A Persistent Web IDentifier (PWID) URN Namespace
draft-pwid-urn-specification-05

Abstract

This document specifies a Uniform Resource Name (URN) for Persistent Web IDentifiers for web material in web archives using the ‘pwd’ namespace identifier.

The main purpose of the standard is to support specification of references that are not covered by other reference techniques: to support references to material in web archives with restricted access. Furthermore, it supports persistent technology agnostic references to web archives in general, in a form that can work as an algorithmic basis for finding web archive resources in general. An additional important benefit is that the standard can be used for specifying web collections, which can then form a persistent computational basis for the extract of the archived collection parts. Since these parts can be specified generally, this further allows collections to be specified with elements from one or more web archives.

The PWID URN is designed to meet requirements for proper referencing needed by researchers. Therefore it is designed as general, global, sustainable, humanly readable, technology agnostic, persistent and precise web references for web materials in web archives.

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 https://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."
1. Introduction

The PWID URN is a supplement to existing reference standards, where the PWID URN will support references to web archives, including areas that are not supported today: support of references to material in web archives with restricted access. Furthermore, the PWID URN enables technology agnostic references to web archives in general, which can be needed, for instance for references to dynamic web material with frequent updates (e.g. a news site) or a specific version of a web material (e.g. specific version of the DOI handbook).

The PWID URN is in a form which can work as an algorithmic basis for finding the resource. This also enables computation of archived web parts to a collection from one or more web archives, if the collection parts are specified by PWID URNs.

Furthermore, the PWID URN includes information about the resource which makes it possible to find alternative resources, in cases where the original precise resource has become unavailable.
The PWID URN is designed to be a persistent reference that is
general, global and technology agnostic in order to enhance its
chances of being sustainable. Furthermore, it is designed to be
humanly readable and with an ability to specify precision about what
the referenced web archive resource covers. This design enables a
PWID URN to:

- be used in technical solutions, e.g. to make them resolvable
- cover references to all sorts of materials in web archives
- cover references to materials from all sorts of web archives

The motivation for defining a PWID namespace is the growing
challenges of references to archived web resources, and the PWID as a
URN can assist in overcoming a lot of these challenges. The standard
is needed to address web materials meeting precision and persistency
issues on par precision in traditional references for analogue
material. Furthermore, it is needed in order to address web archive
resources that are not freely available online. The PWID URN covers
both referencing of web resources from research papers and definition
of web collections/corpora. In detail the challenges are:

- Persistent Identifier systems (like DOI [DOI]) will only cover
  registered resources. In general, citation guidelines do not
cover general and persistent referencing techniques for web
resources that are not registered. However, an increasing number
of references point to resources that only exist on the web, e.g.
blogs that turn out to have a historical impact. In order to
obtain persistency for a reference, the target needs to be stable.
For non-registered web resources, the common rule is that the
resource will change, since the live-web is constantly changing.
Persistency can only be obtained by referring to something stable,
i.e. an archived snapshot of the resource from the web. The PWID
URN is therefore focused on referencing archived web material in a
technology agnostic way (research documented in [IPRES2016] and
[ResawRef]).

- References to materials, which only exist in web archives (i.e. no
  longer on the live web) are not well supported, especially not for
materials that only exists in archives with restricted access.
There are many new initiatives for web archive referencing, - most
of which are centralized solutions offering harvesting and
referencing, but these cannot be used for materials that only
exist in web archives. The PWID URN can be used for all web
archives, including web archives with restricted access.
One of the referencing initiatives for open web archives uses URLs which depend on the current setup of the web archive’s access platform. These URLs are usually technology and placement dependent, and therefore such a reference style is not suited for references that are important to retrace for a long period. The PWID URN can be used for such reference purposes, since it is technology agnostic.

