paper presents the study of the implementation and configuration of HTTP methods on the Web. We distinguish the *implementation* of these methods in Web servers and their *configuration* by websites that use these servers. We analyze the compliance results from both perspectives.

Our results are based on the study of the 100 most popular websites in the world, the 100 most popular websites in the USA (both according to Alexa [1]), and the websites of the top 25 computer science departments in the USA (according to US News [20]). While there is some overlap between the first two sets resulting in 176 unique websites, we selected these three groups, because they represent different flavors of popularity. The most popular websites in the world provide social networking, search, and Web portals in the most-used languages in the world. The US websites additionally include many e-commerce sites. The websites of CS departments represent the popular research institutions.

This paper makes the following contributions to the understanding of the Web:

- It presents a quantitative analysis of the state of HTTP compliance, 2007 A.D. This study focuses on a selection of HTTP methods, headers, and algorithms defined in the HTTP standards.
- It places the results in historical context by discussing trends that have occurred over time, both in the definitions and the implementations of the standard. Thus it provides the basis for understanding the capabilities of HTTP for the designers of new protocols built on it.
- It invites further studies and discussion about the state of the Web by posing questions prompted by the results. More detailed studies, targeting specific aspects of Web protocols can use these findings as a start.

This paper begins with an overview of the related work followed by the overview of our experiments. We present experimental results, compare them to previous studies, and discuss them from the perspective of Web servers and proxies. The results show that the Web servers implement HTTP according to the standard, but few websites are configured according to the standard.

2. The HTTP Standard

Initial implementations of HTTP had only one method, GET [3, 4]. Over time, many methods have been added and removed. Table 1 summarizes the historical development. After an initial explosion, the number of methods has de-

Table 1. HTTP methods defined in subsequent incarnations of the protocol

Protocol Version	Release Date	Supported Methods	Change from previous
Initial implementation	Before 1991	GET	—
Before HTTP/1.0	1991	GET, PUT, POST, HEAD, DELETE, LINK, UNLINK, *CHECKIN**, *CHECKOUT, SEARCH, TEXTSEARCH, SHOWMETHOD, SPACEJUMP*	Added 12 methods.
RFC 1945 (HTTP/1.0)	May, 1996	GET, HEAD, POST, PUT, DELETE, *LINK, UNLINK*	Removed six methods
RFC 2068 (HTTP/1.1) (obsoletes RFC 1945)	Nov, 1998	GET, HEAD, POST, PUT, DELETE, OPTIONS, TRACE	Removed LINK, UNLINK; Added OPTIONS, TRACE
RFC 2616 (HTTP/1.1) (obsoletes RFC 2068)	June, 1999	GET, HEAD, POST, PUT, DELETE, OPTIONS, TRACE, CONNECT	Added CONNECT

creased. The methods have been fairly stable since the standard has been published in RFC (Request For Comments) 1945 [5]. The current definition, RFC 2616 [9], consists of eight methods. For the purpose of this paper, we divide these methods into three groups:

(1) Read-only methods (GET and HEAD) that do not modify the state of the server. These are the only methods that all resources are required to implement.

(2) Write methods (POST, PUT, and DELETE) that modify the data on the server.

(3) Supporting methods (OPTIONS, TRACE, and CONNECT) that were added in version 1.1. These are infrastructure methods that provide access to resource options, debugging support, and secure connections.

RFC 2616 does not define the semantics of the CONNECT method. Unlike the other methods, CONNECT is implemented by Web proxies. The client sends this method to the proxy server to initiate the creation of an end-to-end Secure Socket Layer (SSL) tunnel between a client and an end server. RFC 2817 [13] complements RFC 2616 by describing how to use CONNECT to create an end-to-end SSL tunnel through a proxy from a requesting client to a server.

HTTP requests have the following structure:

```
Method SP Request-URI SP HTTP-Version CRLF
Headers
CRLF
[ message-body ]
```

`Method` is the HTTP method (*e.g.,* GET). `SP` stands for space. `Request-URI` is the relative part of the website address. It is set to "/" if the address is the home page. `HTTP-Version` is HTTP/1.0 or 1.1. `CRLF` is carriage return and line feed. `Headers` is a list of headers, which may be general, request, or entity headers, *e.g.,* Content-Length, Date. They are separated by `CRLF`. `message-body` (also called *entity*) is optional and contains the payload (*i.e.,* data).

The HTTP response consists of a status line, a set of headers, and an optional message-body:

```
HTTP-Version SP Status-Code SP Reason-Phrase CRLF
Headers
CRLF
[ message-body ]
```

Some elements are identical to the request. `Status-Code` is a numerical value of the result and `Reason-Phrase` is a corresponding textual description. For example, in the popular response "404 Not Found", the status code is 400 and reason phrase is "Not Found". Following common terminology, we refer to status codes in the 400 and 500 ranges as *error codes*. `Headers` include general, response, and entity headers.

3. Related Work

Several experimental studies of HTTP have been conducted in the past. These studies have provided motivation for adding new features to the standard, *e.g.,* caching [6], persistent connections [17], or pipelining [16]. The frequency of such studies decreased after HTTP/1.1 became official. Instead, more focus has been given to studying the characteristics of specific network elements, such as Web servers [2] and Web proxies [7].

We are aware of only one experimental work that is similar to ours. PRO-COW [14] studied the compliance of Web servers with HTTP/1.1 in 1999, soon after the standard was published. This study focused on compliant implementation of the mandatory and optional HTTP features, including methods and headers. It analyzed 15 most popular websites of 1999. Our study included 10 of these websites. A follow-up PRO-COW study from 2001 [15] included 500 websites, selected from 3 different services reporting the most popular sites. While the websites of these services still exist today, they no longer list popular websites, so we could not use them. We discuss PRO-COW results to verify our findings in Section 8. We build on their results to provide a more complete view of HTTP compliance of websites, Web servers, and other Web systems.

