Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 14, 2009

M. Nottingham
E. Hammer-Lahav

February 10, 2009

Host Metadata for the Web
draft-nottingham-site-meta-01

Status of this Memo

Copyright Notice

Abstract

This memo describes a method for locating host-specific metadata for the Web.

Table of Contents

1. Introduction

It is increasingly common for Web-based protocols to require the discovery of policy or metadata before making a request. For example, the Robots Exclusion Protocol specifies a way for automated processes to obtain permission to access resources; likewise, the Platform for Privacy Preferences [W3C.REC-P3P-20020416] tells user-agents how to discover privacy policy beforehand.

While there are several ways to access per-resource metadata (e.g., HTTP headers, WebDAV's PROPFIND [RFC4918]), the overhead associated with them often precludes their use in these scenarios.

When this happens, it is common to designate a "well-known location" for such metadata, so that it can be easily located. However, this approach has the drawback of risking collisions, both with other such designated "well-known locations" and with pre-existing resources.

To address this, this memo proposes a single (and hopefully last) "well-known location", /host-meta, which acts as a directory to the interesting metadata about a particular authority. Future mechanisms that require authority-wide metadata can easily include an entry in the host-meta resource, thereby making their metadata cheaply available (indeed, because it can be cached, the more mechanisms that use it, the more efficient it becomes) without impinging on others' URI space.

Note that the metadata provided by a host-meta resource is explicitly scoped to apply to the entire authority (in the URI [RFC3986] sense) associated with it (using the process described in Section 4); it does not apply to a subset, nor does it apply to other authorities (e.g., using another port, or a different hostname in the same domain). However, individual mechanisms (e.g., a relation type in the Link field) MAY reduce or expand this scope. This should only be done after careful consideration of the consequences upon security, administration, interoperability and network load.

Please discuss this draft on the www-talk@w3.org [1] mailing list.

2. Notational Conventions

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 RFC 2119 [RFC2119].

This documnet uses the Augmented Backus-Naur Form (ABNF) notation of [RFC5234], and explicitly includes the following rules from it: CRLF

(CR LF), OCTET (any 8-bit sequence of data), DIGIT, ALPHA, and WSP (white space).

3. The host-meta File Format

The host-meta file format is an extremely simple textual language that allows an authority to convey metadata about itself and its resources.

Its syntax is similar to that of HTTP header-fields [RFC2616], but has a few differences:

- White space is permissible both before and after the block of fields, and
- fields MUST NOT be folded across multiple lines.

Furthermore, this format's use diverges from HTTP header-fields in a number of ways:

- The fields are transferred as the message body, not as headers, and
- rather than being related to a message, the fields in host-meta pertain to the entire associated authority (see Section 4), and
- the permissible field-names are constrained by the host-meta field registry. This specification defines one such field, Link.

```
host-meta      = *( WSP / CRLF )
                 *( meta-field CRLF )
                 *( WSP / CRLF )
meta-field     = field-name ":" [ field-value ]
field-name     = 1*tchar
field-value    = *( field-content / WSP )
field-content  = <field content>
tchar          = "!" / "#" / "$" / "%" / "&" / "'" / "*"
               / "+" / "-" / "." / "^" / "_" / "`" / "|" / "~"
               / DIGIT / ALPHA
```

For example,

```
Link: </robots.txt>; rel="robots"
Link: </w3c/p3p.xml>; rel="privacy"; type="application/p3p.xml"
Link: <http://example.net/example>; rel="http://example.com/rel"
```

As with HTTP headers, field-names are not case-sensitive, unrecognised field-names SHOULD be silently ignored when parsing this format, and ordering of fields SHOULD NOT be considered significant unless specified otherwise. Additionally, although the syntax does

not explicitly allow empty lines between fields, parsers SHOULD silently discard them (i.e., be permissive in what they accept).

Field content is constrained by the specification indicated by its associated field-name.

3.1. The Link host-meta Field

The "Link" host-meta field uses the syntax of the Link HTTP header-field [I-D.nottingham-http-link-header] to convey links whose context is the entire authority, rather than a single resource. For example,

```
Link: </terms>; rel="license"
```

indicates that the URI "/terms" refers to a license for all resources associated with the authority.

The Link host-meta field differs from the Link header in the following respects:

- Its context is defined as all resources that share its authority, by default (although this MAY be overridden by a representation obtained from the indicated resource), and
- When the link URI is relative, its base URI is the root resource of the authority. For example, in the example above, if the authority is "example.com", the full link URI would be "http://example.com/me".