Another referencing initiative, for open web archives, is omitting specification of the web archive where the resource was found. This strategy is used in order to open the possibility of using alternatives from other archives. However, this also adds a risk of imprecision since different archives tend to have different versions even when harvesting at the same time. Therefore, such a reference style is not suited for references where it is important that the reference is precisely the verified reference. The PWID URN can provide an exact reference for where the reference was validated. Additionally, the PWID contains the needed information in order to search for alternative resource, if needed.

For reference of web collections/corpora (possibly across different web archives), recent research have found that various legal and sustainability issues has led to a need of a collection definition of references to their web parts. Furthermore, there is a need for a similar persistent referencing for all parts for calculation and sustainability reasons. So far, there has been no stable standard for definition of such collection parts. The PWID URN can be used for such definitions in order to fulfil these requirements (research documented in [ResawColl]).

The PWID URN is especially useful for web material where precision is in focus and/or there are references to materials from web archives requiring special permissions in order to gain access. The precision regards both pointing to the archive where it was found and validated against its purpose (other archived versions in other web archives may differ both regarding completeness and contents even within short time periods) as well as precision in what is actually referred by the reference (e.g. is it the page or the whole website).

Furthermore, the PWID URN is very useful in specification of contents of a web collection. Definitions of web collections are often needed for extraction of data used in production of research results, e.g. for future evaluations. Current practices are not persistent as they often use some CDX version, which vary for different implementations.

Strict syntax is needed for the PWID URN, in order to ensure that it can act as a reference which can used for computational purposes. This is especially relevant for automatic extraction of parts from
web collection definitions. Furthermore, today’s readers of research papers are expecting to be able to access a referenced resource by clicking an actionable URI, therefore a similar possibility will be expected for references to available archived web material, and this is possible with a strict syntax. Examples of technical solutions that are enabled are:

- Resolving of a reference to a web collection and automatic extraction of the parts of a web collection defined by PWID URNs [ResawRef] [ResawColl]
- Resolving of a PWID URN by resolving services. To begin with, a prototype has been developed for the Danish web archive data and open web archives with standard patterns for the current technologies. Implementations for resolution of PWID URNs for other web archives may be developed.

The purpose of the PWID URN is also to express a web archive reference as simple as possible and at the same time meet the requirements for sustainability, usability and scope. Therefore, the PWID URN is focused on having only the minimum required information to make a precise identification of a resource in an arbitrary web archive. Recent research have shown that this can be obtained by the following information [ResawRef]:

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

The PWID URN represents this information in a human readable way as well as a well-defined way that enables technical solutions to interpret the URN.

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 [RFC2119].
2. Namespace Registration Template

Namespace Identifier:

PWID

Version:

5

Date:

2019-02-27

Registrant:

Eld Maj-Britt Olmuetz Zierau
Royal Danish Library
Soeren Kierkegaards Plads 1
1219 Copenhagen
Denmark
ph: +45 9132 4690
e-mail: elzi@kb.dk

Purpose:

The PWID URN is a supplement to existing reference standards, where the PWID URN will support references to web archives, including areas that are not supported today: support of references to material in web archives with restricted access. Furthermore, the PWID URN enables technology agnostic references to web archives in general, which can be needed, for instance for references to dynamic web material with frequent updates (e.g. a news site) or a specific version of a web material (e.g. specific version of the DOI handbook).

The PWID URN is in a form which can work as an algorithmic basis for finding the resource. This also enables computation of archived web parts to a collection from one or more web archives, if the collection parts are specified by PWID URNs.

Furthermore, the PWID URN includes information about the resource which makes it possible to find alternative resources, in cases where the original precise resource has become unavailable.

The PWID URN is designed to be a persistent reference that is general, global and technology agnostic in order to enhance its
chances of being sustainable. Furthermore, it is designed to be humanly readable and with an ability to specify precision about what the referenced web archive resource covers. This design enables a PWID URN to:

* be used in technical solutions, e.g. to make them resolvable
* cover references to all sorts of materials in web archives
* cover references to materials from all sorts of web archives

The motivation for defining a PWID namespace is the growing challenges of references to archived web resources, and the PWID as a URN can assist in overcoming a lot of these challenges. The standard is needed to address web materials meeting precision and persistency issues on par with precision in traditional references for analogue material. Furthermore, it is needed in order to address web archive resources that are not freely available online. The PWID URN covers both referencing of web resources from research papers and definition of web collections/corpora.

