Persistent Web References – Best Practices and New Suggestions

Eld Zierau
Digital Preservation Dept.
The Royal Library of Denmark
P.O. Box 2149
1016 Copenhagen K - Denmark
+45 9132 4690
elzi@kb.dk

Caroline Nyvang
National Heritage Collection Dept.
The Royal Library of Denmark
P.O. Box 2149,
1016 Copenhagen K - Denmark
+45 9132 4919
cany@kb.dk

Thomas Hvid Kromann
National Heritage Collection Dept.
The Royal Library of Denmark
P.O. Box 2149
1016 Copenhagen K - Denmark
+45 9132 4422
thok@kb.dk

ABSTRACT

In this paper, we suggest adjustments to best practices for persistent web referencing; adjustments that aim at preservation and long time accessibility of web referenced resources in general, but with focus on web references in web archives.

Web referencing is highly relevant and crucial for various research fields, since an increasing number of references point to resources that only exist on the web. However, present practices using URL and date reference cannot be regarded as persistent due to the volatile nature of the Internet, - and present practices for references to web archives only refer to archive URLs which depends on the web archives access implementations.

A major part of the suggested adjustments is a new web reference standard for archived web references (called wPID), which is a supplement to the current practices. The purpose of the standard is to support general, global, sustainable, humanly readable and technology agnostic persistent web references that are not sufficiently covered by existing practices. Furthermore, it can support better practices for precise references in spite of the temporality issues for web material as well as issues related to closed web archives.

In order to explain needed change of practices based on the wPID, the paper includes a thorough description of the challenges in web references. This description is based on the perspectives from computer science, web collection and Digital Humanities.

Keywords

Persistent identification, Web references, Web Persistent Identifiers (wPID), Web elements, Digital Preservation.

1. INTRODUCTION

The main goal of this paper is to suggest needed changes to web reference practices. The approach is to explain the need for changes, and how the suggested wPID standard can assist in achieving better practices by addressing persistency issues that are not properly addressed in current practices.

Today, there are still major issues concerning non-persistent web references. As illustration of the highly relevant need for ways to mitigate these challenges, a 2014 paper [23] found:

> … that more than 70% of the URLs within the Harvard Law Review and other journals, and 50% of the URLs within United States Supreme Court opinions, do not link to the originally cited information.

A persistent web reference is here defined as a persistent identifier (PID) for a web resource. In many cases, web references are not persistent as they consist solely of a web address and an extraction date, where the web address is a Uniform Resource Locator (URL[1]) specifying a web resource location and a mechanism for retrieving it [20]. Such references break as information on the Internet changes.

The subject of persistent web referencing has been discussed almost for as long as the web has existed. As early as 2001, a journal paper about "Persistence of Web References in Scientific Research" was published [13]. Persistent web references are needed in order to avoid the so-called "reference rot" problem, which is a combination of link rot (where a link can become inaccessible on the live web) and content decay (content changes). Examples of causes of reference rot are that a web resource has been changed moved, deleted, or placed behind a pay wall [16,18].

Persistent web referencing is increasingly relevant for research papers, as online resources are increasingly used in scholarly research (e.g. blogs) [9]. Furthermore, the persistency of web referencing is fundamental for preservation of research as well as for documentation and traceability.

There is also an increasing amount of research that is solely based on web resources [6].[2] Such research will in this paper be referred to as web research. Compared to traditional web references to documents from research papers, web researchers face a number of unique challenges, e.g. data management, references to closed archives and identification for precise annotation and referencing. However, as more and more researchers complement traditional sources with web material, these challenges will in the course of time apply to most research. Thus, when considering a general web reference proposal, the issues from web research need to be taken into account. This paper will discuss such issues within the context of Digital Humanities web research, where sustainability of web references is one of the main concerns [5].

The exact definition of a "persistent identifier" is debatable. John Kunze suggests that persistent identifier simply means that "an identifier is valid for long enough" [11]. For references in research papers this could be well over 100 years. As Juha Hakala points out: "persistent identifiers should only be assigned to resources that will be preserved for long term" [11], in other words; an identifier is worthless unless the resource it

[1] Although URL is more or less deprecated, this is the term used in the various citation templates. In order to avoid unnecessary confusion, the URL term is also used for online references

[2] Evidence can e.g. be found in reports from the BUDDAH project. See (wPID reference) wpid:archive.org:2016-03-13T011611Z:http://buddah.projects.history.ac.uk/.

points to is under a preservation program, and the identifier can be used to access the contents.