4. Experimental Setup

We implemented an HTTP client to build, send, and receive the HTTP methods using VisualWorks 7.1

*Methods not included in the next protocol version are shown in italics.

(Smalltalk). The experiments consisted of sending OPTIONS, TRACE, GET, conditional GET, and HEAD methods to the home pages of each of the tested websites. The methods that modify the state of the server (POST, PUT, DELETE) were not tested, because it is not possible to set up the same test for all the websites. See Section 7.1 for more on this. To test CONNECT, we used Fiddler [8], an HTTP debugging proxy that logs all the HTTP method traffic between the sender's computer and the Internet. All the requests go through this proxy and it logs the requests and responses. We browsed the Web pages of the tested websites and looked for the presence of CONNECT messages.

Before presenting the results, it is important to address potential criticisms of our approach. If our requests looked like intruders rather than typical traffic, Web servers may have responded with errors or failed to respond on purpose, *e.g.,* as a protection from a Denial-of-Service attack. We used Web-Sniffer [21], a website that displays headers of various HTTP methods, and Fiddler to verify our findings and got the same results. We considered sending the same requests from multiple locations, but decided against it, following the findings of PRO-COW, which concluded that the location of the requestor has little effect on the responses.

Like most popular sites, the sites targeted by our study need many servers to handle all requests. They use load balancing and caching proxies, which means that subsequent requests may go to different machines or may not even reach the server. It would seem that our test results would be difficult to reproduce, but this is not the case. We analyzed all the results manually and collected them multiple times to verify that they are consistent. All websites generated the same response for each request.

One shortcoming of our approach is that we have tested only publicly accessible websites. It is possible that government or corporate intranets are configured to follow the standard more closely. Unfortunately, we were unable to obtain access to private intranets.

4.1. The Tested Websites

Our tests included 176 unique websites. We started with the 100 most popular websites in the world, the 100 most popular US websites, and the 25 top computer science departments (we refer to them as World sites, US sites, and CS sites in the remainder of the paper). But, due to duplication, we obtained fewer unique results. The top 100 US sites contains 32 of the top world sites. We included the overlapping sites only in the world count. Both world sites and US sites include multiple copies of popular websites, especially Google. The top 100 world sites include 16 versions of Google that correspond to different, country-specific domains. All Google sites are set up exactly the same way – they return the same headers in the same order in all tests – thus we count them only once. The US sites also include 3 copies of Google site, which are not counted. This means that the results for world sites tests contain 85 results, US sites tests have 66 results, and CS sites tests have 25 results.

Among the most popular sites, Yahoo and eBay also have multiple copies corresponding to different countries. However, each of these sites is configured differently (*i.e.,* each site includes different headers in response to the same HTTP method), so each copy is counted individually.

We selected our test websites based on popularity. This might seem counter-intuitive. What is popular need not be the best. However, on the Web, the popularity translates to more stringent availability requirements. Having more traffic means that most popular websites are likely to face more security threats, which makes them more likely to be concerned with the proper use of the HTTP standard. To see how these forces relate to compliance results, we also tested the websites of the highest-ranked computer science departments in the US. These websites represent a different type of popularity and have different requirements. They see lower volume of traffic, fewer market pressures (*i.e.,* no need to make money) and have fewer security concerns. However, the admins of these websites are likely to be well versed in Web server setup.

4.2. The Classification of the Results

The results are classified according to the definitions from the HTTP standard [9]. A method is *unconditionally compliant* if it implements all "must" and "should" requirements. If it implements all "must" requirements, it is *conditionally compliant*. Otherwise it is *non-compliant*. Also, we use the term *correctly* to mean "according to the standard".

Results can be classified in more detail based on the response of a method. *Implemented OK* means that the site is unconditionally compliant and returns all the expected data. *Not implemented* means that the site returns a standard error code (405 Method Not Allowed along with the `Allow` header, or 501 Not Implemented, or 404 Not Found, or 403 Forbidden which "should" include an explanation). This category is also unconditionally compliant. *Error Code in Response* means that the method is not supported/configured, but the response error code was not one of the four listed above. *Runtime Error* means that there was no reply. These two types of responses are non-compliant.

Some websites responded to our tests by requesting a redirection to HTTPS. Since we were not testing HTTPS, we could not classify these responses. Consequently, the counts for some methods do not add up to 176.

This paper presents a summary of results. More detailed results are available at `http://st.cs.uiuc.edu/~padamczy/http_tests.html`.

5. Supporting Methods

The supporting methods are analyzed first:

- OPTIONS is used to request the list of all the methods that a Web server supports. These methods are listed in the Allow header of the response.
- TRACE is used to monitor HTTP messages as they travel through the Web. The destination server sends the request message back to the sender and all the intermediate network elements (proxies, gateways) include their names in the trace by adding the Via header.
- CONNECT is used to change the connection with a Web proxy to an SSL tunnel, resulting in a secure connection.

We discuss OPTIONS first, because it returns the list of methods allowed by a server. Knowing which methods are allowed by a given website helps to determine whether a response with an error code is compliant or not.

5.1. Configuration Compliance Results

OPTIONS method. A standard implementation of OPTIONS (applied to resource "/") "should" return the list of methods supported by the server in the Allow header. Figure 1 summarizes the compliance results of this method.

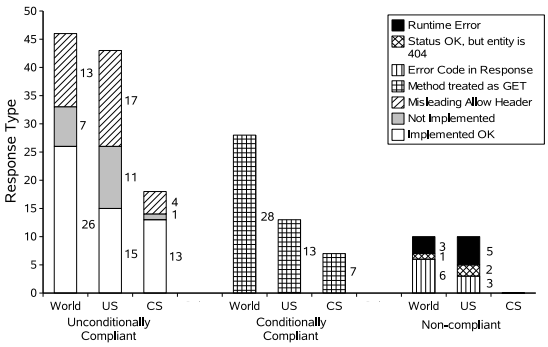

Figure 1. Compliance results for OPTIONS

54 websites implement OPTIONS correctly. The unconditionally compliant category includes 34 sites whose response contains an Allow header listing methods that are *not* allowed by the website. Some of these responses indicate that TRACE or HEAD are allowed, but invoking these methods results in errors "501 Not Implemented", "405 Method Not Allowed" or "403 Forbidden". The wording in the standard makes this case compliant, but we believe that this is an error in the standard. If the contents of the Allow header are inconsistent with other methods, OPTIONS results are misleading. 48 websites treat OPTIONS as if it were a GET, which is also conditionally compliant*.