4. Discovering host-meta Files

The metadata for a given authority can be discovered by dereferencing the path /host-meta on the same authority. For example, for an HTTP URI [RFC2616], the following request would obtain metadata for the authority "www.example.com:80";

```
GET /host-meta HTTP/1.1
Host: www.example.com
```

The semantics of the protocol used for access to the resource apply. Therefore, if the resource indicates the client should try a different request (in HTTP, the 301, 302, 303 or 307 response status code), the client SHOULD attempt to do so; note that this implies that the host-meta file for one authority MAY be retrieved from a different authority. Likewise, if the resource is not available or existent (in HTTP, the 404 or 410 status code), the client SHOULD infer that metadata is not available via this mechanism.

If a representation is successfully obtained, but is not in the format described above, clients SHOULD infer that the authority is using this URI for other purposes, and not process it as a host-meta file.

To aid in this process, authorities using this mechanism SHOULD correctly label host-meta responses with the "application/host-meta" internet media type.

5. Minting New meta-fields

Applications that wish to mint new meta-fields for use in the host-meta format MUST register them in the host-meta field-registry, following the procedures in Section 7.2. Field-names MUST conform to the field-name ABNF Section 3, and field-value syntax MUST be well-defined (e.g., using ABNF, or a reference to the syntax of an existing header field-value). Field-values SHOULD use the ISO-859-1 character encoding. If a field-value applies to a scope other than the entire authority, that scope MUST be well-defined.

6. Security Considerations

The metadata returned by the /host-meta resource is presumed to be under the control of the appropriate authority and representative of all resources contained by it. If this resource is compromised or otherwise under the control of another party, it may represent a risk to the security of the server and data served by it, depending on what mechanisms use /host-meta.

Scoping metadata to a single authority is the default in host-meta. Thus "http://example.com/", "https://example.com" and "http://www.example.com/" all have different host-meta files with seperate and non-overlapping scopes of applicability. Applications that change the scope of metadata can incur security risks without careful consideration.

7. IANA Considerations

7.1. application/host-meta Media Type Registration

The host-meta format can be identified with the following media type:

MIME media type name: application
MIME subtype name: host-meta
Mandatory parameters: None.
Optional parameters: None.
Encoding considerations: field-values may specify any encoding for their contents, although it is expected that most will use ISO-8859-1 or a subset thereof (for both historic and interoperability purposes).
Security considerations: As defined in this specification. [[update upon publication]]
Interoperability considerations: There are no known interoperability issues.
Published specification: This specification. [[update upon publication]]
Applications which use this media type: No known applications currently use this media type.

Additional information:

Magic number(s):
File extension: None.
Fragment identifiers: None.
Base URI: None.
Macintosh File Type code: TEXT
Person and email address to contact for further information: Mark Nottingham <mnot@mnot.net>
Intended usage: COMMON
Author/Change controller: This specification's author(s). [[update upon publication]]

7.2. The host-meta Field Registry

This document establishes the host-meta field registry as the namespace of field-names for use in meta-fields. Although some meta-fields may be similar to message headers, both syntactically and semantically, the host-meta field registry is separate from the message header field registry [RFC3864] See Section 5 for details and requirements for registered meta-fields.

meta-fields may be registered on the advice of a Designated Expert (appointed by the IESG or their delegate), with a Specification Required (using terminology from [RFC5226]).

Registration requests consist of the completed registration template Section 7.2.1, typically published in an RFC or Open Standard (in the sense described by [RFC2026], section 7). However, to allow for the allocation of values prior to publication, the Designated Expert may approve registration once they are satisfied that an RFC (or other

Open Standard) will be published.

Upon receiving a registration request (usually via IANA), the Designated Expert should request review and comment from the apps-discuss mailing list (or a successor designated by the APPS Area Directors). Before a period of 30 days has passed, the Designated Expert will either approve or deny the registration request, communicating this decision both to the review list and to IANA. Denials should include an explanation and, if applicable, suggestions as to how to make the request successful.

7.2.1. Registration Template

Field name: The name requested for the new meta-field. This MUST conform to the host-meta field specification details noted in Section 3

Change controller: For RFCs, state "IETF". For other open standards, give the name of the publishing body (e.g., ANSI, ISO, ITU, W3C, etc.). A postal address, home page URI, telephone and fax numbers may also be included.

Specification document(s): Reference to document that specifies the field, preferably including a URI that can be used to retrieve a copy of the document. An indication of the relevant sections may also be included, but is not required.