In detail the challenges are:

* Persistent Identifier systems (like DOI [DOI]) will only cover registered resources. In general, citation guidelines do not cover general and persistent referencing techniques for web resources that are not registered. However, an increasing number of references point to resources that only exist on the web, e.g. blogs that turn out to have a historical impact. In order to obtain persistency for a reference, the target needs to be stable. For non-registered web resources, the common rule is that the resource will change, since the live-web is constantly changing. Persistency can only be obtained by referring to something stable, i.e. an archived snapshot of the resource from the web. The PWID URN is therefore focused on referencing archived web material in a technology agnostic way (research documented in [IPRES2016] and [ResawRef]).

* References to materials, which only exist in web archives (i.e. no longer on the live web) are not well supported, especially not for materials that only exists in archives with restricted access. There are many new initiatives for web archive referencing, - most of which are centralized solutions offering harvesting and referencing, but these cannot be used for materials that only exist in web archives. The PWID URN can be used for all web archives, including web archives with restricted access.
* One of the referencing initiatives for open web archives uses URLs which depend on the current setup of the web archive’s access platform. These URLs are usually technology and placement dependent, and therefore such a reference style is not suited for references that are important to retrace for a long period. The PWID URN can be used for such reference purposes, since it is technology agnostic.

* Another referencing initiative, for open web archives, is omitting specification of the web archive where the resource was found. This strategy is used in order to open the possibility of using alternatives from other archives. However, this also adds a risk of imprecision since different archives tend to have different versions even when harvesting at the same time. Therefore, such a reference style is not suited for references where it is important that the reference is precisely the verified reference. The PWID URN can provide an exact reference for where the reference was validated. Additionally, the PWID contains the needed information in order to search for alternative resource, if needed.

* For reference of web collections/corpora (possibly across different web archives), recent research have found that various legal and sustainability issues has led to a need of a collection definition of references to their web parts. Furthermore, there is a need for a similar persistent referencing for all parts for calculation and sustainability reasons. So far, there has been no stable standard for definition of such collection parts. The PWID URN can be used for such definitions in order to fulfil these requirements (research documented in [ResawColl]).

The PWID URN is especially useful for web material where precision is in focus and/or there are references to materials from web archives requiring special permissions in order to gain access. The precision regards both pointing to the archive where it was found and validated against its purpose (other archived versions in other web archives may differ both regarding completeness and contents even within short time periods) as well as precision in what is actually referred by the reference (e.g. is it the page or the whole website).

Furthermore, the PWID URN is very useful in specification of contents of a web collection. Definitions of web collections are often needed for extraction of data used in production of research results, e.g. for future evaluations. Current practices are not persistent as they often use some CDX version, which vary for different implementations.
Strict syntax is needed for the PWID URN, in order to ensure that it can act as a reference which can be used for computational purposes. This is especially relevant for automatic extraction of parts from web collection definitions. Furthermore, today’s readers of research papers are expecting to be able to access a referenced resource by clicking an actionable URI, therefore a similar possibility will be expected for references to available archived web material, and this is possible with a strict syntax. Examples of technical solutions that are enabled are:

* Resolving of a reference to a web collection and automatic extraction of the parts of a web collection defined by PWID URNs [ResawRef] [ResawColl]

* Resolving of a PWID URN by resolving services. To begin with, a prototype has been developed for the Danish web archive data and open web archives with standard patterns for the current technologies. Implementations for resolution of PWID URNs for other web archives may be developed.