Currently, there are various approaches to the challenge of persistent web referencing, all of which includes some sort of archiving. These include registration of web resources in PID services like DOI [11], references to web archives, a method that is increasingly being applied [3], and the use of emerging citation services [16]. One of the challenges with today's web archive reference practices is that they refer to the archive resource by including the URL for the web archives access service. This means that the archive URL may break over time due to change of access services, name shift of the archive domain or if the web archive ceases to exist [16,15].

A major obstacle to persistent web referencing for archived URLs is the temporalities not only for a single URL, but also for all the elements contained in a web page located by a URL [1]. These challenges have also been some of the motivation for the creation of the Memento protocol that can assist in finding an archived URL in a limited set of open web archives [19,2]. A recent draft report on *Interoperation Among Web Archiving Technologies* addresses these issues and points services for web archives as part of the solution [15].

The complexity of embedded material in web pages also implies that different web references can be of different quality both regarding the persistence (e.g. the trustworthiness of its survival) and its quality (a web page may not be fully harvested). Thus, in order to make trustworthy persistent references to a web page, one may need to evaluate whether several versions exist in different archives, and which version of the web page (and embedded elements) best fulfils the purpose of the reference. Therefore, this paper will include a discussion of elements to be considered when determining which web reference to use.

In order to accommodate the various challenges and support enhancement of practices for persistent web references, we propose a general global persistent web reference scheme called wPID (**w**eb **P**ersistent **ID**entifier). It is primarily focused on archived web references as a supplement to existing PID services. The wPIDs are designed to be general, global, humanly readable and agnostic regarding specific web archive implementations of access technology. The proposal is based on an analysis made from the perspectives of computer science, web collection and Digital Humanities research. Additionally, the paper describes how to represent the wPID reference scheme as a (Uniform Resource Identifier) URI scheme that can be the basis for future resolution of such identifiers [4].

The paper begins with a walkthrough of the state of the art in persistent web referencing and an introduction of the new wPID. This is followed by explanation of the various challenges in web referencing which are not covered by current best practices. Finally, the new wPID is defined as support to new best practices.

Throughout the paper the term URL will be used when addressing online web addresses (past or present), and the more general term URI will be used in relation to PID standards and archived web resources. Furthermore, any references to web resources will be provided in the new suggested wPID standard, linking to the corresponding current archive URL.

2. STATE OF THE ART AND NEW WPID

As illustrated in Figure 1, we currently have a number of different web referencing techniques and recommendations, all of which rely on the continued existence of the source material on the live web or in some sort of web archive.

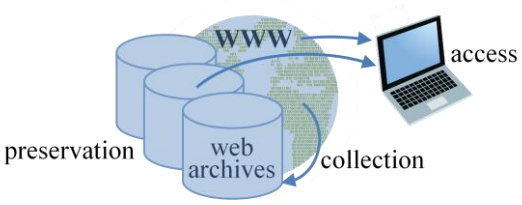

Figure 1: components for persistent references

Below, the current best known ways to make web references are described, regardless of the content that is referred. This covers the following four main ways to make references:

- ***Reference using URL and date***
  A web reference can simply consist of a web address in the live web along with a retrieval date. This is a commonly used (non-persistent) way to make web references.
- ***Reference using existing web PID services***
  A number of existing PID services provides means for content holders to register their resources. Registered web resources can then be referenced via the PID.
- ***Reference using web archives***
  Web archives can offer ways to address their content which can be used as a web reference. For example, URLs to the Internet Archive's Wayback service.
- ***Reference using citation services***
  A number of citation services offer authors a facility to crawl and store relevant resources to be cited, where web references are provided for later access.

This section will describe these four different referencing techniques and end with a short description of the advantages and disadvantages for each technique. It will also shortly introduce the new wPID in order to compare it with current practices. Further description of the wPID will be given later in this paper.

2.1 Reference using URL and date

A commonly used web reference form is to give a URL along with its retrieval date, as for example for reference [7]:

> http://bds.sagepub.com/content/1/1/2053951714540280, retrieved March 2016

This type of reference conforms to the type of website citations using *url* and *accessdate* on "Wikipedia's Template:Citation"[3], and is similar to most scholarly citation styles, e.g. the Chicago style[4], and the APA[5] style that both request the URL and the access date of the cited resource [16]. However, as posited, links can become inaccessible or content can change on the live web. Although commonly used, these types of references do not provide persistent identification of a resource.

2.2 Existing Web PID services

Today, there are a number of PID services offering content holders the ability to register their resources (which the content holders then preserve themselves). PID services for digital objects have been recommended as a way to ensure persistent web references [11].

