Dear all

I finally got through all the commenting and edits needed for the all the previous mails. – Sorry for the late answer, - I just wanted to be sure to make proper work for all comments before answering.

I have uploaded a new version 4 of the PWID URN with updates that meet the comments as described in the following – taking each mail with the newest mail first.

I will send two mails after this one with the following attachments:

* Word file with difference between version 3 and version 4
* This mail with colours making it easier to read

As far as I can see, the most important issues to clear are:

* Possibility and maybe startup of registration of archive identifiers
* Agreement on content coverage

(problem with [, ], ?, # is solved with %-encoding)

A summary of changes is provided in the PWID URN RFC document in the Revision Information section.

And lastly the text for the PWID URN version 4 – only from URN template – is included in the end of this mail

Best regards, Eld

**09-09-2018 from Dale R. Worley** **worley@ariadne.com** **PWID as citation (was: Suggested PWID URN for Persistent Web IDentifiers - version 3)**

*General Answer November 2018:*Thank you Dale, this is very close to my point, which I obviously did not point out strongly enough in version 3. I have tried to make this more explicitly in the new version 4.

Thinking again about PWIDs as citations ...

One feature of a citation is that it's not just an identifier for retrieving a specific instance of a resource. If that instance is unavailable, information can be extracted from the citation that can be used to retrieve other instances of the resource. I.e., if the cited instance is in Library A, information from the citation (author, title,

etc.) can be used to locate another instance in Library B. The citation is \*transparent\* as a source of information for \*partial matching\* against the universe of information resources.

From this point of view, retrieval URLs for web archives are \*opaque\*, at least, unless one knows the specifics of how each archive constructs its URLs.

But PWID URNs are \*transparent\*; the archived URL and timestamp can be extracted from the URN algorithmicly. And those can be used to search other web archives for similar archived resources. (Presumably, other archives don't have archived resources with exactly the same timestamps.)

Answer November 2018:
Yes that is correct, - and doing so, you will be aware that you are not getting the exact referenced resource – but if the resource is truly unavailable, it will be better than nothing.

I have written this point into the RFC as well.

Conversely, if the archives are amenable, we can have straightforward resolution by having each archive register a URL prefix with IANA. Like the current DOI resolver, when a URN is prefixed with its archive's URL prefix and then fetched, the referenced resource is returned.

Answer November 2018:
My first idea (as described under resolution) was to ask web archives to make an interface for PWIDs using their domain followed by PWID like https://<archive-id>/pwid?time=<archival- time>&coverage=<coverage-spec>&item=<archived-item> as something the PWID could resolve to. - As I understand you, your suggestion is something like that, but your suggestion is much better as it is much more flexible for open archives using this kind of “Wayback interface”. However, there are challenges with the date/time, since such web archives does not interpret/resolve readable version of the time (and one of the points is to make it readable), e.g. these examples: [https://web.archive.org/web/20180306090939/https://www.dr.dk/](https://web.archive.org/web/20180306090939/https%3A//www.dr.dk/) and [https://web.archive.org/web/2018-03-06T09:09:39Z/https://www.dr.dk/](https://web.archive.org/web/2018-03-06T09%3A09%3A39Z/https%3A//www.dr.dk/) resolve in different ways, and [http://www.webarchive.org.uk/wayback/archive/2016-01-14T23:31:44Z/http://netpreserve.org/sites/default/files/attachments/2015\_IIPC-GA\_Slides\_18\_Nelson.ppt](http://www.webarchive.org.uk/wayback/archive/2016-01-14T23%3A31%3A44Z/http%3A//netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_18_Nelson.ppt) does not resolve at all while [https://www.webarchive.org.uk/wayback/archive/20160114233144/http://netpreserve.org/sites/default/files/attachments/2015\_IIPC-GA\_Slides\_18\_Nelson.ppt](https://www.webarchive.org.uk/wayback/archive/20160114233144/http%3A//netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_18_Nelson.ppt) does resolve.

Another challenge is that not all web archives and not all Wayback installations work this way. For instance the Danish installation works (with the right access grants) with URIs that points directly to the WARC files. These URIs can only be calculated using the CDX information for the archive. It is no way persistent URIs which has just been shown when the archive just converted from uncompressed WARC to compressed WARC (where filenames, offset and length was changed accordingly). That is why the prototype resolver mentioned in the version 3 works differently for Netarkivet (the Danish web archive), where it requires access to the Netarkivet CDX and calculates the internal URI for the Netarkivet Wayback installation based on the information from the PWID and the CDX.

Still it is a good idea to have this prefix as part of a registry, since it will be easier to make resolvers that can work over a long period of time without maintenance that is hard to automatically control (finding out that a prefix has changed).

Even for the open Wayback web archives, I do also think it must be possible to specify a shorter identifier e.g. UK-WA for the UK web archive which today has the prefix ‘webarchive.org.uk/wayback/archive/’. Both because it is rather long and because it contains information based on today’s technology and therefore is very likely to change when technology change. The reason is that some of the prefixes can be very long and most authors have different reason to prefer to use as short a reference as possible (number of allowed characters in a paper, restrictions by editor etc.)

It is also worth noting that it should be possible to register multiple identifiers for the same archive. It can be very hard to find identification of a web archive over time, if there is no history with different identification. I think the domain name is good because it usually leads you to a description of how to access the web archive or apply for access. However, we cannot be sure that the same domain name is uses over a period of time, and we cannot be sure of other information (name address ….) either – unless the history is there. For example, the Netarkivet has been a virtual organisation until recently, since the responsibility was shared between to libraries. Now the two libraries are merged, and although it is still called Netarkivet with domain netarkivet.dk, I would personally not be surprised if we have a different name in 10 years, e.g. the Danish web archive within the domain of the Royal Danish Library (kb.dk).

Could it be that I request an IANA registry for web archives, ask around for the registration of some of the biggest web archives and then reference it from the RFC, and then mention the domain as an alternative option in case the web archive used is not registered?