An interesting example of non-compliance is sending responses with the status code "200 OK" but including an empty entity named `404.html`. 3 websites return this result. We suspect that this response is meant to circumvent the default behavior of some Web browsers that do not display entities when the response contains an error code. PRO-COW [15] study observed similar behavior.

OPTIONS/* method. The HTTP standard states that OPTIONS applies only to the methods that are defined for the specified Request-URI. To receive the list of methods supported by the server as a whole, the Request-URI must be set to "*". But we found that OPTIONS is interpreted in two different ways. OPTIONS sent to the Request-URI corresponding to the home page can return: (1) all possible methods defined by the server, or (2) all methods defined for the requested page. This inconsistency means that one test is not sufficient to gauge the compliance of OPTIONS.

To perform an accurate analysis of the implementation of OPTIONS, we ran a second set of tests with the Request-URI set to "*". We refer to this version of the method as OPTIONS/* in the remainder of this paper. The results for OPTIONS/*, are shown in Figure 2.

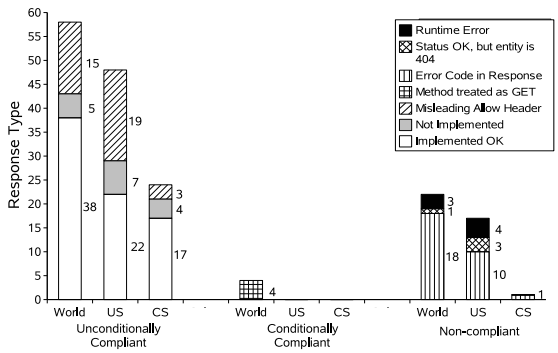

Figure 2. Compliance results for OPTIONS/*

The OPTIONS/* method has more compliant responses than OPTIONS. 130 websites return unconditionally compliant responses. Only 4 responses from the World sites treat OPTIONS/* like GET, but the number of misleading Allow headers increases slightly. Since OPTIONS/* should list all the methods valid for all URIs, its result should be a superset of all the methods listed in the corresponding OPTIONS. But some websites report fewer methods in the Allow header of OPTIONS/* than in response to OPTIONS. The number of error codes in responses increases as well; 4 websites return an empty entity named `404.html`. 69 websites return the same response for both OPTIONS tests.

Allowed HTTP methods. The Allow header (included in OPTIONS or OPTIONS/*) would seem like a good indicator of which methods are configured. Unfortunately,

*RFC 2616, section 9.2 states: *200 response SHOULD include any header fields that indicate optional features implemented by the server and applicable to that resource (e.g., Allow).* The Allow header is not required in the response, so a response without it is conditionally compliant.

not all websites provide this information. Table 2 shows the counts of HTTP methods reported. Depending on the server type, the allowed methods are listed in Allow and/or Public header. For websites where OPTIONS and OPTIONS/* return completely different results, the table reflects the contents of the response where the response lists more methods.

There is a discrepancy between most popular sites and CS department sites. Only about 2/3 of most popular sites (in the US and worldwide) return the Allow header in responses to OPTIONS or OPTIONS/*. In contrast, all but 2 CS department sites provide this information.

All the responding websites list GET and HEAD as supported methods. OPTIONS, TRACE, and POST are listed in most responses. Other methods are listed sporadically, primarily by the US sites. Section 7 has more on this and the relevance of WebDAV. Some of the sites list other methods, which were never in the standard (*e.g.*, INDEX, RMDIR).

TRACE method. A standard implementation of the TRACE method "should" return the original message in the response with the Content-Type header set to message/http. It "should" also include a Via header listing the gateways and proxies that processed this method, but this requirement does not affect the compliance results of the server. (Section 9.2 discusses our findings related to proxies, collected from Via and other non-standard headers). The results for TRACE are summarized in Figure 3.

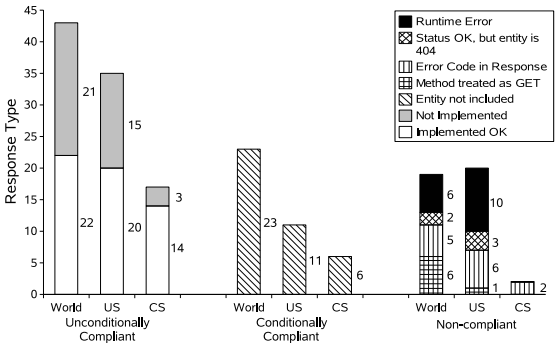

Figure 3. Compliance results for TRACE

Overall, 95 responses to TRACE are unconditionally compliant. About half of the unconditionally compliant World sites do not implement this method and indicate so with the expected error codes. The conditionally compliant responses fail to include the original message in the response entity (23, 11, and 6 websites respectively). Non-compliant responses return wrong error codes to report that the method is not supported (13 websites in total), treat TRACE like GET (7 websites), use the `404.html` entity as the error code (5 websites), or fail to respond (16 sites).

CONNECT method. The end servers implement Secure HTTP (HTTPS), which uses SSL to transport the HTTP methods. The entire communication between the SSL connection endpoints is encrypted. According to RFC 2817 [13], a client can create an SSL connection in two ways. Direct SSL connection can be created if the client communicates directly with the server. Alternatively, a client can create an indirect connection by communicating with a proxy and requesting the proxy to create an SSL tunnel ending at the origin server. This tunnel setup request is initiated by the CONNECT method.