---

[3] wpid:archive.org:2016-03-25T113243Z:https://en.wikipedia.org/wiki/Template:Citation

[4] wpid:archive.org:2015-10-07T053612Z:http://www.bibme.org/citation-guide/chicago/website

[5] wpid:archive.org:2016-03-08T233451Z:http://studygs.net/citation.htm

An example is the DOI PID service where resources can be registered and given a DOI-reference that can later be used for retrieval of the resource, e.g. the above [7] reference has the DOI reference: doi:10.1177/2053951714540280. However, PID services cannot stand alone, since many relevant references are not registered with a PID, and it is solely up to the content holder of the resources to handle registration and preservation.

A chronological list of some widespread PID services is [11]:

1. **Handle**, 1994
2. **U**niform **R**esource **N**ame (URN), 1997
3. **P**ersistent **U**niform **R**esource **L**ocators (PURL), 1995
4. **A**rchival **R**esource **K**eys (ARK), 2001

Handle[6] is a naming service that provides a mechanism for both assigning persistent identifiers to digital objects. It offers resolving of the persistent identifiers and allowing location of the authority that is in charge of the named information.

URN[7] is a concept that creates a common namespace for many different kinds of identifiers, independent of technology and location. The basic functionality of a URN is resource naming that conforms to the requirements of the general URI[8], but a URN will not impede resolution as e.g. a URL does.

PURL[9] relies on a technical solution that allows web addresses to act as permanent identifiers. It is a URL with intermediate resolution service. PURL conforms to the functional requirements of the URI, and PURL uses the HTTP protocols.

ARK[10] introduces a concept combining persistent identification and technical and administrative frameworks. This enables reference to different types of entities, e.g. agents, events and objects with metadata records. The ARK is designed to allow integration with other identifier schemes.

Besides these PID services a number of standards and services have been developed, the best known being:

1. Digital object identifier (**DOI**)
2. International Standard Book Numbering (**ISBN**)
3. National Bibliography Numbers (**NBN**)

DOI[11] makes use of the Handle System for resolving identifiers in a complete framework for managing digital objects along with policies, procedures, business models, and application tools. It is designed to be independent of the HTTP protocol.

ISBN[12] has been around as a 10 (later 13) digit Standard Book Numbering format since the 1960s. In 2001 ISBN was also described as a URN name space.

NBN[13] has no global standard, but has country-specific formats assigned by the national libraries. It is used for documents that do not have e.g. an ISBN. In 2001 NBN was described as a URN name space.

Additionally, there are communities who employ their own PID services, as for example DataCite[14] which is a community based service using DOIs for research data samples, in order to make these searchable and referable.

If a PID is registered for a resource, the idea is that the resource will be accessible through a resolver service (via live www access in Figure 1), and by a set of rules that ensures the preservation of the content that the PID addresses, but where it is the resource holder who holds responsibility for ensuring preservation program for the resource.

2.3 References to Web Archives

An increasing number of references target open web archives like Internet Archive's collection via their Wayback service. This service offers access to a lot of the harvested web pages from the Internet Archives web archive, for example for [9]:

https://web.archive.org/web/20160315035636/http://bds.sagepub.com/content/1/1/2053951714540280

This URL can be used as the *archiveurl* in website citations using *url*, *accessdate*, *archiveurl* and *archivedate* on the above mentioned "Wikipedia's TemplateCitation".

The number of Web archiving initiatives is growing. This is evident from the growth of the member list[15] of the International Internet Preservation Consortium (IIPC[16]), which is dedicated to improving the tools, standards, and best practices of web archiving for research and cultural heritage.

There is no general reference pattern for archived URLs. However, there are similar URL path patterns for web references via *archiveurl* to online web open archives using Wayback for access. All such *archiveurl* include archive date and time (denoted *date* below) and archived original URL (denoted *uri* below). The following is a list of selected open web archives and the URL patterns they use. The path differences are highlighted in bold:

- *Internet Archive (archive.org)*:
  'https://**web.**archive.org/**web**/' + <*date*> + '/' + <*uri*>
- *ArchiveIt service build by Internet Archive (archive-it.org)*
  'http://**wayback.**archive-it.org/**all**/' + <*date*> + '/' + <*uri*>
- *UK Web Archive (webarchive.org.uk):*
  'http://**www.**webarchive.org.uk/**wayback/archive**/' + <*date*> + '/' + <*uri*>
- *Portuguese web archive (arquivo.pt):*
  'http://arquivo.pt/**wayback**/' + <*date*> + '/' + <*uri*>

The differences in the paths are due to differences in the implementation of the access services at the different web archives. Thus Web references via *archiveurl* to online web archives can only be resolved as long as the web archive exists and the access path resolves to an existing access service. However, such patterns may not be valid for future access implementations.