My suggestion would be that the following can be registered (with invented examples for both of Internet archives web archives, the UK web archive and Netarkivet assuming it will change name and domain in 2025):

Global identifier: IA
name: 'Internet Archive' name-start-year: '1996' name-end-year: ''
domain: 'archive.org' domain-start-year: '1996' domain-end-year: ''
prefix: 'web.archive.org/web/' prefix-start-year: '1996' prefix-end-year: ''

Global identifier: IA-IT
name: 'Archive-It at Internet Archive' name-start-year: '2006' name-end-year: ''
domain: 'archive-it.org' domain-start-year: '2006' domain-end-year: ''
prefix: 'wayback.archive-it.org/all/ ' prefix-start-year: '2006' prefix-end-year: ''

Global identifier: UKWA
name: 'UK Web Archive’ name-start-year: ‘2004’ name-end-year: ''
domain: 'webarchive.org.uk' domain-start-year: '2004' domain-end-year: ''
prefix: 'www.webarchive.org.uk/wayback/archive/' prefix-start-year: '2004' prefix-end-year: ''

Global identifier: DKWA
name: 'Netarkivet' name-start-year: '2005' name-end-year: '2025'
domain: 'netarkivet.dk' domain-start-year: '2005' domain-end-year: '2025'
name: 'Danish Web archive' name-start-year: '2025' name-end-year: ''
domain: 'kb.dk' domain-start-year: '2025' domain-end-year: ''

In PWIDs I would say that both the identifier and the domain will be valid identifiers (alternative identifiers). The only way I can think of where this could be a problem would be that a web archive changes their domain and another web archive takes over the old domain name – I must admit that I think that this is more a theoretical than realistic scenario. The argument for allowing the use of the domain name is that people can produce PWIDs without knowing/visiting the registry, and that web archives that have not registered yet can be referred.

I have made changes in the new version of the PWID RFC in this direction, but it is of course missing references to the still not existing registry. For the same reason it does not mention a possibility of registration of an identifier along with a prefix, if that is the way forward.

And I think this combination of features is something new -- PWIDs could be \*algorithmicly resolvable\* (using their specified archives), and yet also be \*citations\* (transparently providing enough information that a human can find similar resources in other archives).

Answer November 2018:
Yes :-)

Dale

**08-09-2018 Dale R. Worley** **worley@ariadne.com** **Re: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

Re-reading this thread, I think Henry's comments are significant in regard to how this proposal will interact with actual archive use. I don't know enough about archiving to speak well to that, but there are significant technical points that seem to be connect with or parallel Henry's archiving considerations.

One critical element is the archive identifier. There seem to be two cases regarding what is intended.

One case is where the PWID URNs are constructed by archive organization itself. Thus, the archive organization can choose the archive identifier, and presumably that will be a domain name that it controls.

That is not a perfect solution (we've had trouble with it in other URN proposals), but it's clear. But in that situation, the archive is probably already minting URLs that can be used to retrieve archived resources, so it's not clear that there is a large benefit to be gained.

Answer November 2018:

It is very useful for the archive, especially for collection documentation. As discussed in the [ResawColl] paper, there are cases where the reference is taken from the CDX. However, since CDX is not a stable format, this is not a persistent reference. Furthermore, the CDX has a lot more information than needed.

The other case is where the URNs are constructed by third parties, that is, neither the archive nor the consumer of the URN. As Henry says, this usage has the nature of a citation. What is unclear in this case is who selects the archive identifier of the archive.

One possibility in this case is that it is expected that there will be only a small number of archives, and there's an IANA registry of archive identifiers, and any person is allowed to register an identifier for any archive. Presumably, expert review could be used to keep the situation under control. This is alluded to by:

 On long term, there should be created a registry that keeps track
 of identifiers of archives over time, since they are likely to
 change names, merge etc. when taking about a 100 year period.

Answer November 2018:
Se comments to mail above (09-09-2018 from Dale R. Worley)

The proposal generally assumes that an archived resource can be identified using just:

- the archive identifier
- the URL whose contents were archived
- the time at which the contents were archived

Embedding the URL into the URN presents syntactic problems. The characters [ and ] can be used in the host-address part of a URL, and ?

is used for queries. The latter I consider to be particularly important, as web links can often include query-parts. This problem needs to be solved, as a syntax that "covers about 80-95% of all cases"

is a syntax that doesn't suffice for the problem at hand.

I don't see a need to worry about fragment-parts of archived URLs, as the fragment structure is inherently embedded in the resource retrieved via the URL. So an archive can archive the URL-without-fragment, the URN can reference that URL, and user can attach the required fragment-part to the URN to specify the desired fragment.

Answer November 2018:
I agree that it seems to be half a solution, although it is a huge step forward to have 80-90%. I therefore propose that occurrences of [, ], ? and # are %-encoded

I have added this to version 4 syntax section (in explanation of the URI token)

There is some lack of clarity about coverage-spec. If the "coverage" is

\*part\* of the archived resource, then it is, or should be, a fragment, and that can be specified by appending a fragment-part to the URN. If the coverage is metadata about the resource, it seems to be undefined what forms the metadata resource could take. But some coverage values seem to suggest that the referenced information is a set of resources which contains as a member the resource designated by the recorded URL.

This concept gets very interesting indeed. There doesn't seem to be any defined resource type for "a web site", or even "all the files needed to display a web page". Also, it's not clear what the archival-time means in this context, since presumably the archive need not contain archived copies of everything in the aggregate that were all made at the specified archival-time.

Answer November 2018:
There seems to be some misunderstanding here, - and I certainly need to be more clear about it. I have therefore added the below quoted text in the syntax section (where coverage is exchanged with precision as explained later).

 “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”

Dale

**21-07-2018 from Dale R. Worley** **worley@ariadne.com** **Re: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

with Eld comments from mail

**RE: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3
10-08-2018 Eld Zierau** **elzi@kb.dk**

The draft seems to me to be prolix, giving a lot of discussion of details that are not reflected in the URN proposal itself. A careful editing will probably fix this.

