Internet Engineering Task Force E. Zierau, Ed.

Internet-Draft Royal Danish Library

Intended status: Informational November 4, 2018

Expires: May 8, 2019

A Persistent Web IDentifier (PWID) URN Namespace

draft-pwid-urn-specification-04

Abstract

This document specifies a Uniform Resource Name (URN) for Persistent

Web IDentifiers to web material in web archives using the 'pwid'

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 it can be used in specifying web

collections, which then can form a persistent computational basis for

the extract of the archived collection parts. Since the parts can be

specified generally, this further allow collections to be specified

with elements from one or more web archives.

The PWID is designed for researchers and 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."

This Internet-Draft will expire on May 8, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the

document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents

(https://trustee.ietf.org/license-info) in effect on the date of

publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as

described in the Simplified BSD License.

1. Introduction

The URN PWID is a supplement to existing reference standards, where

the PWID 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, it enables

technology agnostic references to web archives in general, which can

for instance can be needed for references to web material that is

dynamic (e.g. a news site) or a specific version of a web material

(e.g. specific version of the DOI handbook).

The URN PWID is a form that can work as an algorithmic basis for

finding the resource. This also enables basis for computation of

archived web parts to a collection from one or more web archives.

Furthermore, the PWID includes information about the resource which

makes it possible to find alternative resources, in cases where the

original precise resource have 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 for being sustainable. Furthermore, it is designed to be

humanly readable and with ability to make precision of the web

archive resource covers. This design enables a PWID URN to:

o be used for technical solutions e.g. to make them resolvable

o cover references to all sorts of materials in web archives

o cover references to materials from all sort of web archives

The motivation for defining a PWID namespace is the growing challenge

of references to archived web resources, which the PWID as a URN can

assist in overcoming. The standard is needed to address web

materials meeting precision and persistency issues on par precision

in with 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 collection/

corpus. In detail the challenges are:

o 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 URN is therefore focused

on referencing archived web material in a technology agnostic way

(research documented in [IPRES2016] and [ResawRef]).

o There are many new initiatives for web archive referencing, - most

of them are centralised solutions which offers harvest and

referencing, but these cannot be used for existing materials in

web archives. Other initiatives only cover open web archives,

which does not cover material in archives with restricted access

and where there is a risk of imprecision if a resource in an

alternative archive is the result of resolving such a resource.

The PWID URN is needed in order to fill these gaps where other

techniques are not sufficient.

o There are many different requirements for construction of

collection definitions for web material besides precision and

persistency. Recent research have found that various legal and

sustainability issues leads to a need for a collection to be

defined by references to the web parts in the collection. The

PWID URN is needed in such definitions in order to fulfil these

requirements and to enable a collection to cover web materials

from more archives (research documented in [ResawColl]).

The PWID is especially useful for web material where precision is in

focus and/or there are references to materials from web archives

requiring special grants 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 about what is actually referred by

the reference (e.g. is it the page or the whole website).

Furthermore the PWID is very useful in specification of contents of 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.

Strict syntax is needed for the PWID reference in order to ensure

that it can be used for computational purposes. This is especially

relevant for automatic extraction of parts from web collection

definitions. 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, which strict syntax

can make possible. Examples of technical solutions that is enabled

are:

o resolving of a references and automatic extraction of web

collection defined by PWID URNs [ResawRef] [ResawColl]

o Resolving of a PWID reference by resolving services. As a start,

there is work on a prototype that can work for the Danish web

archive data and open web archives with standard patterns for the

current technologies. There may come different implementations

for resolving which may rely on different protocols and

application

The purpose of the PWID is also to express a web archive reference as

simple as possible and at the same time meeting requirements for

sustainability, usability and scope. Therefore, the PWID URN is

focused on only having the minimum required information to make a

precise identification of a resource in an arbitrary web archive.