Similar patterns may not be found for all closed archives. For example, in the Danish web archive, there are no explicit

---

[6] wpid:archive.org:2016-03-04T031302Z:http://handle.net/

[7] wpid:archive.org:2016-03-07T210340Z:http://tools.ietf.org/html/rfc1737

[8] URNs and URLs denote subsets of URIs [4]

[9] wpid:archive.org:2016-03-04T023751Z:https://purl.org/docs/index.html

[10] wpid:archive.org:2015-09-27T040046Z:https://confluence.ucop.edu/display/Curation/ARK

[11] wpid:archive.org:2016-03-05T022511Z:https://www.doi.org

[12] wpid:archive.org:2016-03-24T051018Z:http://www.isbn.org/ISBN_history

[13] wpid:archive.org:2016-03-31T131818Z:http://tools.ietf.org/html/rfc3188

[14] wpid:archive.org:2016-04-16T144351Z:https://www.datacite.org/about-datacite/what-do-we-do

[15] An even bigger list of web archiving initiatives can be found on wpid:archive.org:2016-03-19T171515Z:https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[16] wpid:archive.org:2015-04-03T190314Z:http://netpreserve.org

archiveurl, instead there is information about the placement of the resource record in a WARC file by WARC file name and offset. As the WARC file name is not part of the bit preservation and thus may change over time, researchers are recommended to supplement a reference with the archived original URL (here http://netarkivet.dk) and harvest time (given in brackets) [14]:

> http://netarkivet.dk 197800-188-20140107085943-00000-sb-prod-har-005.statsbiblioteket.dk.warc/4773261 (9:01:06 jan 7, 2014 i UTC tid)

However, this reference does not include specification of which web archive the resource was retrieved from.

Another aspect of *archiveurl*s is that they may contain inherited information about special functions in web archive's access technology. An example of such a function is the *Identity* Wayback function as used by the Internet Archive. This function is called by placing 'id_' after the <date> in the *archiveurl* [17]. Another example is the function giving a snapshot image of the page[17] [16]. However, such functions may not exist in the future.

2.4 References using Citation Services

In the past years a number of citation services have emerged. These services provide on-demand archiving of a version of a given resource [16]. Examples are:

- **WebCite**[18] is an on-demand archiving system for web references. WebCite is run by a consortium and provides a tool that can archive a web reference as well as provide a new URL in the www.webcitation.org domain, where the harvested and archived referenced resource can be accessed [10].
- **archive.is**[19] (formerly archive.today) is a privately funded on-demand archiving system that takes a 'snapshot' of a web page along with a harvest of the web page (excluding active elements and scripts). The archived web page is assigned a new short URL for subsequent access.
- **perma.cc**[20] is a web reference archiving service that offers users to create a Perma.cc link where the referenced content is archived along with some metadata (URL, page title, creation date). A new link to the archived content is generated [16].

Additionally, certain web archives allow users to nominate a web page for archiving, e.g. the UK Web Archive[21], the Internet Archive[22], and Netarkivet[23]. However, for national archives like the UK Web and Netarkivet, only web pages that are considered to fall within a national scope will be archived.

Other variants exist, e.g. Zotero[24], which allow researchers to archive resources, and Wikisource that specializes in archiving Wikipidia sources[25].

2.5 References Using the New wPID

The suggested new wPID definition is a web archive reference that is independent of current web archive access technology and online access. A wPID consist of three main components, which in general are sufficient to identify any web reference in an arbitrary web archive. These three components are listed in table 1.

Table 1. Web Persistent Identifier (wPID) main parts

Part	Format	Example
Web archive	**Text**	archive.org
Date/time	**UTC timestamp**	2016-01-22T11:20:29Z
Identifier	**URI** (harvested URL)	http://www.dr.dk

For the example of reference [7], the wPID is

> wpid:archive.org:2016-03-15T035636Z:http://bds.sagepub.com/content/1/1/2053951714540280[26]

The wPID is not currently resolvable. However, it would be relatively easy to create services[27], which are based on web archive, <*date*> and <*uri*> from the wPID. This also covers closed web archives (through restricted access interface) as web archives have indexes of contents where <*date*> and <*uri*> can be use as basis for finding the current web archive URL for access.

2.6 Advantages and Disadvantages

Generally, persistency of an identifier depends on the sustainability of locating the resource by use of the identifier and that the resource content is accessible in the intended form. This is applicable to both analogue and digital resources, but the volatile nature of the Internet makes sustainability a more crucial consideration for web references. Thus, for all discussed alternatives, claims of persistency should be measured by the likelihood of a resource being locatable and accessible (with preserved contents) at a later stage.