A major component of pwid-urn is archive-id. My assumption is that the archive-id is the top-level component and defines what abstract "archive" the URN is a reference into. And that the "archive" defines the exact interpretation of archival-tine, coverage-spec, and archived-item. However, in order for PWID URNs to be unambiguous, it must be unambiguous what "archive" a given archive-id refers to. The draft suggests "it is recommended to use the web domain as the identifier for the web archive".

What if an institution decides to create PWID URNs that use an archive-id which is a domain name that the institution does not own? As written, the draft does not forbid \*me\* from creating PWID URNs with archive-id "netarkivet.dk".

**-- Eld (10-08-2018): I think there is a misunderstanding here – it is not the archive that chooses the archive name, it is the creator of the reference, - and that is why it is recommended to use the domain name. It would be best to have a formal registry, but I have to start somewhere, - and with the domain name it is possible to find a reference to any existing web archive (restricted access or not). Within the next 5-10 years, I think there will be evidence enough (e.g. from web archives) to reconstruct a registry in case there are web archives changing domains within that period. I will definitely work on having such a registry formerly established.**

**-- It could be that it should be stronger here and instead say: ‘must use’ – instead of ‘recommend to use’**

Answer November 2018:
1)
I have avoided the use of the “unambiguity”, in the following ways:
Reformulated the “unambiguity” part to address the fact that “the PWID URN represents this information in a well-defined way that enable technical solutions to interpret the URN” (end of introduction)

Removed the word under the purpose section of the template

2)
See also the comments about archive-id above (with later mails)

In regard to archival-time, the draft states "The 'archival-time' [...] can therefore be specified at any of the levels of granularity as described in [W3CDTF]." However, W3CDTF gives a number of formats for time designations, and calls those formats "granularities", whereas this draft allows only one format. I suspect that a more careful description of the meaning of the quoted sentence would fix this problem.

**-- Eld (10-08-2018): absolutely**

Answer November 2018:
I have removed the detailed syntax for archival-time and instead made a description which is a reformulation, so it is fully aligned with the WARC-date description in the WARC standard ISO28500:

“The WARC-date is a UTC timestamp as described in the W3C profile of ISO8601 [W3CDTF], for example YYYY-MM-DDThh:mm:ssZ. The timestamp shall represent the instant that data capture for record creation began. Multiple records written as part of a single capture event (see section 5.7) shall use the same WARC-date, even though the times of their writing will not be exactly synchronized.

WARC-Date may be specified at any of the levels of granularity described in [W3CDTF]. If WARC-Date includes a decimal fraction of a second, the decimal fraction shall have a minimum of 1 digit and a maximum of 9 digits. WARC-Date should be specified with as much precision as is accurately known. This document recommends no particular algorithm for access software to choose a record by date when an exact match is not available.”

The reformulation to be found in the draft for PWID URN version 4 is:

“'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].”

The syntax assumes that knowing archive-id, archival-time, coverage-spec, and archived-item (the archived URI) will always be sufficient to specify exactly the resource in question. However, it seems likely that special situations will arise when two distinct resources will have attributes that are similar enough that the distinction between them is not easily expressed in terms of archive-id, archival-time, coverage-spec, or archived URI. In this case, additional information is needed to create an exact reference, and the URN syntax needs to have an expansion facility to allow this.

**-- Eld (10-08-2018): The very nature of web archives is based on information archived URI and archival time – per archive – if there are such cases the web archive would have a problem with their indexes, as these are based exactly on this information. This is also why it is this information that is chosen and why it is important to specify the archive as well. The content-coverage is another story, which I will go more into in my response to Henry S. Thompson’s mail**

Answer November 2018:
Just adding to previous answer – the needed information to reference was found as a result of the research that is the basis for the suggestion of the PWID URN. If we try to imagine cases where this information is not present, we can firstly rule out identification of the web archive (since the PWID is only concerned with precise references to archived web, i.e. a reference that has been verified to exist in a web archive). Secondly, the PWID is concerned with archived URIs. All Wayback installations have this information. There are certain installations where neither URI nor archival date is easily accessible, which may mean that it might currently be difficult to construct a PWID. However, it is unlikely that the information is not recorded at all and therefore cannot be made available at some stage. Thinking about it, I would say that in such a case, there would be strong reasons to refrain from relying on references to a web archive, since archived URI and archival time is basic information to be recorded in an archive.

archived-item can be a URI, and a URI can contain the characters '?', '#', '[', and ']', among others. However, those 4 characters may not appear in the NSS part of a URN. Note that the first 2 of these characters are used in URIs only to introduce the "query" and "fragment" parts, but as the draft is written, the archived-item URI is not restricted to not having a query or fragment part.

**-- Eld (10-08-2018): that is a problem if the archived URI does have fragments**

Answer November 2018:
See comments and changes with %-encoding above (with later mails)

Dale

**11-08-2018 Dale R. Worley** **worley@ariadne.com** **Re: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

Including Eld comments from mail

**13-08-2018 Eld Zierau** **elzi@kb.dk** **RE: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

From: Dale R. Worley <worley@ariadne.com>
Sent: Saturday, August 11, 2018 3:55 AM
To: Eld Zierau <elzi@kb.dk>
Cc: juha.hakala@helsinki.fi; ht@inf.ed.ac.uk; urn@ietf.org
Subject: Re: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3

Eld Zierau <elzi@kb.dk> writes:
> Dale R. Worley worley@ariadne.com writes:
> > A major component of pwid-urn is archive-id. My assumption is that
> > the archive-id is the top-level component and defines what abstract
> > "archive" the URN is a reference into. And that the "archive"
> > defines the exact interpretation of archival-tine, coverage-spec,
> > and archived-item. However, in order for PWID URNs to be
> > unambiguous, it must be unambiguous what "archive" a given
> > archive-id refers to. The draft suggests "it is recommended to use
> > the web domain as the identifier for the web archive".
> >
> > What if an institution decides to create PWID URNs that use an
> > archive-id which is a domain name that the institution does not own?
> > As written, the draft does not forbid \*me\* from creating PWID URNs
> > with archive-id "netarkivet.dk".
>
> -- Eld: I think there is a misunderstanding here - it is not the
> archive that chooses the archive name, it is the creator of the
> reference, - and that is why it is recommended to use the domain name.

