Abstract

This document specifies a Uniform Resource Identifier (URI) for Persistent Web IDentifiers to web material in web archives using the ‘pwid’ scheme name. The purpose of the standard is to support general, global, sustainable, humanly readable, technology agnostic, persistent and precise web references for such web materials.

The PWID URI can assist in two ways: First, by providing potential resolvable precise and persistent reference scheme for documents, which is not sufficiently covered by existing web reference practices. Second, by providing a standardized way to specify web elements in a web collection also known as web corpus. Definitions of web collections are often needed for extraction of data used in production of research results, e.g. for evaluations in the future. Current practices today are not persistent as they often use some CDX version, which vary for different implementations.

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 December 11, 2017.
1. Introduction

The purpose of the PWID URI is to represent general, global, sustainable, humanly readable and technology agnostic web archive resource references - in a scheme that can be used for technical solutions. The motivation for defining a PWID URI scheme is the growing challenge of references to web resources, - both regarding referencing web resources from papers and regarding definition of web collection/corpus.

   - Citation guidelines generally do not 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, e.g. blogs that turned out to have a historical impact. In order to obtain persistency for a reference, the target need to be stable. As the live web is ‘alive’ and in constant change, persistency can only be obtained by referring to archived snapshots of the web. The PWID URI is therefore focused on referencing archived web material in a technology agnostic way (research documented in [IPRES] and [ResawRef]).

The PWID URI scheme is focused on only having the minimum required information to make a precise identification of a resource in an arbitrary web archive. Recent research have found that this is obtain by the following information [ResawRef]:

- Identification of web archive
- Identification of source:
  - Archived URI
  - Archival timestamp
- Intended coverage (page, part, subsite etc.)

The PWID URI scheme represents this information in an unambiguous way, and thus enabling technical solutions to be defined based on this scheme.

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 purpose of the PWID URI is to represent needed referencing information (as listed in the introduction) in a scheme that can be used for technical solutions. As described in [ResawColl] such references can be represented in a textual way. However, strict unambiguous syntax is needed in order to ensure that it can be used for computational purposes. This is relevant for web collection definitions, which will need a strict scheme in order to be a basis for automatic extraction. Furthermore, readers of research papers are today expecting to be able to access a referenced resource by clicking an actionable URI, therefore a similar facility will be expected for references to available archived web material.

The interest for this new PWID URI scheme has already been shown, a 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. At the RESAW 2017 conference there are two related papers: One on referencing practices [ResawRef] and one on research data management practices [ResawColl]. The interest for the PWID URI so far indicates that this is a recognized issue, and that the PWID URI can fill a gap.

The PWID URI could function as a URN RFC 2141 [RFC2141], but is not defined as such as the ambition is to make an easily understandable and technology independent persistent identifier, where the prefixing of "urn:" will be disturbing. At the same time the PWID definition can enjoy the same common syntactic, semantic, and shared language benefits that the URI presentation confers.

It should be noted that for closed web archives, the PWID URI can be used to resolve within a closed environment. Likewise, the PWID can be resolved within coming web archive research infrastructure, which is currently being proposed in the RESAW community [RESAW].

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-scheme  = "pwid"
pwid-spec  = archive-id "::" archival-time "::" coverage-spec
: archived-item
archive-id = +( unreserved )

archival-time = full-date datetime-delim full-pwid-time
datetime-delim = "_" / "T"
full-pwid-time = time-hour ["."] time-minute ["."] time-second "Z"

coverage-spec = "part" / "page" / "subsite" / "site"
               / "collection" / "recording" / "snapshot"
               / "other"

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

where

- ‘unreserved’ is defined as in RFC 3986 [RFC3986]
- ‘coverage-spec’ values are not case sensitive (i.e. "PAGE" / "PART" / "PaGe" / ... are valid values as well.)
- ‘archival-time’ is a UTC timestamp conforming to the W3C profile ISO8601 ISO 8601 [ISO8601] (also defined in RFC 3339 [RFC3339]), with a few exception for the ‘datetime-delim’ and ‘full-pwid-time’, as well as using "." is used instead of ":" in order not to collide with ":" used for delimitation of URI parts. The ‘full-date’ is defined as in RFC 3339 [RFC3339]. The ‘archival-time’ must represent the time specified in the archive, and can therefore be specified at any of the levels of granularity as described in [W3CDTF] and in accordance with teh WARC standard ISO 28500 [ISO28500].