Resent research have found that this is obtain by the following

information [ResawRef]:

o Identification of web archive

o Identification of source:

\* Archived URI or identifier

\* Archival timestamp

o 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:

4

Date:

2018-11-03

Registrant:

Eld Maj-Britt Olmuetz Zierau

Royal Danish Library

Soeren Kierkegaards Plads 1

1219 Copenhagen

Denmark

ph: +45 9132 4690

email: elzi@kb.dk

Purpose:

The URN PWID is a supplement to existing reference standards,

where the PWID 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 it

enables technology agnostic references to web archives in general,

which can for instance can be needed for references to web

material that is dynamic (e.g. a news site) or a specific version

of a web material (e.g. specific version of the DOI handbook).

The URN PWID is a form that can work as an algorithmic basis for

finding the resource. This also enables basis for computation of

archived web parts to a collection from one or more web archives.

Furthermore, the PWID includes information about the resource

which makes it possible to find alternative resources, in cases

where the original precise resource have become unavailable.

PWID URN is designed to be a persistent reference that is

general, global and technology agnostic in order to enhance its

chances for being sustainable. Furthermore, it is designed to be

humanly readable and with ability to make precision of the web

archive resource covers. This design enables a PWID URN to:

\* be used for technical solutions e.g. to make them resolvable

\* cover references to all sorts of materials in web archives

\* cover references to materials from all sort of web archives

The motivation for defining a PWID namespace is the growing

challenge of references to archived web resources, which the PWID

as a URN can assist in overcoming. The standard is needed to

address web materials meeting precision and persistency issues on

par precision in with 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 collection/corpus. In detail the challenges

are:

\* 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 URN is therefore focused on referencing archived web

material in a technology agnostic way (research documented in

[IPRES2016] and [ResawRef]).

\* There are many new initiatives for web archive referencing, -

most of them are centralised solutions which offers harvest and

referencing, but these cannot be used for existing materials in

web archives. Other initiatives only cover open web archives,

which does not cover material in archives with restricted

access and where there is a risk of imprecision if a resource

in an alternative archive is the result of resolving such a

resource. The PWID URN is needed in order to fill these gaps

where other techniques are not sufficient.

\* There are many different requirements for construction of

collection definitions for web material besides precision and

persistency. Recent research have found that various legal and

sustainability issues leads to a need for a collection to be

defined by references to the web parts in the collection. The

PWID URN is needed in such definitions in order to fulfil these

requirements and to enable a collection to cover web materials

from more archives (research documented in [ResawColl]).

The PWID is especially useful for web material where precision is

in focus and/or there are references to materials from web

archives requiring special grants in order to gain access. The

precision regards both regards precise reference where there can

be no doubt about that you have the correct web material as well

as precision about what is actually referred by the reference

(e.g. is it the page or the whole website)

Furthermore, the PWID is very useful in specification of contents

of 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.

Strict syntax is needed for the PWID reference in order to ensure

that it can be used for computational purposes. This is

especially relevant for automatic extraction of parts from web

collection definitions. 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, which

strict syntax can make possible. Examples of technical solutions

that is enabled are:

\* resolving of a references and automatic extraction of web

collection defined by PWID URNs [ResawRef] [ResawColl]

\* Resolving of a PWID reference by resolving services. As a

start, there is work on a prototype that can work for the

Danish web archive data and open web archives with standard

patterns for the current technologies. There may come

different implementations for resolving which may rely on

different protocols and application

The purpose of the PWID is also to express a web archive reference

as simple as possible and at the same time meeting requirements

for sustainability, usability and scope. Therefore, the PWID URN

is focused on only having the minimum required information to make

a precise identification of a resource in an arbitrary web

archive. Resent research have found that this is obtain 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 it 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

\* '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' shall

represent the timestamp that the web archive have recorded for

the referenced archived URI. The archival-time may be

specified at any of the levels 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]

\* 'precision-spec' values are not case sensitive (i.e. "PAGE" /

"PART" / "PaGe" / ... are valid values as well.)

\* '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 the reference is to an html web

element, this element can be interpreted in several ways:

\* As just one web part

Meaning the file containing the html, and precisely this file

\* A web page

Meaning that an application like Wayback shows result in a

browser, and calculates referenced web parts (display

templates, images etc.) and use these found web parts in the

result.