If you are forced to say "it is not the archive that chooses the archive name", then I suggest that the "archive name" is not actually the name \*of the archive\*, but an identifier of some other entity.

-- Eld (**13-08-2018)**: If it says '**archive name**' anywhere, it should be corrected - it is an identifier we are taking about.

Answer November 2018:
I have not been able to find any occurrences of archive name, except from where it states that the archive name is not precise enough, - therefore I have not made any changes regarding this comment

When you say "the domain name" here, \*of what\* is it the domain name?
Do you mean "the archive's domain name", or "the domain name of the creator of the reference"?

-- Eld (**13-08-2018)**: the domain name of the archive - it is to identify in which archive you can find the resource in - I actually realized that it should say top level domain name of the web archive.

Answer November 2018:
Sorry, it is certainly not top level domain, - it is the domain name (and not subdomain name – like web.archive.org is a subdomain to archive.org)

I have made it explicit that it is the domain of the web archive supplied with an example in the Assignment section.

> It would be best to have a formal registry, but I have to start
> somewhere, - and with the domain name it is possible to find a
> reference to any existing web archive (restricted access or not).
> Within the next 5-10 years, I think there will be evidence enough
> (e.g. from web archives) to reconstruct a registry in case there are
> web archives changing domains within that period. I will definitely
> work on having such a registry formerly established.

You seem to be describing how you expect these URNs to be processed, but it is not clear to me what the process is. Can you describe how this process works, starting with "I am looking at a PWID URN" and continuing to the point where "I have discovered the particular archive in question". (I assume that after that point, looking up the resource in the archive's index is straightforward, at least conceptually.)

-- Eld ((**13-08-2018)**): It is indirectly described under resolution, - but it is a good point that there should be a construction as well. This would be:
-- PWID URNs can be constructed in different ways. In case a single reference is wanted, it can be created by having/finding/creating the wanted archived resource and then construct the PWID URN from that. Or it can be created for a collection by using tooling producing PWIDs.
-- In the first case, there can be three situations where the creator of the reference:
-- \* already has the resource from a web archive that the creator wants to reference
-- \* finds resource from a web archive by using tools like Memento and choses the exact web resource to reference from a specific archive
-- \* creates a resource by asking a web archive/citation service to archive a live URI, and afterwards verifies the reference in the web archive
-- In these three cases the PWID URN is constructed by filling out the PWID URN pattern
-- urn:pwid:<archive-id>:<archival-time>:<coverage-spec>:<archived-uri>
-- By setting
-- The <archive-id> to the top level domain of the web archive holding the reference. For open web archives like Internet Archive’s publicly available web archive collection (with current wayback on <https://web.archive.org/>) the top level domain is archive.org. For web archives with restricted access, choose the top level domain for where the web archive is publicly described, e.g. the Danish web archive is described on <http://netarkivet.dk/in-english/> which has the top level domain netarkivet.dk.
-- The <archival-time> (on the form described in the syntax and given as UTC time) to the archival time for the selected web archive resource. In some wayback solutions, this archival time can be found in the online URI for the resource, e.g. in the online URL for an archived version of <https://tools.ietf.org/html/rfc3986>: [https://web.archive.org/web/20180418152929/https://tools.ietf.org/html/rfc3986](https://web.archive.org/web/20180418152929/https%3A/tools.ietf.org/html/rfc3986), the archival time is represented in the long number 20180418152929, where year is the first four digits, month the next two digits etc. In other Wayback solutions or web archive interfaces the archival date is given as part of the metadata for the web resource.
-- The <coverage-spec> to what the creator of the reference wants the reference to cover, e.g. page or subsite.

The <archived-uri> to the URI that is archived, which means what the web archive has registered as the archived URI. For URIs that still exists online, the online URI and the archive URI will usually be the same. However, there are case (e.g. for redirects) where they differ. For instance for archive.is archived in archive.org on 2018-01-06T14:08:01 UTC is archived wiith the archived URI: <http://archive.is:80>.

-- An example of constructing an PWID URI for [https://web.archive.org/web/20180418152929/https://tools.ietf.org/html/rfc3986](https://web.archive.org/web/20180418152929/https%3A/tools.ietf.org/html/rfc3986) would therefore be
-- <archive-id> = archive.org (to level domain of web.archive.org)
-- <archival-time> = 2018-04-18T15:29:29Z (deduced from 20180418152929 which is always UTC time)
-- <coverage-spec> = webpage (depends on style sheets references. In case the creator of the reference wants to indicate that the contents is covered fully on this reference without dependencies to calculation of dependencies, he/she could choose to say part, - although this may result in a resolution of the html depending on the resolution tool)
-- <archived-uri> = <https://tools.ietf.org/html/rfc3986>

-- Which results in the PWID URN:
-- urn:pwid:archive.org:2018-04-18T15:29:29Z:webpage:https://tools.ietf.org/html/rfc3986

Answer November 2018:
I have extended the descriptions of assignment and resolution in a way that hopefully make this much more clear.
> > archived-item can be a URI, and a URI can contain the characters
> > '?', '#', '[', and ']', among others. However, those 4 characters
> > may not appear in the NSS part of a URN. Note that the first 2 of
> > these characters are used in URIs only to introduce the "query" and
> > "fragment" parts, but as the draft is written, the archived-item URI
> > is not restricted to not having a query or fragment part.

> -- Eld: that is a problem if the archived URI does have fragments

Yes, it is a problem. But your specification allows the archived URI to have fragments, because it refers to URIs but does not forbid them to have fragment parts. How do you propose to resolve this contradiction?

-- Eld (**13-08-2018)**: I will certainly make the reader aware of it, e.g. by writing:
-- "It should be noted that a valid URN is not allowed to contain the characters ?', '#', '[', and ']'. Therefore, PWID URNs for archived URIs cannot be constructed as legal URNs when following the instructions given."
-- Eld (**13-08-2018)**: I would like to discuss different alternative:
-- 1: we can just leave it like that - which means that that valid PWID URNs cannot cover these cases. - In all cases the PWID URN will be a huge step forward, and cover about 80-95% of all cases. In this case, it will be up to the users if they construct them anyway - just like the web archives online URIs are not valid URIs since they contain "//" as part of the path.
-- 2: we can consider encoding of these characters (which will complicate things a bit)

Answer November 2018:
I missed to put in the 3. Solution which is the one with %-encoding. Please see comments and changes with %-encoding above (with later mails).