Reference using URL and date: This reference can never be persistent. The contents can change several times during the specified date. Thus when the resource is retrieved at a later stage, there is no way to check whether it has indeed changed, and whether its contents are the intended contents.

Reference using existing web PID: Persistency relies first of all on whether a resource is registered with a PID. Of further concern is the future existence of resolver services (e.g. cases like the outage of the DOI resolver service in early 2015)[28] and whether content holders maintain the accessibility of their resource. Accessibility will rely on whether the resource holder has ensured that the resource is covered by a digital preservation program. Furthermore, for services like DOI,

[17] A snapshot example is: wpid:archive.org:2013-06-19T224334Z:https://archive.is/J4I1a/image.

[18] wpid:archive.org:2016-03-06T000304Z:http://webcitation.org/

[19] wpid:archive.org:2016-02-19T153542Z:http://archive.is/

[20] wpid:archive.org:2016-03-05T093301Z:https://perma.cc/

[21] wpid:archive.org:2016-03-04T052011Z:http://www.webarchive.org.uk/ukwa/info/nominate

[22] wpid:archive.org:2016-03-01T085607Z:http://archive.org/web/

[23] wpid:archive.org:2016-03-03T220018Z:http://netarkivet.dk/

[24] wpid:archive.org:2016-03-06T080434Z:https://www.zotero.org/

[25] wpid:archive.org:2016-02-27T212014Z:https://en.wikipedia.org/wiki/Wikisource

[26] Omission of ":" in date/time is described later.

[27] Discussion on APIs (including Open Wayback) includes mentioning of APIs for such services [15].

[28] wpid:archive.org:2016-03-10T044938Z/http://blog.crossref.org/2015/03/january-2015-doi-outage-followup-report.html.

persistency hinges on ongoing payment of service charges. On the positive side, fees for lack of maintenance of the DOI mean that there is a strong motivation for maintaining the DOI as long as it exists.

Reference using Web Archives: Persistency relies on the continued existence of a web archive, and the preservation program that the archive has for its resources. As mentioned in [16]: "one link rot problem was replaced by another" if the archive ceases to exist. Furthermore, future existence of compatible access services as archive links with inherited service and service parameters may be at risk due to future changes in access tools or archive ownership.

Reference using Citation services: These services are in many aspects similar to web archives, and so the persistency of references depends on the continuation of the given service and the future existence of compatible access services as well as preservation program for the resources. An example of a vanished citation service is the former mummify.it citation service mentioned in [16], which in the Internet Archives web archive was used in the period from 2013-08-30 to 2014-02-14. In 2015, it had changed to an online shoe shop and is now inactive.

Reference using the new wPID: As for web archive and citation services references persistency rely on the existence of the web archive and its preservation program. The advantage is that a wPID has sufficient information to identify a referred source in any web archive independent on access implementations and/or generated IDs like shortened URLs. Current lack of resolving may be seen as a disadvantage, but services can easily be made and these services can be maintained to point to access platforms as they change due to change in technologies.

Logical preservation of resources needs to be part of the required preservation program for all resources pointed to by persistent web references. Logical preservation covers aspects of keeping the resource accessible in spite of technology and format changes. For controlled web resources (e.g. handled by PID systems) this can include migration of formats. One of the solutions for web archives that is now being investigated is emulation, e.g. oldweb.today.[29]

It should be noted that a major difference between PID services and web archives is the placement of responsibility of preservation management. PID services only provide identifiers where the resource holders are responsible for content preservation, while it is the web archives that have this responsibility for archive references (which is the same for most citation services).

In the rest of this paper, we will leave out further analysis of the ***URL and date*** reference type, as it can never become a persistent way of referencing a web resource. As the aim here is to focus on references to the archived web as a supplement to existing practices, where there may not be a holder of the resource, further analysis of ***existing Web PID services*** is also left out.

The focus in the rest of the paper will be on what a web reference actually means, taken into account the needs from researchers, the quality of a web reference according to its purpose and the ambiguities that can be inherited in a web reference.

[29] wpid:archive.org:2016-03-08T205232Z:http://oldweb.today/

3. RESEARCH IN WEB MATERIAL

In many ways, web researchers using web references face challenges that are similar to referencing to digital papers and resources. However, in the field of web research it is more obvious that there are additional requirements, which must be taken into account in order to make the best possible proposal for general web references.