A client that is connected directly to the server can issue an optional or mandatory upgrade request (asking the server to switch to the SSL protocol) using GET or OPTIONS. However, direct connections are not supported by any of the websites we tested. The clients communicate with a proxy. They send a CONNECT request to the proxy to set up a tunnel. The proxy communicates with the end server, creates the tunnel, and, once the tunnel is set, silently forwards the packets from the client to the end server.

We successfully sent the CONNECT method to a proxy and created an SSL tunnel with 32 out of 85 World sites. We did not see CONNECT in 53 other cases, because these end servers do not implement HTTPS. For the 66 US sites, we found that proxy servers are configured to use CONNECT in 43 cases; the remaining 23 do not implement HTTPS. All the CS sites implement HTTPS. The fact that CONNECT is not supported in many World and US websites does not mean that it is unnecessary. It may be that the origin servers only contain public material (*e.g.*, CNN, Washington Post, IMDB), or they contain publicly editable material (*e.g.*, Wikipedia), or they use some other encryption scheme to achieve confidentiality (*e.g.*, Livejournal clients send their data using POST, but the authentication data is encrypted; Yahoo China sends username in plaintext, but password in encrypted format).

The Allow header of OPTIONS indicates that only 4 websites claim to implement CONNECT (see Table 2): Earthlink, Statcounter, Digitalpoint, and the CS Department of the University of Pennsylvania. Surprisingly, Statcounter and Digitalpoint do not implement CONNECT.

5.2. Discussion

Neither OPTIONS nor TRACE is configured sufficiently well to perform the functionality defined in the HTTP standard. In practice, only CONNECT, defined outside of RFC 2616, is configured according to its specification.

Ideally, the clients could use OPTIONS to dynamically discover what methods are supported by a resource and then select the best method to invoke. However, only few websites configure this method to return unambiguous, compliant results indicating that dynamic discovery is not a good idea. The OPTIONS method appears to be used mainly to "confuse the enemy" (*i.e.*, hackers) by providing misleading results in the Allow header. It prevent hackers from learning

Table 2. Allowed HTTP methods count, as reported in OPTIONS

	GET	HEAD	POST	PUT	DELETE	OPTIONS	TRACE	CONNECT	Other	WebDAV
World	65	65	42	4	3	65	61	0	3	1
US	42	42	34	18	17	41	39	3	17	7
CS	23	23	17	2	2	23	23	1	2	1
Total	130	130	93	24	22	129	123	4	22	9

the capabilities of the system. Browsers never send the OPTIONS method to the server. Hence it is natural to obscure the information to achieve security through obscurity.

The TRACE method appears to be more useful, because message tracing and debugging is very difficult in distributed systems, such as the Web. TRACE is reminiscent of the `traceroute` command. Yet, our results show that few websites handle this method correctly. We suspect that the reason for low compliance of these methods is security.

Many websites do not use CONNECT, because security is not an important requirement for them. CONNECT is used by 65% of the US sites, but by only 37% of the World sites. This is because the top ranking websites in the world represent different regions; and typically the social networking and Web portal sites are highly ranked in all regions. These sites have low security requirements and do not use CONNECT. In contrast, many e-commerce sites are featured in the top US sites (along with the more popular social networking sites). Security is a key requirement for these websites, so their providers configure the end servers and the proxies correctly to execute the protocol.

6. Read-Only Methods

The read-only methods retrieve the data from the server:

- GET is used to retrieve the representation of a resource.
- HEAD is used to retrieve the headers (*i.e.*, metadata) of the corresponding GET method for a given Request-URI, but without the entity.

6.1. Configuration Compliance Results

GET method. A standard implementation of the GET method "must" contain an entity and two headers, Content-Length and Date. Our results indicate that all websites are fully compliant with this definition regardless of the Web server or the protocol version they are using.

Since all Web servers implement the GET method correctly, we also tested one of the flavors of "conditional" GET to see how the caching is supported. A conditional GET is a GET method that includes a header whose name begins with "If", *e.g.*, If-None-Match. The conditional GET returns an entity only if the specified condition is true.

Conditional GET. A standard implementation of conditional GET "should" return an entity only if the resource on the server has been modified since the time specified by the condition. But all the websites generate the content of their homepage dynamically – each subsequent response has a different Date header and different Content-Length.

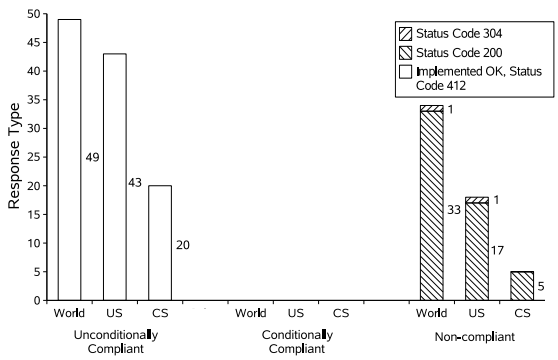

Figure 4. Compliance results for GET with If-Unmodified-Since header

For this test we needed an entity that changes less often – instead of the homepage, we were requesting the *favicon*, which is a small icon logo of a website; Firefox displays it to the left of the URL in the address bar. We tested the websites that do not define a favicon by requesting a small image from their homepage. We excluded from our test 7 websites that store all their graphics at a different domain.

To test the conditional GET, we selected the header If-Unmodified-Since. PRO-COW [14] also tested this condition. This method "must" return status code "412 Precondition Failed" if the requested resource has not been modified since the given date. We used a date 5 years ago to ensure that all favicons were modified since. The results are summarized in Figure 4.

The majority of websites in each category (49, 43, and 20 respectively) respond to this method correctly. There are two types of non-compliant responses. 55 websites (33, 17, and 5 for each category) include an entity that has been modified after the date we requested. Two websites respond with status code "304 Use Local Copy," which the standard defines as the expected response to a conditional GET with a different header, If-Modified-Since. There are no conditionally compliant responses.