Dale

**27-07-2018 Henry S. Thompson** **ht@inf.ed.ac.uk** **Re: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

Including Juha Hakala’s comments from mail

**27-07-2018 Hakala, Juha E** **juha.hakala@helsinki.fi** **RE: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**

And Eld comments from mail

**10-08-2018 Eld Zierau** **elzi@kb.dk** **RE: [urn] Suggested PWID URN for Persistent Web IDentifiers - version 3**
Eld Zierau writes:
> Please note that I have updated the suggested PWID URN for Persistent
> Web Identifiers with the following changes:
>
> ...
>
> It is available as version 3 from
> <https://datatracker.ietf.org/doc/draft-pwid-urn-specification/>, and it
> is attached to this email as a pdf.

[Before my comments, please note that I don't believe it's best practice to attach drafts in mail to this list -- your original was classified as spam as a result of the attachment, and I only found out about it as a result of Juha's mail.]

**-- Eld (10-08-2018): Got that**

Much of what follows is a critique of the requirements analysis and the presumed use cases, rather than the specifics of the namespace itself, and the overall conclusion is negative.

This is a bit unusual, in that I wouldn't normally bother with an extended critique whose TL;DR would be "It doesn't seem likely to me that this will solve anyone's problems". After all, if it doesn't solve anyone's problems, no-one will use it, let the 'market' decide, right?

This might be so, for a narrowly focused namespace targeting a specialist community. But this proposal has \_very\_ large-scale use in view.

**-- Eld (10-08-2018): Yes it is aimed big, and there is already a big interest for it – and there are several reasons for this.**

**-- Let me start by pointing out that I have a background as a computer scientist (from 1989) with a PhD in digital preservation from 2011. With this background, four years ago, I would probably raise some of the same questions.**

**-- The background for the PWID is a research project, which I undertook with a historian and a literary scientist, and they have much harder requirements for references than what I have been used to in my field. It is this difference that has made it clear to me that there is a lack of ways to reference sources in web archives for digital humanist working with web archives. Please note that the only way to make such a reference today is via URLs, which only works for online web archives. A lot of European web archives are not online, and we are unable to employ such references. At some point, the Danish web archive recommended the ‘URL’ used in a protected environment addressing the files directly. These references are now useless after we have compressed all the data. Please also note that there is nothing sustainable about a web archive URL – as noted in the draft, the paths are in many cases tools dependent and are therefore very likely to change in 5-10 years.**

**-- Especially for the humanities there is great focus on the sustainability of references. It must be evident that all referenced sources can be retrieved later on in order to re-evaluate the findings (for more than a 5-10 year horizon). As many studies shows (referenced in the draft) there is a great link rot just over a short number of years. My personal experience is that 15 % of the referenced links in my PhD (from 2011) are now dead. What we found in our studies was that within the literary field, there were quite a few works that did not reference the web at all – we suspect that this could point at a tendency that they do not want to use the insecure web references - which in practice can mean that the internet might not be the subject of relevant research.**

**-- A present need for the URN PWID is for definition of web corpora. A definition of a corpora (web collection) is for example the European initiative to select a part of the archived web for each year – cleaned for duplicates etc. – in order to have a basis for statistical measurements. In order to be able to do recalculation of the measurements and look into the basis for the measurements, it will be required that you somehow preserve a representation of it. It will be too expensive to preserve the data separately because of the volume, therefore it must be references that is saved. There have been ideas about saving CDX indexes, but these are not standard either. Therefore, specification of PWID URNs as references is a good choice.**

**-- This leads to the use of PWID as a general form, if you have your resource archived or find the resource in an archive: Either by getting a possibility to refer to the large volumes of material in European web archives (with restricted access) or by getting a sustainable technology agnostic way to address resources that have been archived for the purpose of creating a persistent reference.**

**-- I hope I with this have made it clear that this is NOT just narrowly focused namespace, but will be a huge help especially for any researcher that wants to make acceptable research in web archive materials.**

Answer November 2018:
See below

\_And\_, the legitimacy of discourse around various aspects of 'persistent' reference has been so compromised by flawed analyses and failed solutions that there is actually real reputational risk for the urn infrastructure itself from yet another example, which I'm afraid I think this is, particularly given its very high aims.

**-- Eld (10-08-2018): I hope that the above gives some clarification on this, - and about the infrastructure: this only relies on the definition, - anyone can build resolvers at any point of time (as long as we have the history of domains, which I will come back to). The reason is that the PWID is based on the very basic elements from registering a “located” thing in an archive: the archive, the location, and the time of archiving. Therefore, with the right indexes and grants, a resource can be found based on this information.**

**-- I suspect that I may have confused things by both talking about the textual and the technical URN PWID. It was meant as a way to explain this, but that can be changed if it is confusing things instead.**

Answer November 2018:
See below

So, I'm sorry but for these reasons, and the many specific problems discussed below, I don't personally see a future for this document.

Answer November 2018:
As written earlier, there is more evidence of interest for this PWID URN apart from the above listed evidence:

* The PWID was presented on a poster at iPRES 2018 and won the “best poster” award
* A poster about the PWID has been accepted for iDCC 2019 – although I did write to them that a similar poster had been at iPRES
* The Hungarian web archive has adjusted the SOLR Wayback installation to produce PWIDs for their archived material.

And please note that this is supplementary to existing solutions, - BUT the only solution for references to web archives with restricted access, which covers a lot of European web archives.

I have reformulated the introduction to be more precise about it being supplementary and needed.

============

Specific comments, by section

\*Purpose\*

As noted above, the goal stated in the open paragraph is very ambitious.