Related information: Optionally, citations to additional documents containing further relevant information.

7.2.2. The Link host-meta field

This specification registers one host-meta field.

Field name: Link
Change controller: IETF
Specification document(s): [[this document]]
Related information: [I-D.nottingham-http-link-header]

8. References

8.1. Normative References

[I-D.nottingham-http-link-header]
Nottingham, M., "Link Relations and HTTP Header Linking", draft-nottingham-http-link-header-03 (work in progress), November 2008.

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

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

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.

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

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

8.2. Informative References

[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.

[RFC3864] Klyne, G., Nottingham, M., and J. Mogul, "Registration Procedures for Message Header Fields", BCP 90, RFC 3864, September 2004.

[RFC4918] Dusseault, L., "HTTP Extensions for Web Distributed Authoring and Versioning (WebDAV)", RFC 4918, June 2007.

[W3C.REC-P3P-20020416] Marchiori, M., "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification", W3C REC REC-P3P-20020416, April 2002.

URIs

[1] <http://lists.w3.org/Archives/Public/www-talk/>

Appendix A. Acknowledgements

We would like to acknowledge the contributions of everyone who provided feedback and use cases for this draft; in particular, Phil Archer, Dirk Balfanz, Tim Bray, Paul Hoffman, Barry Leiba, Ashok Malhotra, Breno de Medeiros, and John Panzer. The authors take all responsibility for errors and omissions.

Appendix B. Frequently Asked Questions

B.1. Is this mechanism appropriate for all kinds of metadata?

No. The primary use cases are described in the introduction; when it's necessary to discover metadata or policy before a resource is accessed, and/or it's necessary to describe metadata for a whole authority (or large portions of it), host-meta is appropriate. In other cases (e.g., fine-grained metadata that doesn't need to be known ahead of time), other mechanisms are more appropriate.

B.2. Why not use OPTIONS * with content negotiation to discover different types of metadata directly?

Two reasons; a) OPTIONS is not cacheable -- a severe problem for scaling -- and b) it is not well-supported in browsers, and difficult to configure in servers.

B.3. Why not use a META tag or microformat in the root resource?

This places constraints on the format of an authority's root resource to be HTML or similar. While extremely common, it isn't universal (e.g., mobile sites, machine-to-machine communication, etc.). Also, some root resources are very large, which would place additional overhead on clients and intervening networks.

B.4. Why not use response headers on the root resource, and have clients use HEAD?

The headers on a root resource pertain to that resource, not the whole site. While it is possible to mint new message headers that apply to the whole site, such a header would need to be sent on every response for the root resource, whether it was useful or not, with the potential for substantially increasing the size of those responses (which are often popular, and not very cacheable).

B.5. Why scope metadata to an authority?

The alternative is to allow scoping to be dynamic and determined locally, but this has its own issues, which usually come down to a) an unreasonable number of requests to determine authoritative metadata, b) increased complexity, with a higher likelihood of implementation and interoperability (or even security) problems. Besides, many mechanisms on the Web already presume a single authority scope (e.g., robots.txt, P3P, cookies, javascript security), and the effort and cost required to mint a new URI authority is small and shrinking.

B.6. Why /host-meta?

It's short, descriptive and according to search indices, not widely used.

B.7. Aren't you concerned about pre-empting an authority's URI namespace?

Yes, but it's unfortunately a necessary (and already present) evil; this proposal tries to minimise future abuses.

B.8. Why use link relations instead of media types to identify kinds of metadata?

A link relation declares the intent and use of the link (or inline content, when present); a media type defines the format and processing model for those bits.

B.9. What impact does this have on existing mechanisms, such as P3P and robots.txt?

None, until they choose to use this mechanism.

B.10. Why not (insert existing similar mechanism here)?

We are aware that there are several existing proposals with similar functionality. In our estimation, none have gained sufficient traction. This may be because they were perceived to be too complex, or tied too closely to one use case.

Appendix C. Document History

[[RFC Editor: please remove this section before publication.]]

- -01
  - Changed "site-meta" to "host-meta" after feedback.
  - Changed from XML to text-based header-like format.
  - Remove capability for generic inline content.
  - Added registry for host-meta fields.
  - Clarified scope of metadata application.
  - Added security consideration about HTTP vs. HTTPS, expanding scope.

Authors' Addresses

Mark Nottingham

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

Eran Hammer-Lahav

Email: eran@hueniverse.com
URI: http://hueniverse.com/