HEAD method. A standard implementation of HEAD "must not" contain an entity and it "should" contain the same headers as the corresponding GET method. The results are summarized in Figure 5.

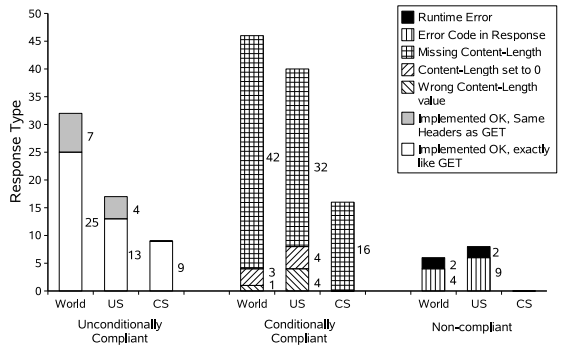

Figure 5. Compliance results for HEAD

All 3 website categories return more conditionally compliant responses (102 in total) than unconditionally compliant ones (58 in total) for HEAD. Most unconditionally compliant responses include exactly the same headers for GET and HEAD (25, 12, and 9 respectively). Others return the same headers, but with slightly different values (7 World sites and 4 US sites). For example, CNN and AOL return a different value of Connection ("close" for HEAD and "keep-alive" for GET), but all other header values match.

All the conditionally compliant responses have problems with one header, Content-Length, which specifies the size of the entity in the corresponding GET method. Recall that all GET responses "must" include Content-Length. Half or more of all the websites in each category (42, 32, and 16 respectively) do not include the Content-Length header[†]. Other responses have this header set to a value unrelated to the value in the corresponding GET method (these 5 results are counted under *Wrong Content Length* in the figure). Yet other responses set it to 0 (3 World and 4 US responses).

Since HEAD is a required method, every website "must" configure it. Otherwise, it is non-compliant and counted under *Error Code in Response*.

6.2. Discussion

It is not surprising that all the tested websites configure the GET method correctly, because GET is the essence of the Web. Even more advanced versions of GET, such as conditional GET, are configured correctly by the majority of websites. This is because the conditional GET allows Web clients to take advantage of caching.

In contrast, the HEAD method is seldom configured correctly. This result is surprising for three reasons. First, HEAD is a required method. Every website is required to support GET and HEAD. Second, HEAD seems to be easy to implement by reusing the code for GET. In fact, Apache uses the same code to handle both methods. The only code where the handling differs is the check whether to include the entity in the response. But even Apache Web servers sometimes (due to bugs) fail to generate unconditionally compliant responses to HEAD. One amusing problem is setting Content-Length to 0, which suggests that the code for GET is reused and the size of the entity is calculated dynamically before it is sent. Since HEAD has no entity, the server sets the value to 0. Third, HEAD performs useful tasks. The metadata retrieved with HEAD is used (1) to determine the size of entity before requesting a large resource, (2) to verify that a resource still exists, and (3) to check when the resource has been last modified. Over time, the first task has become less significant, because improving network speeds have made downloading data easier. The second task will always be important. However the third task can be better accomplished with conditional GET. When the client uses HEAD to check if a resource was modified and the resource was indeed changed, it must then send a GET to retrieve it. With conditional GET, the resource is returned in one message exchange, but only if it was changed.

It would seem that HEAD should be used often, yet the test results show that it is not. Had it been used often, the website admins would have to configure it correctly. We suspect that this method is underutilized, because clients do not know about it.

[†]We shared all our test results with Apache developers. They informed us that this is a known bug in Apache 2.x. They did not provide any insights about the reasons for other non-compliant results.

7. Write Methods

The HTTP standard defines three methods for modifying resources stored on the server:

- POST is used to update the state of an existing resource.
- PUT is used to create a new resource or to replace the entire contents of an existing one.
- DELETE is used to delete a resource.

7.1. Configuration Compliance Results

Unfortunately, the write methods cannot be tested uniformly. It is not possible to set up the same test for all websites, because each website has its own policy for updating resources. To set up a comprehensive test, we would need to find unique modifiable resources on each website. To modify these resources, we would also need to gain access rights (*e.g.*, register with the website). In the end, all tests would be unique rather than uniform.

PRO-COW collected compliance results for the POST method by checking if Web servers report that POST is not allowed using the standard-defined error codes. Such a test is not convincing, because it tests the configuration of the error reporting, rather than the configuration of the method itself. Error reporting in the methods discussed in the previous two sections shows that most websites do not configure

error codes correctly. This is another reason why we did not test POST responses the way PRO-COW did.

Instead, we use our test data to determine the relative popularity of write methods by analyzing the contents of the Allow header. The results collected from this header indicate that POST is configured by 71.5% of websites that return Allow. (See Table 2). Much fewer websites report that they support PUT or DELETE (18.5% and 16.9% respectively). Moreover, some websites do not include POST in the Allow header even if they support it. For instance, Yahoo does not list any write methods in the Allow header, but it must support at least one; otherwise the users could not send messages through Yahoo email.

7.2. Discussion

The POST method is configured by most websites, while PUT and DELETE are not. One reason for this is that the definition of write methods in the standard is too complex.

The difference between PUT and POST is subtle. The standard states that to set up or update a resource, a client should send it to the server via PUT. Alternatively, to allow the server to determine how to update the resource, a client would provide the data via POST. In practice, POST is used almost always, because it makes the server responsible for updating the resource. If the server determines that the data submitted via POST should replace the current data, it can do so anyway. A client wishing to update an existing resource using PUT, needs to first obtain the current data stored in the resource (via GET), make the updates, and then send the updated representation to the server via PUT.

Similarly, POST is used to delete resources. This is typically done by sending POST to a Request-URI that includes the word "delete." This is not a good interpretation of the standard, because the URI is supposed to identify a resource. Including "delete" in the URI means that the URI is used to identify the delete method on the server.