Here, the additional web research requirements are illustrated by investigating current issues in Digital Humanities. Today, Digital Humanities is used to describe at least two entwined processes:

1. With the advent of new computational techniques researchers are able to process a hitherto unseen amount of data (whether born-digital or reformatted), and
2. As the hegemony of conventional media is being challenged, scholars must now trace a number of cultural processes and interactions in the form of digital artefacts [8]

These new circumstances call for new measures, yet the lack of a shared and stringent methodology is a well-recognized cause for concern within Digital Humanities [6,7]. This is particularly true when it comes to research in web materials – a budding empirical field within both the Social Sciences and the Humanities. Web researchers have to cope with unique issues due to the dynamic content of their empirical field.

Web research, whether using the live or archived web, is faced with challenges related to both data management and identification for annotation and referencing (figure 2). Identification is here understood both as the actual search for material as well as the means to identify the precise content of a web reference.

Figure 2: components for web research

It should be noted that a lot of such references have a potential problem regarding access, as usage of non-public references might be restricted (e.g. denied or limited to on-site users) due to regional legal frameworks.

3.1 Analogue Standards and Current Needs

From the perspective of an institution dedicated to cultural heritage preservation and research, this paper grapples with one of the cornerstones of sound methodology, namely the ability to give citations in keeping with current scientific standards.

The purpose of accurate referencing is – first and foremost – to give readers an opportunity to assess the reliability and provenance of a given source material as well as to retrace and reproduce research steps. As such, the current inability to provide reliable references touches on all components – identification, data management and access – for web research (figure 2).

Within the Humanities and Social Sciences, reference systems are structured to provide the most accurate link to a given object. In original sources this entails pointing to distinct passages, foot notes, a word or even marginalia. For published material, which can appear in a number of different editions, citation styles often require users to include unique identifiers

such as the former mentioned ISBN and the related Serial Item and Contribution Identifier (SICI[30]) for periodicals.

Yet for web pages, the most commonly used style guides (e.g. formerly mentioned Chicago style) request nothing beyond the URL and a date indicating the date a URL was "last modified" or merely "accessed".

The discrepancy between these standards and the requirements of conventional research means that researchers might shy away from incorporating web materials or that web research will in itself be discredited due to methodological inadequacies.

In conclusion, there is a present and urgent need for a persistent web referencing scheme on par with that for analogue materials.

4. WEB REFERENCING CHALLENGES

The differences between referencing scheme for analogue and web references are mainly due to the dynamic nature of the web and the temporalities within complex web pages. The differences and related challenges are especially pertinent for researchers referring to complex web resources, as is often done in web research.

Determining whether a link is "alive" or "dead" poses an additional challenge. There are notions of dead links in connection with a citation that points to the live web, but there is no clear definition of what "dead" entails if we take into account that a link can potentially live on in an – possibly off-line – archive. Since persistency does not necessarily rely on what is online, this needs to be taken into account in regards to the challenges of persistent web references.

The following sections describe the dynamics and context of "dead" links, and are concluded by a section discussing the quality of a persistent web reference with respect to these issues.

4.1 A Relative Whole with Temporalities

One of the major challenges with persistent references of web pages is that these are mostly comprised of separate parts as illustrated in Figure 3. In this example the URL only refers to an html element, which includes style sheets, text, links etc. The links are new URLs to elements embedded in the web page (e.g. images) or URLs to elements in form of other resources (e.g. link to PDF files).

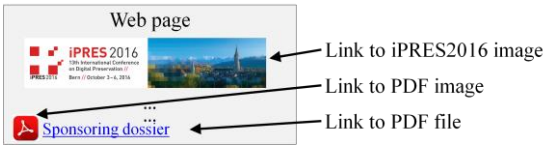

Figure 3: Elements in a web page

Different elements are harvested at different times, and some elements may only partially be harvested or not at all. This causes troublesome temporalities or incompleteness in the web archives.

For single self-embedded elements like a PDF file the temporalities are not an issue. However, for web pages with dynamic contents the temporalities can be crucial, see for example slide 8 of the *Evaluating the Temporal Coherence of archived pages* IIPC presentation [1] in which a weather forecast predicting storm and rain is depicted with a clear radar image extracted 9 months after the harvest of the main page.

[30] wpid:archive.org:2016-03-04T102536Z:http://www.niso.org/apps/group_public/download.php/6514/Serial%20Item%20and%20Contribution%20Identifier%20(SICI).pdf

The challenges of temporalities and coverage make web archives a rather difficult academic resource [6].

The temporality challenge implies that an archived web reference may be ambiguous. Traditionally, it is the archive software that picks the version of page elements, but for an exact research reference it may be necessary to specify each of the elements. Consequently, all parts should be denoted with wPIDs, which in some cases may incorporate wPIDs for parts found in separate web archives (also mentioned in [15]).