If the full reference only contains the PWID URN for the page,

this may mean that the archived page can change look over time,

e.g. in case that parts referred by the page did not exist at

reference time, but are harvested at a later stage, - or in

case the web archive's algorithm for calculation of the

referred web parts are changed and given a different result.

In order to make a precise reference to a picture in context of

a web page, the most precise reference will 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 result in a

browser showing the web page, - and if there are restricted

access according to the reference, the application also needs

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 an URN PWID in

order to enable the person making the above described precision in

the reference. Furthermore this precesion specification will make

it possible for resolvers to display the referred source in a way

that corresponds to the precision specification.

Especially for web materials, there can be 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, there are also web

archives that archive snapshots of the web pages for the archived

URI. A third option can be 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.

As consequence of

the above, there are following valid precision-

spec values:

\* part

the single archived web part harvested as a file from the

specified URI, e.g. a pdf, an html text, an image

\* page

the web page represented by the web page file (e.g. html)

harvested from the specified URI, where this contents 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') where it 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') where it is

possible to browse to all references in the domain specified in

the archived URI

\* collection

Representation of a collection specification, where it is up to

the web archive applications to find out 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 it is up

to the web archive applications to find out how it is rendered

(where interpretation could depends on file-suffix for the web

recording), an example is 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 does not have to be assigned by an authority, as

they are based on the information created at the time of

archiving. In other words: the PWID URNs are created

independently, but following an algorithm which ensures that the

referred item can be found if it is still available. It also has

the benefit that it includes information to look alternative

resources e.g. via Memento for some open web archives [MEMENTO] or

via possibly coming web archive infrastructures.

A PWID URN is created by finding the relevant information of the

syntax parts of the PWID on form:

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

":" archived-item

The PWID URN for an archived item in 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 Archives open web archive).

This is recommended, since browsing of this domain page

typically will lead to description of how to access the web

archive, e.g. online 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 in the same organisation, e.g. archive.org and

archive-it.org are both covered by Internet Archive. When a

registry of web archives are 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 domain netarkivet.dk)

\* archival-time (archival timestamp):

The archival time for the archived item in hand may be

displayed along with the archived item, but there are different

implementation where it is important to be aware of whether a

more precise timestamp can be found, and that it is the correct

timestamp that is used. For many Wayback implementation the

precise time can be found as part of the URI used for viewing

the archived item, e.g. in the example of

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

viewable by the Internet Archives Wayback installation, the

number 20160122112029 represents the archival time

2016-01-22T11:20:29Z. In other installations. In other

installations, the most precise time 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).

Since the archived URI can have different forms indicated by

the precision specification, this information may be used in

resolution and location.

For most imprecision types are the ones that involves

calculation, i.e. page, site or subsite. For items like an

image that have no references to calculate the precision is

best described by part, since it also tells that it is a

precise reference.

\* 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 the archived item in

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 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-

siameser-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-

29T00:41:42Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_Master\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:39:47Z:part:http://www.susanlegetoej.dk/shop/css/

print.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:06Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_Basket\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:00Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_TopMenu\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:00Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_SearchPage\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:35Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_Productmenu\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:22Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_SpaceTop\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:24Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_SpaceLeft\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:23Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_SpaceBottom\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:40:25Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_SpaceRight\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:37:23Z:part:http://www.susanlegetoej.dk/images/ddcss/

SK113\_ProductInfo\_NF.css

urn:pwid:netarkivet.dk:2008-11-

29T00:37:24Z:part:http://www.susanlegetoej.dk/Shop/js/

Variants.js

urn:pwid:netarkivet.dk:2009-03-

03T11:53:00Z:part:http://www.susanlegetoej.dk/Shop/js/Media.js

urn:pwid:netarkivet.dk:2009-03-

03T11:53:02Z:part:http://www.susanlegetoej.dk/images/design/

print.gif

urn:pwid:netarkivet.dk:2009-03-

03T11:54:19Z:part:http://www.susanlegetoej.dk/Shop/js/Scroll.js

urn:pwid:netarkivet.dk:2009-03-

03T11:54:09Z:part:http://www.susanlegetoej.dk/Shop/js/

Shop5Common.js

urn:pwid:netarkivet.dk:2006-11-

20T20:16:03Z:part:http://www.susanlegetoej.dk/images/602551.jpg

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 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 RFC