The easiest way to avoid this complexity when configuring a website is to use one method for all the cases. In practice, all writes (including creation and deletion of resources) are done with POST.

It might seem beneficial to remove PUT and DELETE from the HTTP standard, because they are rarely used and their tasks can be accomplished with POST. However there are HTTP extensions that use these methods. WebDAV, a standard for distributed authoring [10] is one example. It uses existing HTTP methods (including PUT and DELETE) and some new ones to define tasks for Web authoring: management of resource versions, access to collections of resources, and access control. WebDAV is relatively popular; 7 of the top 100 US sites support it. So, it is important to preserve PUT and DELETE as part of HTTP, because they are relevant in other contexts.

8. Changes in Compliance

Our results show that the majority of the websites configure only the GET method correctly. All other methods are often misconfigured. To put these results in perspective, they should be compared with prior experiments. PRO-COW project was a HTTP/1.1 compliance study that tested 13 different features (algorithms, methods, and headers) defined in HTTP/1.1. Our experiments tested the same methods as PRO-COW. Other PRO-COW tests, such as the presence of mandatory headers in methods, or persistence and pipelining are outside of the scope of our study. They tested low-level details of the protocol, while we are focusing on methods that the clients can call. Table 3 compares the results of the 6 tests common to both studies. Each method shows the percentage of unconditionally compliant, conditionally compliant, and non-compliant configurations. They are derived from values described in the previous sections.

There are two PRO-COW papers. The first one presents results for 15 sites in 1999 [14]. The second one tested 500 websites and produced similar results in 2001 [15]. Except for OPTIONS/*, we use the results from the first paper, because they show a single value for each test.

The results, summarized in Table 3, show two trends: (1) increased compliance of the GET method, and (2) unchanged or decreased compliance of other methods.

The results of GET, very good in PRO-COW, are perfect in our tests. The results of the conditional GET show improvement in unconditional compliance, but still show a lot of non-compliance. The increased compliance of conditional GET indicates that the configuration of features like caching, which are useful to clients, will improve over time.

In contrast, TRACE and HEAD show large decrease in unconditional compliance, and TRACE shows significant increase in non-compliance. OPTIONS shows an increase of unconditional compliance, but a larger increase in non-compliance. This indicates that these methods are not used in practice. Since clients do not use them, website administrators have no incentive to configure them correctly (recall mismatched headers of GET and HEAD, or responses missing the original request in TRACE). Ambiguity in the standard definition (as in OPTIONS) is also a cause of bad configurations. Lastly, the results of OPTIONS, and especially TRACE, illustrate a change of culture prompted by the security concerns that occurred after the PRO-COW study.

9. Compliance of Implementation

We sent the test messages to websites, but the data we collected says a lot about the Web servers that generated the responses and Web proxies through which our tests passed. This section discusses how Web servers and Web intermediaries (proxies, caches, etc) implement the HTTP standard.

Table 3. Website compliance results comparison with PRO-COW (All values are in percentages)

	OPTIONS				OPTIONS/*				TRACE			
	PRO-COW	World	US	CS	PRO-COW	World	US	CS	PRO-COW	World	US	CS
Unconditional	59.8	54.8	65.2	72.0	26.8-32.3	69.0	73.8	96.0	97.3	50.6	53.0	68.0
Conditional	39.4	33.3	19.7	28.0	65.0-72.4	4.8	0.0	4.0	2.5	27.1	16.7	24.0
Non-compliant	0.8	11.9	15.1	0.0	0.8-2.7	26.2	26.2	0.0	0.2	22.3	30.3	8.0
	HEAD				**GET**				**Conditional GET**			
	PRO-COW	World	US	CS	PRO-COW	World	US	CS	PRO-COW	World	US	CS
Unconditional	72.9	38.1	26.2	36.0	83.5	100.0	100.0	100.0	41.7	59.0	70.5	80.0
Conditional	9.4	54.8	61.5	64.0	16.1	0.0	0.0	0.0	1.2	0.0	0.0	0.0
Non-compliant	17.7	7.1	12.3	0.0	0.4	0.0	0.0	0.0	57.1	41.0	29.5	20.0

9.1. Web Servers

To determine the compliance of Web servers, we extracted the names of Web servers from the responses. Typically the Server header identifies the type and version of the Web server. Only 10 websites that we tested do not include this header or send meaningless values (*e.g.*, "server"). Among them are Yahoo and Amazon. Table 4 shows the count of Web servers used by the tested websites.

Table 4. Web servers used by tested websites

Vendor	Server Type	World	US	CS
Apache	Apache/1.3	15	10	7
	Apache/2.x	14	8	12
	Apache (Unspecified)	24	19	4
	Total Apache	**53**	**37**	**23**
Microsoft IIS	IIS/5.0	5	6	1
	IIS/6.0	9	8	1
	Total IIS	**14**	**14**	**2**
Other	Netscape	1	4	0
	Sun-ONE	1	3	0
	Other (AOL, lighttpd)	9	5	0
	Total Other	**11**	**12**	**0**
	Not Specified	7	3	0
	Total	**85**	**66**	**25**

Apache is the most popular Web server in our tests. Most non-US sites use Apache, as do almost all top CS departments. Microsoft's IIS is more popular in the US, but less known Web servers are also more often used in the US sites.

We mapped the correctly configured websites to specific Web server versions to find which servers implement the standard correctly. Most Web servers versions listed in Table 4 have a corresponding, correctly-configured website. Table 5 lists some examples.