The only parts of the argument that follows that I can make sense of, as to why the proposed scheme is needed, consist of two mistaken and/or unsupported high-level claims:

 1) Citations of web-hosted resources should refer to archives, not to the resources themselves;

**-- Eld (10-08-2018): I have included several references to conference papers in the draft to support this claim, - I will of course gladly extend this if it is needed**

Answer November 2018:
See below

 2) Citation of archived material needs help to be sufficiently 'precise'.

**-- Eld (10-08-2018): There are also references to support this – mainly from the research group that I refer to above - but in these papers there are also other references - I will of course also gladly extend this if it is needed**

Answer November 2018:
I think these comment should be covered by the precision in the introduction of the PWID

The first is at best contentious as a general claim, and even in cases where an archive reference is preferred, why packaging it in a PWID urn will add value is not explained. After all, if I know that what I want to cite is available in some archive, it's because I've found it there, in which case I already \_know\_ the URI I can use to retrieve it. More on this when below when we come to the question of uniqueness.

**-- Eld (10-08-2018): This is due to the fact that an URI to a web archive is not persistent – and only existing for online web archives. I thought I made it clear, but I can of course elaborate on this as well.**

**Juha (27-07-2018): ISO TC 46/SC 9 is currently revising ISO 690 (Guidelines for bibliographic references and citations to information resources). One of the challenges is how to deal with reference rot, which consists of link rot (404) and content drift (retrieved documents change). A large percentage of all Internet resources suffer from these problems.**

**Eld has been most helpful in writing the Annex that discusses the usage of Web archives in detail, and the current version is based on what she has written, with two important exceptions:**

**- I don't think that referring just to a copy of the cited resource in a Web archive is always the best solution. Archive reference only must be used when the resource is dynamic (say, https://www.ed.ac.uk/) and a specific version is cited. But when the page is supposed to be persistent (e.g. https://www.w3.org/Provider/Style/URI) the added value provided by the archive link may be modest, because the archived copy may not always be exactly the same, and web archives themselves may not last forever. The current draft of ISO 690 (which is still only a working draft) does promote the use of web archives, but not as the only solution. And if web archives are used, utilization of Memento is something that should be considered since it provides some level of protection to disappearance of web archives.**

**- the current ISO 690 draft does not mention PWID. There are three reasons for this. First, and most importantly, PWID does not (yet) have a registered URN namespace. An ISO standard cannot refer to PWIDs as long as namespace registration has not been completed. And IMO it is not obvious that it will be completed. Second, every persistent identifier that will be included in the examples or elsewhere in the standard must be actionable (expressed, for the time being, as HTTP URI). So all DOIs that were in the form doi:<DOI string> have been converted into https://doi.org/<DOI string>. This will not possible with PWIDs, since only the Danish web archive is planning to implement PWIDs, in an archive which has strict access limitations.**

**When you consider what to do with PWID, keep in mind that as a national library the Royal Danish Library can always use the NBN namespace for items in the web archive. For instance, something like urn:nbn:pwid:<pwid> can be constructed, as long as the <pwid> string conforms to the syntactical requirements of the NBN namespace.**

Answer November 2018:
Concerning when web archive references are ‘best’, I have added the point about the dynamic part or versions in the introduction(s).

Concerning ISO 690, I have deleted all references to this in order to avoid any kind of confusion.

Concerning potential use of the PWID URN, I can tell that the Hungarian web archive are now having installations of the SOLR Wayback (<http://193.6.201.202/solrwayback>) – and they have adjusted it to produce URN PWIDs for pages in their archive with domain mia.oszk.hu (see for example <http://193.6.201.202/solrwayback/services/generatepwid?source_file_path=/var/local/wct/wayback/store/WEB-20180118104816162-00000-6875~webadmin.oszk.hu~8443.ver1.warc.gz&offset=5140225> for page [http://193.6.201.202/solrwayback/services/web/20180118114834/http://gyulavara.hu/indexen.html](http://193.6.201.202/solrwayback/services/web/20180118114834/http%3A//gyulavara.hu/indexen.html) – choose “view source” if it looks strange in the browser you have chosen)

The second surely applies to any kind of citation at all, and seems to depend on a confusion between web identity, media type and the ontology of citable things and loci associated with them -- all of FRBR lurks just around the corner here.

**-- Eld (10-08-2018): The precision is needed because references to web materials are very imprecise. Especially within the humanities, you need to specify what you refer to and thereby also what the author has cheked that the specific reference points to. There is a big difference between referencing a web page or a website or just the html for the web page. The reason that this is part of the PWID URN is to have all the information there and to make it possible (later on) for the web archives to supply services that correspond to the reference**

Answer November 2018:
Just adding – regarding FRBR – the PWID URN provides means to be able to make more precision, because web reference or web archive references today – where in fact a URI can be interpreted in different ways. We could probably have long discussion about FRBR in this context, - but I do not think it would add anything to this specific RFC, - especially because FRBR was intended for the physical world, which does not always translate well into the digital domain in general, and for web materials in particular.

Just because some representation is hosted on the Web does not make it easier (or harder) than it ever was to make clear in a citation what part/aspect/property of what you \_name\_ in a citation is what you actual are \_referring to\_. More on this below when we come to the 'coverage-spec'.

**Juha (27-07-2018): trying to express in PWID coverage-spec which part of the of the resource has been cited may be** **more difficult that providing this information elsewhere in the citation. ISO 690 draft specifies such an alternative approach. And even if PWID really has to be used for this purpose, usage of f-, r- and q-components may also be considered as an alternative.**

**As Henry says below, coverage-spec would need a lot more alternatives to be truly useful. Coverage-spec "other" does not add much value, and it would be necessary to use it for every image, video, etc. And if there are several images on the page, how to indicate the right one (except with f-component)?**

Answer November 2018:
I have realized that the purpose of the coverage-spec has far from been explained well enough. And I have realized that it is easily confused with the purpose of a file-extension, - which is not intended.

I have therefore renamed it to ‘precision-specification’ and made a more thorough explanation in the version 4 syntax section (inserted text as provided in answer to comment about coverage-spec in mail from Dale **08-09-2018 above**).This will also explain the case with an image in the context of a web page.