Another challenge is that web archives – open as well as closed – will never be able to contain snapshots of the entire Internet. One reason is the continuous change in content and the challenge of temporality, but also the fact that the amount of data is simply too big. Today, a number of web archives cover different parts of the web. Typically, national web archives systematically harvest the Top Level Domains of the country, but Top Level Domains like .com, .org and .nu are not covered in full by any web archives.

4.2 Variety of Errors in Web Page Search

When looking for or looking up a web reference in a web archive, it is important to be aware of the possible reasons why a page seems to be missing from the archive.

In general, a "missing" reference can either be caused by limitations or errors in how the related URL was collected or how it is accessed. This is illustrated in Figure 4.

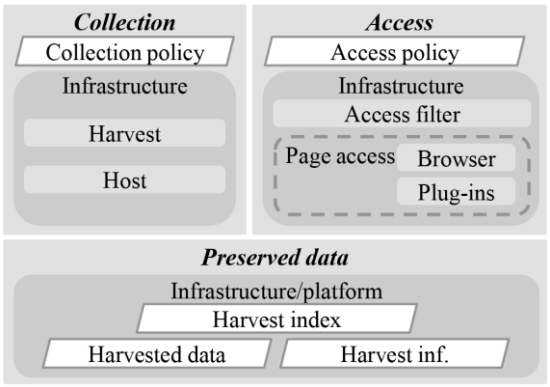

Figure 4 Web Archive Infrastructure

Below, each of the three components is described with its name in ***bold/italic***, and subcomponents are highlighted by being written in *italic*. Archived URLs are denoted URIs, as they exist within a web archive and therefore do not represent locators. The description includes all possible error sources, including sources that do not exist in the web archive.

Collection causing a missing web reference may be due to:

Collection Policy: In case the web reference is not covered by policy and thus never collected by the web archive in question.

Harvest of a host web reference can fail for a number of reasons:

- Errors in infrastructure (e.g. missing network connection)
- Bad harvest settings (e.g. stopped by max bytes)
- Cannot harvest inherited material (e.g. Flash)
- Cannot harvest scripting (e.g. java scripts and AJAX)
- Harvester fails (e.g. due to crawler traps) or was killed
- Host or part of host is down or unavailable
- Host does not allow the harvest

In most cases, the above harvesting errors mean that a reference is not usable.

Access causing a missing web reference may be due to:

Access policy enforced by an *Access filter*: In some case there is limited access, e.g. respecting robot.txt by disallowing access or filtering of illegal material, special granted access to the web archive may be required to check if the reference is correct.

Page access can fail for a number of reasons:

- Unavailable preserved data due to infrastructure problems (e.g. network or access application is down)
- Errors or limitations in access program (e.g. cannot show https & ftp URIs or URIs with special characters like æ, ñ)
- Misunderstood date/time as it is specified as UTC time-stamp or errors in local conversion of time
- Errors in the index used for look up of data (e.g. wrongly generated index or problems with de-duplication)
- Normalization of URI doesn't conform with indexed URI
- Access programs may interfere with the display[31]

Browser used for access does not render the page correctly

- because the browser does not comply with standards used (or exceptions from standards)
- because the web page is from a time period requiring special browsers (e.g. Netscape) or special versions of browser

Plug-ins needed for access do not exist or are not supported on rendering platform (or no longer supported).

Furthermore, the ***Preservation data*** can cause access errors, either by having an erroneous *Harvest index*, by errors in *Harvested data* (e.g. packed with wrong ARC file offset) or by *Infrastructure / platform errors* (e.g. server with preserved data is down).

Understanding these potential error sources, it is now possible to classify whether a link has truly died, and what sort of "death" we are encountering.

4.3 Different types of "Dead" links

A 'dead link' is commonly associated with link or reference rot, however, there are many ways that a link can 'die', therefore we need to look closer at the variations of what link rot means.

The archived content for a URI depends on harvest and consequently resolving of the URLs. Thus, a proper analysis of a persistent web reference must include consideration of the different types of "deaths" of both web URLs and archived URIs.

The following description relates to the search for a web reference in the form of a URL/URI (and possible date/time) with *expected* contents and may refer to HTTP codes[32] resulting from URL/URI requests.[33]

The following lists possible types of "deaths" for URLs on the live web:

- *Indistinguishably dead*: The page does not seem to exist, e.g. HTTP return a code indicating net or host errors
- *Instantly dead*: The page cannot be found, e.g. it resolves to an HTTP 404 code "page not found" generated by the server or a web page offering you to buy the domain or just redirects to some random domain.
- *Identity dead*: the page is not the expected page, e.g. due to new domain owner.
- *Simulated dead*: the page cannot be accessed due to some sort of blocking such as content filters or firewalls (also called soft errors in [15]).