3986 [RFC3986] and the URN standard RFC 8141 [RFC8141]

\* the 'archival-time' of the PWID URN conforms 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 RFC 8141 [RFC8141]) or an URI which conforms to the

URI standard defined as in RFC 3986 [RFC3986], with %-encodings

of "[", "]", "#", and "?" in order to conform to the the URN

standard RFC 8141 [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 either an identifier where location of the

web archive can be found by looking up the identifier in a

registry, - or it is the domain name for the web archive, where

browsing this domain page typically will lead to description of

how to access the web archive, e.g. online or by applying for

access grants

\* 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 the 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:

urn:pwid:archive.org:2016-01-22T11:20:29Z:page:http://www.dr.dk

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

Based on the current (2018) knowledge of Internet Archive's open

access web interface, which has the pattern:

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

If the web archive has registered an identifier for the web

archive along with the prefix before <time> and <uri>, then this

identifier can be used to manually (or automatically) deduce the

prefix via this register

we can manually (or automatically) deduce an actual (current 2018)

access https address for Internet Archives Wayback application

(where only digits from the date is included):

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

The same recipe can be used for other Wayback platforms for open

web archives.

Another manual resolution would be 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.

It is also noteworthy that the information in the PWID can help in

finding an alternative resource, in case the original referred

resource is not available anymore. The archived URI can be

searched in other web archives, where the date and time can help

to find the best match found, e.g. via Memento (for some open web

archives) or via possibly coming web archive infrastructures.

Regarding the precision specification, there are not yet any

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.

prototype which run at the Royal Danish Library [PWIDresolver] and

is planned to be more broadly available. This service currently

covers both the Danish web archive (with the proper rights) and

open web archives with access services based on a patterns

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.

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 archive 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 do not yet exists. 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, where the suggestion

was based on research between a computer science researcher with

knowledge of web archiving and researchers from humanity subject

(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, 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 PWID instead.

The comments on the drafted PWID URI ([DraftPwidUri]) has been

that is seems to be a URN rather than a URI. Which is the reason

why it is now suggested as a URN.

At the RESAW 2017 conference there are two related papers: One on

referencing practices [ResawRef] and one on research data

management practices [ResawColl]. This practice is also planned

to be used for Danish web collections.

The interest for this new PWID has already been shown. There was

a lot of response at iPRES. Especially at the RESAW 2017

conference, web researchers from digital humanities have expressed

strong interest in the PWID, since it can fill a gap and make it

possible for them to make all the references they need to make.

Therefore, the ambition is to make the PWID URN namespace

definition a constituent part of a standard being developed in the

IETF or some other recognized standards body.

At iPRES 2018, the PWID URN was presented in a digital poster,

which had a lot of interest around it, and it won the "best

poster" award [IPRES2018]. A more researcher-oriented version of

this poster has been accepted to iDCC 2019.

Revision Information:

This is the fourth version of PWID as a URN, where remarks from

the URN PWID reviews have been incorporated. This large covers

the following:

\* It has been more clear clear that the PWID URN is a needed

supplement to existing standards (especially in Abstract and

Introduction of RFC, as well as Purpose of URN template)

\* It has been made more clear that the PWID URN also can be used

as basis for search of resources that has become unavailable

(especially in the Introduction of RFC, as well as Purpose and

Resolution sections of URN template)

\* The Introduction section of the RFC and the Purpose section of

the URN template has been aligned.

\* 'Coverage' has been renamed to 'precision' and it has been

explain in much more details (especially in the Syntax,

Assignment and Resolution sections)

\* Use of the term "ambiguity" have been rephrased in order to be

more correct

\* 'archival-time' and 'URI' have been decribed in more details

and more correctly (in the Syntax section)

\* Description of Assignment has been expanded to provived more

thorough and precise description (in the Assignment section)

\* Description of Resolution has been expanded to provived more

thorough and precise description (in the Resolution section)

\* The Interoperability descriptions have been adjusted to reflect

the descrions in the Syntax section (in the Interoperability

section)

Furthermore the Security and Privacy section has been edited in

order to become more clear, and the Additional Information section

has been extended with mentioning of the price winning iPRES 2018

poster and coming iDCC 2019 poster.

