Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: June 11, 2017

E. Zierau, Ed.
The Royal Library of Denmark
December 8, 2016

Scheme Specification for the pwid URI
draft-pwid-uri-specification-00

Abstract

This document specifies a Uniform Resource Identifier (URI) for Persistent Web IDentifiers to web archives using the 'pwid' scheme name. The purpose of the standard is to support general, global, sustainable, humanly readable and technology agnostic persistent web references that are not sufficiently covered by existing web reference practices. Since only archived web can reach a degree of persistency. The 'pwid' URI primarily aim at references into web archives.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The motivation for defining a pwid URI scheme is the growing challenge of references to web resources, which are poorly supported in citation guidelines. Citation guidelines generally don't cover general and persistent referencing techniques for web resources that are not registered by Persistent Identifier systems (like DOI [DOI]). However, an increasing number of references point to resources that only exist on the web. Such web referencing is highly relevant and crucial for various research fields. For example blogs that shows out to have a historical impact.

Today there are different ways to refer to web references that are not registered:

- By specifying http/https address and the date it was visited

- Via Citation services - which constructs citation http/https addresses that are not following a general scheme and that will change or be lost in case the citation service change domain, service

- Via Web Archive http/https access addresses - which can only be used for open web archives. Furthermore, Web archive http/https

addresses are not following a general scheme, and they will change or be lost in case the web archive service changes domain, or changes path to the web archive service

Http/https address and date is in no way persistent, and is the main reason for studies showing that a large percentage of links in research studies are dead after a relatively short period. Citation services can sometimes be used, but responsibility of preservation and collection is not fully clear, and they often use http/https address shorteners for access, which complicates preservation of source and metadata even more.

Finally, there are the web archives that offer access openly or locally, but where access http/https addresses depends on domains for the web archive as well as differing paths to their access service.

The 'pwid' URI Scheme is another step in facilitating, supporting, and standardizing the problem of persistent web references to resources in web archives. Accessing a referenced web resource will require APIs from web archives no matter whether they are open web archive or not. There are different solutions for the resolving of a 'pwid' URI, which needs to be investigated and implemented as use and support of the 'pwid' URI evolves.

According to RFC 3986 [RFC3986]], a Uniform Resource Identifier (URI) is "a compact sequence of characters that identifies an abstract or physical resource". The 'pwid' URI Scheme defined in this document identifies web archive resources (abstract resources) in a general, global, stainable, humanly readable and technology agnostic way. An example of such a 'pwid' URI follows:

```
pwid:archive.org:2016-01-22T11.20.29Z_page:http://www.dr.dk
```

In this example the domain of the archive has been used as identifier. However, an archive identifier does NOT need to be a domain. The choice in the example is only to use a short archive identifier that is already associated with the archive.

For the sake of usability and sustainability, the definition of the 'pwid' URI scheme is focused on only having the minimum required information in order to precisely identify a resource in an arbitrary web archive.

1.1. Requirements Language

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].

2. Demonstrable, New, Long-Lived Utility

The 'pwid' URI scheme allows identification of web archive resources in a general, global, sustainable, humanly readable and technology agnostic way. No matter whether it will become resolvable, it can be used at any time to identify the web archive resource, as long as the material exists.

The 'pwid' is defined as a URI as there are great potentials for making it resolvable. This means it could function as a URN RFC 2141 [RFC2141], but is not defined as such as the ambition is to make it resolvable. At the same time the 'pwid' definition can enjoy the same common syntactic, semantic, and shared language benefits that the URI presentation confers.

The interest for this new 'pwid' URI scheme has already been shown, a long paper about the invention of the 'pwid' URI "Persistent Web References - Best Practices and New Suggestions" [IPRES] was accepted for the iPres 2016 conference and nominated as best paper.

There is no question about the need for a standard to address web materials meeting precision and persistency issues on par precision in with traditional references for analogue material. The interest for the 'pwid' URI indicates that this is a recognized issue, and that the 'pwid' URI can fill a gap.

The 'pwid' URI will benefit from becoming resolvable to some extent, but as a start it has a value even without being resolvable.