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’ are 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]

The ‘coverage-spec’ defines the type of archived item, serving as a precision to what is referred:
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 enable a general, technology agnostic, persistent representation to be resolvable at any time.

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

Note that the ‘coverage-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 ‘coverage-spec’ was specified as a query, since it would not be clear whether the query belonged to the ‘pwid-uri’ or the ‘archived-item’.
o Date and time associated with the archived URI/item
   as part of precise identification of the material

o Coverage of what is referred
   as part of clarification of what the referred material covers
   (page, part etc.)

For example the PWID URI:


has the information:

o archive.org
   currently known identifier in form of the Internet Archive domain
   name for their open access web archive

o 2016-01-22_11.20.29Z
   date and time associated with the archived URI

o page
   clarification that the reference cover the full web page with all
   its inherited parts selected by the web archive

o http://www.dr.dk
   archived URI of item

With knowledge of the current (2017) Internet Archive open access web
interface having the form:

   https://web.archive.org/web/<time>/<uri>

We can manually (or technically) deduce an actual (current 2017)
access https address:


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

The PWID URI Scheme is another step in facilitating, supporting, and standardizing the problem of persistent web references to resources in web archives. There is not a specific definition of computational operation yet. It is expected that there may be different implementations in pace with needed use and available technology and infrastructures.

Automatic access of a referenced web resource may work on the open net for open web archive or in restricted environments for the closed web archives. There may be a need for varied operation depending on the available technology and applications, e.g.:

- Via locally installed browser plug-ins or applications forming http/https URIs:
  
  * http/https URIs for standard web archive interfaces
  At this stage there are initiatives on streamlined and standardize APIs to web archives interfaces, - and in case such APIs will be implemented generally, it may be used for resolving of the PWID URIs. This could be on form (denoting pwid parts in <> using syntax names):

    https://<archive-id>/pwid?time=<archival-time>&coverage=<coverage-spec>&item=<archived-item>

  The example from previous section would then resolve by


  * http/https URIs for archive material for individual web archives
  Using the current open access http/https address pattern for the individual web archives, which for the example is


  This would require a registry of the different patterns for the individual web archives

- Via web research infrastructures this is a future solution scenario as a web archive research infrastructure do not yet exists. However, it is a likely future scenario, as it is
currently being proposed in the RESAW community [RESAW]. The PWID URI resolving could in such cases be a question of starting a special application, as for the ‘mailto’ scheme RFC 6068 [RFC6068].

Use of URIs for standard web archive interfaces is preferred as dependency on registries and infrastructures may pose too many limits.

6. Context of Use

The PWID URI scheme facilitates, supports and standardise a scheme for specification of identification of web archive resources in a general, global, sustainable, humanly readable and technology agnostic way. The standard is needed to address web materials meeting precision and persistency issues on par precision in with traditional references for analogue material.

The purpose with the PWID URI is to represent this information in a scheme that can be used for technical solutions, for example for resolving of a references and automatic extraction of web collection defined by PWID URIs [ResawRef] [ResawColl]. As described above, there may come different implementations for resolving which may rely on different protocols and application.

7. Internationalization and Character Encoding

Internationalization and character encoding for PWID URIs are relevant for the ‘webarchive-id’ and ‘archived-item’ syntactical units of the scheme-specific-part of the PWID URI. The rest of the main syntactical units (‘archival-time’ and ‘coverage-spec’) are only constructed by a very limited set of characters, and do therefore need internationalization and character encoding.

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-item’ follows the rules of URIs in general (currently for http and https URIs archived in web archives). The ‘archived-item’ 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 ‘archival-time’ conforms to the W3C profile ISO8601, except for minor modification in order to make it fit into a URI. Furthermore, the ‘archived-item’ conforms to the URI standard.

10. Acknowledgements

A special thanks to Caroline Nyvang and Thomas Kromann who have contributed to the research identifying the minimum information required in a persistent web reference, and to Bolette Jurik who contributed with supplementary research concerning requirements for web collection/copora definitions. Also thanks to all that have contributed to this work with the research and reviewing this RFC.

11. IANA Considerations

The URI scheme name ‘pwid’ 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


13.2. Informative References


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