The purpose of the PWID URN is also to express a web archive reference as simple as possible and at the same time meet the requirements for sustainability, usability and scope. Therefore, the PWID URN is focused on having only the minimum required information to make a precise identification of a resource in an arbitrary web archive. Recent research have shown that this can be obtained by the following information [ResawRef]:

* Identification of web archive

* Identification of source:
  + Archived URI or identifier
  + Archival timestamp

* Intended precision (page, part, subsite etc.)

The PWID URN represents this information in a human readable way as well as a well-defined way that enables technical solutions to interpret the URN.

Syntax:

The syntax of the PWID URN is specified below in Augmented Backus-Naur Form (ABNF) [RFC5234] and conforms to URN syntax defined in [RFC8141]. The syntax definition of the PWID URN is:
pwid-urn = "urn" "." pwid-NID "." pwid-NSS

pwid-NID = "pwid"
pwid-NSS = archive-id "." archival-time "." precision-spec 
"." archived-item

archive-id = +( unreserved )

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

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

where

* All parts of the pwid-NSS are case insensitive, except for archived-item in cases where the archived-item is an URI with case sensitive parts. According to [RFC8141] (section 3.1) this means that the PWID URNs in general are case insensitive, except from cases where it includes a case sensitive URI as archived-item.

* 'archival-time' is a UTC timestamp as described in the W3C profile of [ISO8601] [W3CDTF] (also defined in [RFC3339]), for example YYYY-MM-DDThh:mm:ssZ. The 'archival-time' must represent the timestamp that the web archive have recorded for the referenced archived URI. The archival-time may be specified at any level of granularity described in [W3CDTF], as long as it reflects exactly the granularity of the timestamp recorded in the web archive, which is in accordance with the WARC standard [ISO28500].

* 'unreserved' is defined as in [RFC3986].

* 'URI' is defined as in [RFC3986] but where occurrences of "[", "]", "?" and "#" are %-encoded in order not to clash with URN reserved characters [RFC8141].

The precision specification is expressing the intended precision of the reference. For example, if it refers to an html web element, this element can be interpreted in several ways:

* As one web part only
  Meaning the file containing the html, and precisely this file

* As a web page
Meaning that an application like Wayback shows a resulting web page in a browser based on calculated referenced web parts (display templates, images etc.).

If the full reference contains only the PWID URN for the page, this may mean that the archived page can change its appearance over time, e.g. if parts referred by the page did not exist at reference time, but are harvested at a later stage, - or if the web archive’s algorithm for calculation of the referred web parts are changed and consequently returns a different result. Therefore, the most a precise reference to a picture in context of a web page would be to provide the PWID URN for the page (with page precision) and the PWID URN for the image file part which contains the referred picture (with part precision)

* As a site or subsite
Meaning that an application like Wayback shows the result in a browser showing the web page. If access is limited to the referenced part (the html page), then the application would also need to make sure that all parts/pages belonging to the site/subsite is available.

If the full reference only contains the PWID URN for the site/subsite, this may mean that the site/subsite can change its appearance over time in the same way as for the web page described above

The precision specification needs to be part of a PWID URN in order to enable the making of the above described precision in the reference. Furthermore, this precision specification will make it possible for resolvers to display the referred source in a way that corresponds to the precision specification.

There are different ways to represent e.g. a web page, which provides different precision of the source as well. The above examples with part, page, subsite and site are addressing the most common access via browser functionality like in Wayback. However, some web archives archive snapshots of the web pages for the archived URI. A third option is to produce a collection of archived URIs as basis for browser access instead of letting the web archive calculate sub items (which may change over time). An example of the production of such a collection is provided in the section about assignment. Lastly, a web page may be archived via a web recording.