3. Syntactic Compatibility

The syntax of the 'pwid' URI Scheme is specified below in Augmented Backus-Naur Form (ABNF) RFC 5234 [RFC5234] and it conforms to URI syntax defined in RFC 3986 [RFC3986]. The syntax definition of the 'pwid' URI is:

```
pwid-uri  = pwid-scheme ":" pwid-spec
pwid-spec = archive-id ":" archived-date [ content-spec ]
            ":" archived-item
```

where

```
pwid-scheme = "pwid"
archive-id  = +( unreserved )

archived-date   = full-date datetime-delim full-pwid-time
datetime-delim  = "_" / "T"
```

```
     full-pwid-time  = time-hour ["."] time-minute ["."] time-second "Z"
     content-spec    = "_page" / "_part" / "_coll" / "_snapshot"
                       / "_rec" / "_other"

     archived-item = URI / archived-item-id
     archived-item-id  = +( unreserved )
```

where

- 'unreserved' is defined as in RFC 3986 [RFC3986]

- 'content-spec' values are not case sensitive (i.e. "_PAGE" / "_PART" / "_PaGe" / ... are valid values as well.)

- 'archived-date' is a UTC timestamp conforming to the W3C profile ISO8601 also defined in RFC 3339 [RFC3339], with a few exception for the datetime-delim and full-pwid-time, and "." is used instead of ":" in order not to collide with ":" used for delimitation of URI parts. This means that full-date is defined as in RFC 3339 [RFC3339].
  The datetime-delim "_" is accepted in order to make it more readable, in the same way as the W3C profile accepts " ", but where "_" is used here in order to use allowed URI characters in an URI. In line with RFC 3339 [RFC3339] the "T" may alternatively be lower case "t".
  time-hour, time-minute and time-second is defined as in RFC 3339 [RFC3339].
  In line with RFC 3339 [RFC3339] the "Z" may alternatively be lower case "z".

- 'URI' is defined as in RFC 3986 [RFC3986]

Note that the 'content-spec' is a parameter that could have been specified as a query. However, since the 'pwid-uri' can include an URI as 'archived-item', it would introduce ambiguities if the 'content-spec' was specified as a query, since it would not be clear whether the query belonged to the 'pwid-uri' or the 'archived-item'.

The 'content-spec' defines the type of archived item. This serves as a precision to what is referred:

- when a URI is specified, since it can be what was harvested for the specific URI, or it can be the URI interpreted as a web page within the context of the specified archive

- it can specify the type of contents expected for an identifier ('archived-item-id'), which can be a collection a snapshot, a recording or other identifier

4. Well Defined

The information in a 'pwid' URI can be used for locating a web archive resource, for any kind of web archive. It includes the minimum information for web archive materials which enables resolvability, manually or by a resolver. One of the reasons for defining 'pwid' as a URI is to open the possibility to make a generally resolvable representation.

The information needed is:

- Web archive identification
  to find the archive holding the material

- Archived URI or identifier of item
  as part of identifying the material

- Date and time associated with the archived URI/item
  as part of precise identification of the material

- Specification of what is referred
  as part of clarification of what is referred

For example the 'pwid' URI:

```
pwid:archive.org:2016-01-22T11.20.29Z_page:http://www.dr.dk
```

has the information:

- archive.org
  current known identifier of Internet Archives open access web archive

- 2016-01-22T11.20.29Z
  Date and time associated with the archived URI

- page
  Clarification that it is the web page that is being referred

- http://www.dr.dk
  Archived URI of item

With knowledge of the current (2016) Internet Archive open access web interface having the form:

```
https://web.archive.org/web/<digit date>/<uri>
```

We can manually (or technically) deduce an actual (current 2016) access https address:

https://web.archive.org/web/20160122112029/http://www.dr.dk

and regard the referred web page as the reference.

The same recipe can be used for other Wayback platforms - and possibly also other web archive access tools platforms, as the crucial information is date and URI which are requested to be looked up in a specified archive.

Note that this also includes access to archives that are only accessible via a local proxy to a restricted environment. Here the difference is that the archive information is used to identify the local environment used (possibly on-site) and then construct local http/https address based on knowledge from the local access installation.

5. Definition of Operations

There is not a specific definition of computational operation yet, but there will be ongoing work to see if it can be put into operation in different ways.

There may be a need for varied operation depending on whether a web archive is open online, or whether it is a closed archive that only works in a restricted environment.

At this stage there are initiatives on streamlined APIs to web archives, - and in case such an API will be implemented generally, it may be used for resolving of the 'pwid' URIs.

Because of the case of closed archives, the 'pwid' URI resolving can in such cases be a question of starting a special application, as for the 'mailto' scheme RFC 6068 [RFC6068].

For open archives resolving could be a matter of creating an http/https address based on knowledge of the archive and access interfaces to the archive. In the latter case this would require:

1. An archive registry
   as a start the current archive domains could be used, but as soon as domains are changed the validity of a 'pwid' URI will be dependent on such a registry.