------------

One more specific issue with this section

 "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)"

There's nothing in the proposal which follows to back up the "you have the correct web material" claim, and the subsequent inventory of 8 'type[s] of archived item[s]' is clearly not sufficient to unambiguously determine "what is actually referred [to] by the reference".

**-- Eld (10-08-2018): I just realize that there is a context missing and precision is needed. It relates to the way Memento argues for URL and date being sufficient information in order to identify a web resource, i.e. that it does not matter which archive you get the resource from as long as it is approximately from that time. That should definitely be more clear. I have numerous examples of differences in archived material of the same URL around the same time, where a reference can make sense in one case but not the other (typically from often updated news sites, or due to bad harvest settings or warnings requiring additional information before entering the site). A way to make the precision could be to say that it is precise about which archive it was evaluated in.**

**-- I agree that you cannot be sure of your reference as long as it is not entirely based on parts, since it will be up to e.g. Wayback software to determine each of the parts (usually the closest in time, which may change if included materials from other archives or harvest parts that did not originally exist in the archive). I gladly include description of this as well.**

Answer November 2018:
I have reformulated in order to make more clear what is meant

From

"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)"

to

"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)."

\*Syntax\*

I see no problems in the ABNF.

------------

The definitions of the 'coverage-spec' values are hopelessly underspecified, and heterogenous. The 'part'/'page' distinction is particularly unclear/non-obvious/ambiguous/medai-type dependent.

**-- Eld (10-08-2018): hopefully it is more clear from the above. – the point is to make it possible to present the resource as reference - or for a user of the reference to make the right choices in order to extract the referred information. This means in case of ‘part’ for html, asp, jpg … it will be the actual source of the archived file, in case of web page it will be the “Wayback” software version (and actually the same for subsite and site). Some archives have possibility for choosing between snapshots or harvested web files, which is where the snapshot comes in. A comment from the iPRES 2013 conference was that it was missing web recordings, which is why recording is in, and finally there is a need for addressing collection (corpuses) also in order to make a hierarchi of collections (e.g. each year of the Danish example corpuses, another containing those years, a third containing the Danish years and some years from UK web archive etc.). if it is useful I gladly include a modified version of this explanation also.**

Answer November 2018:
Please see comments above (two comments up – after Juha comment)

\*Assignment\*

The initial claim here and its subsequent gloss taken together don't hold up:

 "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 that itself guarantees uniqueness."

On a strict reading of "uniqueness" (i.e. a one-to-one relation between items and PWIDs), this amounts to a claim that \_any\_ 3rd party considering \_any\_ item in \_any\_ archive will always construct the \_same\_ PWID. This reading is obviously false: the presence of the the two "+( unreserved )" expansions in the ABNF amount to a \_guarantee\_ that there will be multiple distinct PWIDs for the same item.

**-- Eld (10-08-2018): I never meant to say that it was a two-ways uniqueness, and I can see it can be read as that, so that should be clarified. For example, if internet archive exchange their domain name with ai.org then future PWIDs with ai.org as archive identifier will point to the same as archive.org for the same archived items (which covers the first "+( unreserved )". There may also be rare cases with the other one, which represents a generalization (to include e.g. collections specified by an identifier) – and yes in some cases an archive have more identifiers for the same thing. But as you say there should be no problem with that.**

**-- If the assignment is confusing I can also expand that, to clarify that it is the basic metadata for archiving (archive, got from location, at time) combined with precision of what is referenced.**

Answer November 2018:
I believe this comment is met by the expansion of the explanation in section ‘Assignment’ (avoiding use of the word unique) and the additional description of item identifier and archive identifier.

RFC8141 does not require one-to-one relations between URNs, even URNs within the same namespace. But it does require, in the case of "URNs . . . created independently" that they be created "following an algorithm that itself guarantees uniqueness". The underspecification of the 'coverage-spec' already alluded to above makes independent \_and\_ inconsistently understood coinages of identical PWIDs not just possible but likely. Consider for example

urn:pwid:web.archive.org:2018-01-01T17:03:53Z:part:http://www.ltg.ed.ac.uk/~ht/

One person might coin this to mean the text/html character sequence which was served from <http://www.ltg.ed.ac.uk/~ht/> (my homepage at the University of Edinburgh) at the beginning of 2018, whereas someone else might coin it to mean the text/html character sequence you get from the Web Archive today if you do an http GET on [https://web.archive.org/web/20180101170353/http://www.ltg.ed.ac.uk/~ht/](https://web.archive.org/web/20180101170353/http%3A/www.ltg.ed.ac.uk/~ht/),

which are different in important ways.

**-- Eld: This is only about archived materials, with the limitations that web archiving has. It is only meant to mean the latter case, which can be made more precise, of course**

Answer November 2018:
As stated in a previous comment, the PWID URN is only for archived web, so it is uniquely identifying the latter case.

I have adjusted description to make this clear

In general, it seems that knowing whether to use 'part' or 'page' would depend on a detailed understanding of the media type of the retrieved representation, which the average creator of a citation is unlikely to have.

**-- Eld (10-08-2018): This referencing mechanism is meant for users that verify their source before referring to it, - it is already part of standards to make clear whether it is page or site, so I would not see any problems with that. I can expand the explanation of part with several examples to make clear that that should not be an option if you mean page or site. For the rest I would assume that you would know if you were using it.**

**-- It is also meant as a future basis for making tools to respect and provide what the reference is covering – so you do not need to know that you have to select snapshot, - or use a get for requesting the source of a page, Also to have tools for corpora where can be made clear when a reference is to a subcorpora. The inspiration for this way of doing it came from the Internet Archives use of function e.g. the Identity Wayback function which is called by placing ‘id\_’ after the <date> in the archive url as for example** [**https://web.archive.org/web/20180101170353id\_/http://www.ltg.ed.ac.uk/~ht/**](https://web.archive.org/web/20180101170353id_/http%3A/www.ltg.ed.ac.uk/~ht/) **(different functions described on https:/en.wikipedia.org/wiki/Help:Using\_the\_Wayback\_Machine).**

Answer November 2018:
I believe this is met by the more thorough description of coverage-spec (renamed to precision spec) in assignment and resolution.