Because of the above, the following valid precision-spec values are exists:

* part
The single archived web part harvested as a file from the specified URI, e.g. a pdf, an html text or an image

* page
The web page represented by the web page file (e.g. html) harvested from the specified URI, where its content is interpreted as a web page with all referred parts relevant to display the web page (but where referred parts must be calculated as described above), e.g. an html page with referred images

* subsite
The referred web page (as described under ‘page’) from which is possible to browse to all references starting with the same path as the archived URI

* site
The referred web page (as described under ‘page’) from which is possible to browse to all references in the domain specified in the archived URI

* collection
Representation of a collection specification, where the web archive applications will decide how it is rendered (e.g. collection specification in the XML format enabling interpretation as in the example provided in [ResawColl])

* snapshot
A snapshot (image) representation of web material, e.g. a web page

* recording
Representation of a web recording specification where the web archive applications will decide how it is rendered (interpretation could e.g. depend on file-suffix for the web recording), an example is a web recording coded in a WARC file

* other
This is a placeholder to allow reference of a resource of any kind with an assigned identifier (by the archive). In all cases, it will be up to the application serving the web archive to interpret how this item should be rendered

Assignment:
The PWID URNs do not have to be assigned by an authority, as they are based on the information created at the time of archiving. In other words: a PWID URN is created independently, but following an
algorithm which ensures that the referred item can be found if it is still available. A PWID URN also has the benefit that it includes information to look at alternative resources e.g. via Memento for some open web archives [MEMENTO] or via possible future web archive infrastructures.

A PWID URN is created by finding the relevant information of the syntax parts of the PWID:

"urn:pwid:" archive-id ":" archival-time ":" precision-spec ":" archived-item

The PWID URN for an archived item at hand can be constructed by exchanging the unspecified PWID parts with relevant information, as explained in the following:

* archive-id (identification of web archive):
  In this version of the standard, it is recommended to use the domain of the web archive as the identifier for the web archive (e.g. archive.org for Internet Archive’s open web archive and netarkivet.dk for the Danish web archive with restricted access). This is recommended, since browsing the domain page will typically lead to a description of how to access the web archive, e.g. by online access or by applying for access grants. Furthermore, it is more precise than e.g. the name of the archive, since there may be more than one installation of web archives at the same organization, e.g. archive.org and archive-it.org are both covered by Internet Archive. When a registry of web archives is established, it will be more precise and persistent to use the web archive identifier specified in this registry (e.g. DKWA for the Danish web archive with the domain netarkivet.dk)

* archival-time (archival timestamp):
  The archival time for the archived item at hand may be displayed along with the archived item, but there are different implementations where it is important to be aware of whether a more precise timestamp can be found, and whether the correct timestamp is used. In many Wayback implementations, the precise timestamp can be found as part of the URI used for viewing the archived item. For example, the archive http URI https://web.archive.org/web/20160122112029/http://www.dr.dk for an archived resource viewable via the Internet Archive’s Wayback installation, the number 20160122112029 represents the archival time 2016-01-22T11:20:29Z. In other installations, the most precise timestamp may be found in the URI from a search result leading to the resource (which usually redirects on basis of a call to the underlying archive index).
Especially for web pages with frames, there may be cases where the actual time is not displayed with the source, since only the times for the contents of the frames are displayed.

* precision-spec (precision as represented page, part, site, snapshot etc.):
The precision specification specifies how the user should view the referred item - either as a specific representation (with inherited precision) or by use of tools (e.g. browse web site based on calculations or browse on basis of collection of specific parts).
Inherited precision is implicitly indicated by the precision specification from how the information is used in resolution and location. The most precise reference is part, e.g. for an image which can be located and accessed independently. Less precise references are references where calculation of other parts are needed in order to resolve and view it, e.g. page, site or subsite.

* archived-item (archived URI or identifier):
The archived item will be the URI (or identifier assigned for a resource by the archive) of the displayed archived item at hand.