Table 5. Example websites that pass all tests

Server type	Website	Description
Apache/1.3	`www.apple.com`	Apple Inc.
Apache/2.0	`www.cs.ucsd.edu`	CS Dept., University of California, San Diego
Apache/2.2	`www.cs.utexas.edu`	CS Dept., University of Texas
IIS/5.0	`www.realtor.com`	Realtor real estate
SunONE/6.1	`www.nytimes.com`	NY Times newspaper

Moreover, some websites using IIS/6.0 and Netscape/6.0 come very close to passing all our tests. MySpace, using IIS/6.0, is unconditionally compliant with all methods except OPTIONS. The Allow header of OPTIONS lists the TRACE method, but when it is called, it returns error code "501 Not Implemented". Similarly, Comcast (using Netscape/6.0) responds to all methods, but it returns wrong status codes – 413 instead of 501 for the unsupported TRACE and 304 instead of 412 for GET If-Unmodified-Since. Had we tested more websites that use these two servers, we would likely find perfect configurations.

These results show that the type of Web server used does not influence the compliance of a website. Since we are able to find at least one example of a fully-compliant website for almost every server type, these server types *are* fully-compliant. Our results show that well-implemented servers are configured to be non-compliant on purpose.

The PRO-COW study does not consider the relationship between Web server implementation and configuration. It counts the number of compliant results per Web server type, implying that website configuration and Web server implementations are closely related. This is not correct. There are many reasons why websites that use a certain version of a Web server are misconfigured more often. This may be the most popular server, so everyone wants to use it, and it is easier to find non-compliant configurations. Or the server may have many security flaws and people misconfigure it on purpose. Or the server might be very hard to configure.

It is not possible to infer server compliance from percentages of website results. However, it is possible to determine which Web servers tested by PRO-COW were implemented correctly. They show the percentage of servers that pass all tests sorted by server types. Some Apache/1.2, Apache/1.3 and IIS/4.0 servers passed all tests, while none of Netscape/3.5 and 3.6 did. This is consistent with our results: most Web servers implement the standard correctly.

Web server security. It might seem that a website identifying the type and version of its server creates a security vulnerability, but this is not the case. Although there are few major server vendors, they offer multiple versions of the software and each version has many configurations.

9.2. Web Intermediaries

Our tests of OPTIONS, OPTIONS/* and TRACE produced responses with spoofed Server headers. Spoofing is typically done by Web intermediaries that override the Server header with their own name. SquidCache and AkamaiGHost are the two most popular Web intermediaries in our tests, each of them overriding the Server header of 9 different websites. SquidCache is a caching proxy, while AkamaiGHost is a content delivery system. Most spoofed headers are in responses with error codes. For example, if a method is not supported, the intermediary, not the server, sends an error response. We know this because the Server header is changed to the name of the intermediary and the HTTP version is set to 1.0. That is not to say that all websites are HTTP/1.1-compliant. Some of today's most popular websites (Wikipedia, AOL, Mapquest) use HTTP/1.0. 13 websites return all responses as version 1.0. Other non-standard ways of including the information about the intermediary in the response are adding new headers (*e.g.*, S, X-Server, X-cache) or using existing headers (*e.g.*, From, which is supposed to return an e-mail address).

This behavior is non-compliant. The HTTP standard states that gateways and proxies "should" identify themselves in the Via header and "must not" modify the Server header. HTTP errata [12] further clarifies that the Via header "must" be used in all methods, but this rule is rarely followed. We collected only 13 responses, from 5 different websites, that include Via. Possibly the Via header is not used, because clients are not interested in this information. Web browsers hide this information from the users. Proxies may also be configured not to report their identity in Via to overcome the problem of finding open proxies on the Web.

Although some of the HTTP methods were defined only in HTTP/1.1, all our results include the data obtained from both HTTP/1.0 and HTTP/1.1 responses. This is because many of the 1.0 responses are 1.1-compliant. The HTTP version header is not end-to-end, which means that Web intermediaries that handle the response may change its value, even without modifying the contents of the response. Most intermediaries we observed implement only HTTP/1.0. But HTTP/1.0 is still in use by websites as well. Some of today's most popular websites (Wikipedia, AOL, Mapquest) use it. 13 websites return HTTP/1.0 in all responses.

The above results seem to imply that Web proxies still do not implement the HTTP/1.1 standard correctly. But the results of the CONNECT method show that proxies implement secure connections (defined after HTTP/1.1) quite well. Web proxies implementing CONNECT support SSL 2.0, SSL 3.0, and TLS 1.0 (*a.k.a.* SSL 3.1). We used Internet Explorer 6.0 on Windows XP for the tests. Internet Explorer sends a TLS 1.0 compatible ClientHello handshake request to initiate the protocol. The server returns an SSL 3.0 compatible response through the proxy and the tunnel is set up according to SSL 3.0. This is possible, because TLS 1.0 can be downgraded to SSL 3.0. We obtained the same results with Firefox 1.5.

Proxy non-compliance with HTTP is not caused by a lag in implementation, but rather it indicates that some HTTP features are not important from the proxies' perspective.

Web proxy security. Security vulnerabilities of TRACE have an indirect effect on proxies. Websites that enable TRACE can become targets of 'cross site tracing' attack that could reveal user information [19]. A script on the client machine can forward a response to a TRACE method with Cookie headers to a malicious server [11]. To prevent this, many servers return the length of the TRACE request instead of the whole request to the client. We suspect that proxies not include the Via header due to such vulnerabilities.

10. Analysis of the Results

By now, the key result of our experiments should be obvious: HTTP methods do not behave according to the standard. The PRO-COW paper [14] concludes: "The results of our experiment show that the situation on the Web must first be improved at the origin server before we can worry about end-to-end improvements." Yet 6 years after the PRO-COW experiments, our tests show that the results have not changed much. But the Web is doing just fine. Why is that?

Perhaps website admins are incompetent and their setups are incorrect. The websites we tested generate the most traffic and probably face the most security threats. For many of these websites (just consider Google), the Web presence is their entire business. To ensure availability, they must be configured well. Therefore this hypothesis is not true.

Perhaps it is difficult to set up Web servers, because they come with bad defaults. While this hypothesis seems plausible, it is not true either. We installed two versions of Apache to see what their default configurations are. In Apache 1.3 for RedHat, all the methods (even PUT and DELETE) are configured correctly out of the box. The same is true for Apache 2.0 for Windows XP. A website is operational in minutes after the Web server software is downloaded. It is harder to disable correct configurations.