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

contributed with supplementary research concerning requirements for

web collection/corpora definitions. Also thanks to all that have

contributed to this work with the research and reviewing this RFC.

4. References

4.1. 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,

<https://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,

<https://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,

<https://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,

<https://www.rfc-editor.org/info/rfc5234>.

[RFC8141] Saint-Andre, P. and J. Klensin, "Uniform Resource Names

(URNs)", RFC 8141, DOI 10.17487/RFC8141, April 2017,

<https://www.rfc-editor.org/info/rfc8141>.

4.2. Informative References

[DOI] International DOI Foundation, "The DOI System", 2016,

<https://web.archive.org/web/20161020222635/

https:/www.doi.org/>.

urn:pwid:archive.org:2016-10-20T22:26:35:site:https://www.

doi.org/

[DraftPwidUri]

Zierau, E., "DRAFT: Scheme Specification for the pwid URI,

version 4", June 2018, <https://datatracker.ietf.org/doc/

draft-pwid-uri-specification/>.

[IPRES2016]

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

[IPRES2018]

Zierau, E., "Precise and Persistent Web Archive References

- Status, context and expected progress of the PWID",

September 2018", September 2018.

In: proceedings of the 15th International Conference on

Preservation of Digital Objects (iPres) 2018

[ISO28500]

International Organization for Standardization,

"Information and documentation -- WARC file format", 2017,

<https://www.iso.org/standard/68004.html>.

[ISO8601] International Organization for Standardization, "Data

elements and interchange formats -- Information

interchange -- Representation of dates and times", 2004,

<https://www.iso.org/standard/40874.html>.

[MEMENTO] Memento Development Group, "About the Memento Project",

January 2015, <http://mementoweb.org/about/>.

urn:pwid:archive.org:2018-11-

01T15:26:28Z:page:http://mementoweb.org/about/

[PWIDprovider]

Royal Danish Library (Netarkivet), "SolrWayback 3.1",

2018, <https://github.com/netarchivesuite/solrwayback>.

urn:pwid:archive.org:2018-06-

11T02:00:05Z:page:https://github.com/netarchivesuite/

solrwayback

[PWIDresolver]

Royal Danish Library (Netarkivet), "Date and Time Formats:

note submitted to the W3C. 15 September 1997", 2018,

<https://github.com/netarchivesuite/NAS-research/releases/

tag/0.0.6>.

urn:pwid:archive.org:2018-07-

16T06:53:51Z:page:https://github.com/netarchivesuite/NAS-

research/releases/tag/0.0.6

[RESAW] The Resaw Community, "A Research infrastructure for the

Study of Archived Web materials", 2017,

<https://web.archive.org/web/20170529113150/

http://resaw.eu/>.

pwid:archive.org:2017-05-29T11:31:50Z:site:http://resaw.eu

/

[ResawColl]

Jurik, B. and E. Zierau, "Data Management of Web archive

Research Data", 2017,

<https://archivedweb.blogs.sas.ac.uk/files/2017/06/

RESAW2017-JurikZierau-

Data\_management\_of\_web\_archive\_research\_data.pdf>.

In: proceedings of the RESAW 2017 Conference, DOI:

10.14296/resaw.0002

[ResawRef]

Nyvang, C., Kromann, T., and E. Zierau, "Capturing the Web

at Large - a Critique of Current Web Referencing

Practices", 2017,

<https://archivedweb.blogs.sas.ac.uk/files/2017/06/

RESAW2017-NyvangKromannZierau-

Capturing\_the\_web\_at\_large.pdf>.

In: proceedings of the RESAW 2017 Conference, DOI:

10.14296/resaw.0004

[W3CDTF] W3C, "Date and Time Formats: note submitted to the W3C. 15

September 1997", 1997,

<http://www.w3.org/TR/NOTE-datetime>.

W3C profile of ISO 8601 urn:pwid:archive.org:2017-04-

03T03:37:42Z:page:http://www.w3.org/TR/NOTE-datetime

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