A much easier way to construct PWID URNs is to use tools that construct them. Currently, there is also a prototype for a SOLR-Wayback tool (Source at https://github.com/netarchivesuite/solrwayback) [PWIDprovider], which can assist in finding the most precise reference to an archived web page. This Wayback version can provide all PWID URNs belonging to a shown page (with the page PWID URN at the top). For example, in netarkivet.dk, the archived URI for the web page http://www.susanlegetoej.dk/shop/handskedyr-siamesser-killing-8681p.html archived 2008-11-29 01:19:16 UTC, has the following parts calculated by the SOLR-Wayback tool:

```
```
urn:pwid:netarkivet.dk:2008-11-29T00:40:00Z:part:http://www.susanlegetoej.dk/images/ddcss/SK113_TopMenu_NF.css


Security and Privacy:

Security and privacy considerations are restricted to accessible web resources in web archives. Resolvers to PWID URNs will usually only be possible using the web archives’ access tools, where security and privacy are covered by these tools. In such cases security and privacy will be covered by such tools, since the information used for access has no security and privacy issues. In the cases where resolution is made around the archives’ access tools, there should be made separate analysis.

Interoperability:

This is covered by comments in the Syntax description:

* the PWID URN conforms to the URI standard defined as in [RFC3986] and the URN standard [RFC8141]

* the ‘archival-time’ of the PWID URN conforms to the UTC timestamp as described in the W3C profile of ISO 8601 [ISO8601] [W3CDTF] and is in accordance with the WARC standard ISO 28500 [ISO28500].

* the ‘archived-item’ is either an assigned identifier (the URN standard [RFC8141]) or an URI which conforms to the URI standard defined as in [RFC3986], with %-encodings of ",", ",", ",", and "," in order to conform to the URN standard [RFC8141]

Resolution:

The information in a PWID URN 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. Resolution of a PWID URN is the primary motivation of making a formal URN definition, instead of just textual representation of the for needed parts of a PWID.

Resolution (manually or automatically) is done based on the PWID parts:
Web archive identification for web archive holding referred resource
The identifier is typically the domain name for the web archive, where browsing this domain page typically will lead to description of how to access the web archive. For example, "archive.org" is the domain name leading to the Internet Archive’s interface to their online web collection, and "netarkivet.dk" is the domain name leading to the website for the Danish web archive with information about how to apply for access permission to the web collections. A future possibility is to have a registry for archive identification, with archive identifiers along with their current location on the internet. Such a registry will be needed for persistent reference to the archive, since an archive may change their location and name or archives may merge. There is work in progress to define such a registry, but no details yet.

Archived URI or identifier of archived item
If the resource is an archived URI, this URI must be used in search for or construction of location of the resource. If the resource is an identifier assigned to the resource (by the archive), it is this identifier that must be used in search for or construction of location of the resource.

Date and time associated with the archived item
The archival date and time must be used in search for or construction of location of the resource.

Precision of what is referred
The precision can either contribute to the guidance of activating tools to view the referred item e.g. browse the referred item as a page on basis of computed closest past, browse the referred item on basis of parts specified in a collection, or view the referred item as a snapshot. In the example of the snapshot, it also contains a specification of which resource to display.

In the following the different resolution techniques are explained (manual as well as via a service).

An example of a PWID URN is:


has the information:

* archive.org
Currently known identifier in form of the Internet Archive domain name for their open access web archive. If Internet Archive registered their open web archive in an IANA web archive register, this identifier could currently be "web.archive.org/web/" for Wayback resolution, or it could be "archive.org/pwid/" if a PWID interface was created as described below

* 2016-01-22T11:20:29Z
UTC date and time associated with the archived URI

* page
Clarification that the reference cover the full web page with all its inherited parts selected by the web archive

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

Resolution of this URN PWID can be deduced based on the current (2019) knowledge of Internet Archive’s open Wayback access web interface, which has the pattern:

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

Using this pattern (where only digits from the timestamp is included) it is possible manually (or automatically) to deduce the online https URI:


The same recipe can be used for other Wayback platforms for open web archives.

Another manual resolution is to find the resource by use of the specified web archive’s search interface. This will work for both open web archives and web archives with restricted access onsite.