2. Open access http/https address pattern registry

this would only make sense for the open web archives, and it does not need to be a formal registry, since the pattern can be found (manually) as long as the archive is identifiable. Thus the validity of a 'pwid' URI does not depend on such a registry.

In all cases the 'pwid' URI can be used for 'manual' look up as described in the previous section.

6. Context of Use

Typically, 'pwid' URIs will be used for references to web resources in web archives, e.g. in research or scholarly work. However, it may also be used for research data management specification (specifying specific target of archived contents from an http/https address) or applications that are restricted to access a specific set of archived contents from http/https addresses in a web archive. When the references are listed in hypertext documents, these will become resolvable in case the pwid URI becomes resolvable.

As described above, there may come different implementations for resolving which may rely on different protocols and application; - from redirects to the http/https protocol to call of locally installed browser plug-ins or applications.

7. Internationalization and Character Encoding

Internationalization and character encoding for 'pwid' URIs are relevant for the webarchive-id and archived-uri parts of the scheme-specific-part of the 'pwid' URI, since both archived-date and content-spec only can be constructed by a very limited set of characters.

The webarchive-id will not be case sensitive, but can allow for percent encodings, although for simplicity reasons, it may turn out that the coming establishment of an archiving registry will recommend using letters that do not need encodings.

The archived-uri follows the rules of URIs in general (currently for http and https URIs archived in web archives). The archived-uri is only case sensitive to the extent that the web archive can handle archived case sensitive URIs.

8. Scheme Name Considerations

The scheme name is "pwid" - short for Persistent Web Identifier. Initially the scheme name "wpid" was reserved. However, one of the feedbacks has been a concern that "wpid" was interpreted as a PID related to a PID-system, e.g. as the DOI. All though PID does not

have a precise definition that makes it wrong to call it a "wpid", the danger is that it is confused with PID systems which is not the intension. Consequently, this suggestion names the scheme "pwid" instead.

9. Interoperability Considerations

This is covered by comments on the date in the section of Syntactic Compatibility, where the archived-date conforms to the W3C profile ISO8601, except for minor modification in order to make it fit into a URI. Furthermore, the archived-uri conforms to the URI standard.

10. Acknowledgements

Thanks to all that have contributed to this work in creating the iPres paper, commenting at the iPres conference and reviewing this RFC

11. IANA Considerations

The pwid URI scheme is reserved as a provisional URI as result of request IANA #938449

12. Clear Security and Privacy Considerations

Security and privacy considerations are restricted to accessible web resources in web archives. If resolvers to 'pwid' URIs are created, there should be made an analysis of whether they can be restricted to the former mentioned registry of web archives. Security and privacy will then be a question of security and privacy considerations related to the web archive resources.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC3339] Klyne, G. and C. Newman, "Date and Time on the Internet: Timestamps", RFC 3339, DOI 10.17487/RFC3339, July 2002, <http://www.rfc-editor.org/info/rfc3339>.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, DOI 10.17487/RFC3986, January 2005, <http://www.rfc-editor.org/info/rfc3986>.

[RFC5234] Crocker, D., Ed. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, DOI 10.17487/RFC5234, January 2008, <http://www.rfc-editor.org/info/rfc5234>.

13.2. Informative References

[DOI] International DOI Foundation, "The DOI System", 2016, <https://web.archive.org/web/20161020222635/ https:/www.doi.org/>.

pwid:archive.org:2016-10-20_22.26.35_page:https://www.doi.org/

[IPRES] Zierau, E., Nyvang, C., and T. Kromann, "Persistent Web References - Best Practices and New Suggestions", October 2016, <http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf>.

In: proceedings of the 13th International Conference on Preservation of Digital Objects (iPres) 2016, pp. 237-246

[RFC2141] Moats, R., "URN Syntax", RFC 2141, DOI 10.17487/RFC2141, May 1997, <http://www.rfc-editor.org/info/rfc2141>.

[RFC6068] Duerst, M., Masinter, L., and J. Zawinski, "The 'mailto' URI Scheme", RFC 6068, DOI 10.17487/RFC6068, October 2010, <http://www.rfc-editor.org/info/rfc6068>.

Author's Address

Eld Maj-Britt Olmuetz Zierau (editor)
The Royal Library of Denmark
Soeren Kierkegaards Plads 1
Copenhagen 1219
Denmark

Phone: +45 9132 4690
Email: elzi@kb.dk