Perhaps bad website configuration is done on purpose. As noted in the previous paragraph, the Web servers have standard-compliant default settings, yet the compliance of configurations varies. The non-standard setup of less popular methods could be a way to achieve security through obfuscation. Security vulnerabilities mentioned in the previous section support this hypothesis.

Perhaps it is not important to be compliant with the entire HTTP standard. Even the most popular sites seem to be satisfied that the key features are working. If GET or POST

method were to suddenly stop working, they would need to be fixed immediately. The other methods are not used often enough to demand proper configuration.

Perhaps most Web traffic is handled by HTTP-agnostic systems, such as content delivery systems. Such Web intermediaries do not operate at the HTTP layer, but use different protocols in their communication with the servers. When they produce HTTP responses on behalf of the servers, such responses are seldom HTTP-compliant.

There are other explanations for the low compliance results, but the last three reasons - security concerns, the limited use of most HTTP methods, and HTTP-agnostic systems - shed some light on this problem. There is a disconnect between the theory (HTTP standard) and practice (system compliance) on the Web. In theory, HTTP is a simple protocol for the Web. It was designed to be extended with specialized protocols, such as WebDAV. In practice, only a small subset of HTTP is used. Web systems built on top of the simplified HTTP add capabilities that already exist in the full-fledged HTTP, but not in the commonly-used subset.

11. Conclusion and Future Work

Our study of different types of popular websites shows that most of them are not compliant with HTTP. While Web servers implement the standard very well, few of the websites are configured correctly. This is a continuing trend, yet it has not affected the growth of the Web.

Our results provide experimental evidence for the debates about the future of the Web. More specific experiments are needed to address them fully, but some debates, *e.g.*, REST vs. SOAP [18], lightweight vs. standards-based security for Web services, can benefit from our results.

The relationship between theory and practice of building Web systems is very complex. Studying only one small aspect of it is not likely to produce comprehensive results, but it is an important step toward a better understanding of the Web. We provide three hypotheses explaining the HTTP non-compliance, but obtaining a clearer picture requires more studies, including surveys of website admins and server implementers. We hope to have raised enough interesting questions for others to join this conversation.

References

[1] Alexa. `http://www.alexa.com`.

[2] M. Arlitt and C. Williamson. Understanding Web server configuration issues. *Software Practice and Experience*, 34(2):163–186, Feb. 2004.

[3] T. Berners-Lee. The original HTTP as defined in 1991. `http://www.w3.org/Protocols/HTTP/AsImplemented.html`, 1991. W3C webpage.

[4] T. Berners-Lee. Is there a paper which describes the WWW protocol. `http://lists.w3.org/Archives/Public/www-talk/1992JanFeb/0000.html`, Jan 9 1992. WWW-talk mailing list.

[5] T. Berners-Lee, R. T. Fielding, and H. Frystyk Nielsen. Hypertext Transfer Protocol — HTTP/1.0. Internet informational RFC 1945, May 1996.

[6] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: A live study of the World Wide Web. In *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems (USITS-97)*, Monterey, CA, 1997.

[7] B. M. Duska, D. Marwood, and M. J. Freeley. The measured access characteristics of World-Wide-Web client proxy caches. In *1997 USENIX Symposium on Internet Technologies and Systems (USITS-97)*, Monterey, CA, 1997.

[8] Fiddler. `http://www.fiddlertool.com/fiddler`.

[9] R. T. Fielding, J. Gettys, J. C. Mogul, H. Frystyk Nielsen, L. Masinter, P. J. Leach, and T. Berners-Lee. Hypertext Transfer Protocol — HTTP/1.1. Internet proposed standard RFC 2616, June 1999.

[10] Y. Y. Goland, E. J. Whitehead, A. Faizi, S. Carter, and D. Jensen. HTTP Extensions for Distributed Authoring — WebDAV. Internet proposed standard RFC 2518, Feb. 1999.

[11] J. Grossman. Cross-Site Tracing (XST). `www.cgisecurity.com/whitehat-mirror/WH-WhitePaper_XST_ebook.pdf`, Jan 2003.

[12] HTTP/1.1 specification errata. `http://skrb.org/ietf/http_errata.html`, Oct 2004.

[13] R. Khare and S. D. Lawrence. Upgrading to TLS within HTTP/1.1. Internet proposed standard RFC 2817, May 2000.

[14] B. Krishnamurthy and M. Arlitt. PRO-COW: Protocol compliance on the Web. Technical Report 990803-05-TM, HP Labs, 1999.

[15] B. Krishnamurthy and M. Arlitt. PRO-COW: Protocol compliance on the Web — A longitudinal study. In *Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems (USITS-01)*, San Francisco, CA, 2001.

[16] H. F. Nielsen, J. Gettys, A. Baird-Smith, E. Prud'hommeaux, H. W. Lie, and C. Lilley. Network performance effects of HTTP/1.1, CSS1, and PNG. In *Proceedings of the ACM SIGCOMM '97 conference on Applications, technologies, architectures, and protocols for computer communication*, Cannes, France, 1997.

[17] V. N. Padmanabhan and J. C. Mogul. Improving HTTP latency. *Computer Networks and ISDN Systems*, 28(1–2):25–35, Dec. 1995.

[18] P. Prescod. Roots of the REST/SOAP debate. In *EML 2002: Proceedings of the Extreme Markup Languages 2002 conference*, Montreal, Canada, 2002.

[19] US-CERT Vulnerability Note VU867593. Multiple vendors' Web servers enable HTTP TRACE method by default. `https://www.kb.cert.org/vuls/id/867593`, Jan 2003.

[20] US News and World Report. `http://http://www.usnews.com/usnews/home.htm`.

[21] Web-sniffer v1.0.24. `http://web-sniffer.net/`.