It is also noteworthy that the information in the PWID can help in finding an alternative resource, in case the original referred resource is no longer available. The archived URI can be searched in other web archives, where the date and time can help to find the best match, e.g. via Memento (for some open web archives) or via possible coming web archive infrastructures.

If an open web archive has registered an identifier for the web archive along with the current pattern for access with <time> and <uri> - and where the latter information is updated when the pattern change - then such a register can be used to deduce
location on long term. Likewise, for web archives with restricted access, the registry will be able to provide information of where to apply for access permissions.

Regarding the precision specification, there are so far no implementations which support distinctive rendering depending on such a parameter, e.g. only providing html for an html page specified as part and the page with calculated elements if specified as page etc. Therefore, the precision specification will initially be ignored by a resolution to a Wayback interface.

A resolving service is currently available in form of code for a prototype which run at the Royal Danish Library [PWIDresolver] and is planned to be more widely available. This service currently covers both the Danish web archive (with the proper rights) and open web archives with access services based on a pattern including archive, archival time and archived URI. In other words, for open web archives it covers conversion of PWID URNs for: archive.org, archive-it.org, arquivo.pt, bibalex.org, nationalarchives.gov.uk, stanford.edu and vefsafn.is. For the Danish web archive with restricted access, the prototype works locally accessing the CDX of the library, and providing access via a local proxy to a restricted environment. The source code for this prototype is available from https://github.com/netarchivesuite/NAS-research/releases/tag/0.0.6.

Automatic access of a referenced web resource may work on the open web for open web archives or in restricted environments for the web archives with restricted access. 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 as described above

* Via web research infrastructures
  This is a future solution scenario as a web archive research infrastructure does not yet exist. However, it is a likely future scenario, as it is currently being proposed in the RESAW community [RESAW]

Documentation:

None relevant

Additional Information:
The PWID was originally suggested as a URI, based on research between a computer science researcher with knowledge of web archiving and researchers from humanity subjects (History and Literature). This resulted in the paper "Persistent Web References - Best Practices and New Suggestions" [IPRES2016] from the iPres 2016 conference. In this paper, the PWID is referred to as WPID. However, feedback was received displaying a concern that WPID was interpreted as a PID related to a PID-system, e.g. as the DOI. Although the definition of a PID does not contradict the name "WPID", there would still be a danger of confusing it with PID-systems, which is not the intension. Consequently, this suggestion names the PWID instead.

Comments on the drafted PWID URI ([DraftPwidUri]) have suggested that it should be a URN rather than a URI, which is why the PWID URN is defined here.

At the RESAW 2017 conference there were two related papers: One on referencing practices [ResawRef] and one on research data management practices [ResawColl]. These practices are also planned to be used for Danish web collections.

There has been expressed interest for the PWID URN at several occasions. There were lots of response at iPRES 2016. Especially at the RESAW 2017 conference, web researchers from digital humanities have expressed a strong interest in the PWID, since it will fill a gap and make it possible for the researchers to make the necessary references.

At iPRES 2018, the PWID URN was presented as a digital poster, which gained a lot of interest and won the "Best poster" award [IPRES2018].

A more researcher-oriented poster was presented at iDCC 2019 [IDCC2019].

Revision Information:

This is the fifth version of PWID as a URN, where remarks from the recent PWID URN reviews have been incorporated along with some minor updates. The changes includes the following:

* It is made explicit that the PWID URN is not case sensitive, except for case sensitive URIs identifying an archived item.
3. 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/corpora definitions. Also thanks to everybody who has contributed to this work with the research parts and with reviewing of this RFC.

4. References

4.1. Normative References


4.2. Informative References


Poster at 14th International Digital Curation Conference (IDCC) 2019


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


In: proceedings of the 15th International Conference on Preservation of Digital Objects (iPres) 2018, DOI: 10.17605/OSF.IO/U5W3Q


Author’s Address

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

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