-------------

Some of this is corrigible, but only at the expense of a vast increase in detail and clarity. Others not-so-much, insofar as so much is \_necessarily\_ left underspecified, to allow for \_anything\_ to be considered as an archive ('archive-id' value), with the consequent requirement to allow for arbitrarily idiosyncratic internal naming mechanisms ('archived-item' value).

**-- Eld (10-08-2018): It should be clarified that this is only in case an identifier can be found in the archive – actually it should probably also be mentioned under assignment that in these cases, it is identifiers assigned to objects by the archive.**

Answer November 2018:
I have explicitly mentioned that in case of the item identifier, it will be identifiers assigned by the archive holding the item.

I'm curious in this connection to wonder if we actually have any URN namespaces where assignment is done "independently but following an algorithm that itself guarantees uniqueness"? Ah, there are registrations for urn:uuid:, which qualifies, and urn:oid:, which sort of does. OK, is there such a namespace in active use? Which provides for resolution, at least in principle?

**-- Eld (10-08-2018): we use UUIDs for identifiers for our library material, and we use WARC as packaging format, therefore it is essential for us to use the urn:uuid: since WARC only accepts URIs as identifiers**

Answer November 2018:
Not sure of what the question is here, - we surely do use the urn:uuid name space to a large degree, since we make WARC packages for our UUID-registered none-web materials – not sure of the relevance here either.

A crucial point here is that in the OID case a urn:oid:... carries a chain of responsibility with it. If you run across one, you know how to find out who takes responsibility for every level in the tree that's involved in interpreting it. In the UUID case, a urn:uuid: doesn't travel well/at all, so the question doesn't arise. But for PWIDs, if I find one I have \_no way\_ to figure out who's responsible, or whom I can ask for clarification, or whom to blame if it doesn't appear to 'work'.

**-- Eld (10-08-2018): I am not sure I understand this – or what is causing a possible misunderstanding. As said before, anyone can make a resolver for open web archives and any archive with restricted access can make their own (only having the challenge with archive identifiers as mentioned later). You can take the PWID and “manually use it to locate you resource. If the person making the reference made mistakes or if the archive lost its resources, I don’t think it is the PWID that should be blame – which I see a direct parallel to for any referencing systems – digital or analogue.**

Answer November 2018:
From what I have understood of this comment, I think that the much more thorough description of construction and resolution should cover it.

----------------

I don't see what the intended value of the discussion of the "SOLR-Wayback tool" is for the spec.

**-- Eld (10-08-2018): It is to say that you are not on a huge job if you want to collect all parts for a web page, - there actually exists tools to help you. If it is not appropriate to have it here, I can take it out.**

Answer November 2018:
I have changed the text of the assignment section, in order to be more clear why it is mentioned.

\*Interoperability\*

The above comments about ambiguity/lack of identity apply here too: if different implementors interpret e.g. a coverage-spec of 'subsite'

differently, their resolvers will not interoperate.

**-- Eld (10-08-2018): I can make a more clear definition if wanted, - admittedly, I have not gone into depth with this, since (as far as I know) there are no tooling for this yet, but it is a point that it should be more precise, - also to make clear what was validated when making the reference.**

Answer November 2018:
I have added more precise definition and better explanation

\*Resolution\*

None of this is very much to the point. In particular, it completely ignores the precedents set by doi.org, identifiers.org and n2t.net

(<https://doi.org/10.1038/sdata.2018.95> exemplifies the first and its content discusses the second and third).

**-- Eld (10-08-2018): I don’t understand this, - as far as I know these are dealing with registered identifiers. There are no way that all web materials will have a registered identifier, - the volume is too big. That is the reason why the PWID is needed – … maybe I have misunderstood some things here.**

Answer November 2018:
This should be covered by emphasizing that the PWID is a supplement to reference of web (archive) resources.

==============

In conclusion, I think a useful way to think about this proposal is as a misguided attempt to define a urn type which \_is\_ a citation. This makes sense of PWID's 3rd party nature, and clarifies that the archive-id and -time, the embedded URI and the coverage spec are just a pretty arbitrary, certainly impoverished, subset of what you would expect to find in a \_real\_ citation. Trying to pack a citation into a URN just doesn't make sense to me. Neither does trying to determine exactly the \_right\_ subset of citations of web-hosted resources to pack up in order to be both useful and unambiguous.

**-- Eld (10-08-2018): I hope I have made it more clear in these comments**

(There is work underway in various venues (see e.g. [1]) to \_objectify\_ citations and give PIDs to \_those\_, which might address at least some of the goals of this work.)

Answer November 2018:
I must admit that I find that this URN PWID suggestion does actually fulfil all of the points:

* being definable in a machine-readable manner – requiring ontology modifications;
*(that is actually why we need the PWID as an URN)*
* being storable, searchable and retrievable – requiring a well-structured open database;
*(except from the database part – which does not practically apply for web archives – then it is the web archives that makes this possible by various search and Wayback interfaces)*
* being identifiable – requiring a new a global Persistent Identifier; and
*(that is why we need the PWID in general, as you cannot in practice assign identifiers for each harvested web element – the PWID utilize the data recorded at archival point to provide an identifier anyway)*
* having a Web-based resolution service that takes the identifier as input and returns a description of the citation.
*(that is not practically possible for web material – but via the prototype you can be taken to your resource)*

ht

[1] <http://sched.co/DJ3P>

--

 Henry S. Thompson, School of Informatics, University of Edinburgh

 10 Crichton Street, Edinburgh EH8 9AB, SCOTLAND -- (44) 131 650-4440

 Fax: (44) 131 650-4587, e-mail: ht@inf.ed.ac.uk

 URL: <http://www.ltg.ed.ac.uk/~ht/> [mail from me \_always\_ has a .sig like this -- mail without it is forged spam]

Answer November 2018:
I have of course also added reference to the awarded PWID poster at iPRES 2018, and that there is an accepted more researcher-oriented version of this poster for iDCC 2019