Note that the classifications are conceptual and cannot necessarily be linked to specific technical traceable HTTP codes. This means that it can be hard to verify whether a page is *Instantly* or *Indistinguishably dead*. For example, a disappeared domain can resolve with the same error as missing network connection.

Today, live link checkers can search for dead links mainly by relying on technical HTTP codes. That means a "page not found" generated internally from a server may be regarded as a successful unbroken link as it will not return an HTTP 404 code. *Identity dead* links will also be reported alive and the link checker will not be able to determine whether a link is *Indistinguishable dead* or *Instantly dead*.

It becomes even more complicated when searching for content in an archive due to the possible harvest/access/preservation errors described in the previous section:

- *Archival dead*: A URI (and date/time) doesn't exist in the archive
- *Partially dead*: A URI (and date/time) does exist in the archive, but cannot be correctly displayed because pieces are missing due to harvest limitations
- *Technology dead*: A URI (and date/time) does exist in the archive, but it is not correctly displayed because of access limitations, e.g. due to browser or plug-in limitations
- *Apparently dead*: A URI (and date/time) cannot be found in the archive, due to errors in the access part, e.g. cannot access HTTPS URIs, wrong indexes etc.
- *Temporarily dead:* A URI (and date/time) can be found in the archive, but infrastructure problems or limitations like robot.txt make it temporarily inaccessible

Again these death types are conceptual classifications, and they are not necessarily easy to recognize, as symptoms of errors may differ for different access applications.

Finally, there is the *Ultimate dead* meaning that the URL/URI is neither in any archives nor on the live web. This will probably be impossible to verify, as we can never be sure whether we know all archives and whether all possible errors are taken into account.

4.4 Quality of a Persistent Web Reference

In general, use of web references as part of research or articles needs to be carefully evaluated for the intended purpose of the reference and its persistency quality both for the identifier and the resources identified. Specifically, for ***Reference using Web Archives*** the various mentioned web referencing challenges should be taken into account.

When choosing a web reference, the first task is to ***identify*** the needed reference in a web archive (or citation service) and verify that the resource can be accessed. For example, the reference is not *Apparently dead*, e.g. due to errors in the access application, and it is not *Instantly dead*, because of

[31] In the case of the Internet Archive Wayback, however, there are options to mitigate this challenge [17].

[32] wpid:archive.org:2016-02-29T024353Z:https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

[33] As noted in [18] contents from a dead link can be provided by other channels, e.g. contacting the author of a thesis referred.

reconstruction of the web archives access platform and the fact that the resource therefore needs to be found under another URL.

The next task would be to evaluate the referred ***contents*** with respect to referencing purpose. For example, it is not a case of *Simulated dead*, e.g. that the harvested resource is not just a login screen for password protected content.

Furthermore, it must be checked that the referred resource is of the right *Identity*, as could be the case for the mentioned mummify.it example, which at one stage was a citation service and at another stage a shoe sales site. In this example it is easy to recognize, but differences may be subtle and thus harder to recognize.

The purpose of the reference is crucial, since *Partially dead* referred content may fulfil its purpose, e.g. a web page containing complicated java script and flash can be harvested incompletely, yet the rest of the content might still be accessible and adequate for the referencing purpose [18].

Finally, an evaluation of the ***persistency*** should be performed in terms of future accessibility of the resource. This includes evaluation of the identifier as well as the contents referred.

The referenced resource may suffer *Archival dead* if the web archive partly or fully ceases to exist, i.e. an evaluation of the sustainability of the web archive(s) should be included. As an example, this paper will have a lot of invalid wPIDs in the future if the Internet Archive web archive is shut down.

The referenced resource can suffer a *Technology dead* if the web archive does not have a proper preservation program, and thus fails to keep the resource's existence or resource's functionalities available over time. Sustainability of access services should also be evaluated, in particular for web archives in the form of citation services relying on shortened URLs as persistent identifiers. Business and funding models are crucial elements in this evaluation.

5. SUGGESTED WPID REFERENCES

The suggested wPID aims at simplicity, readability, sustainability and transparency. The definition is based on analysis of the state of the art of persistent referencing; relevant web standards and the need for web research and the various challenges described in the previous sections. Furthermore, it takes into account that it could benefit from becoming an accepted permanent URI scheme [4] as described and explained in the last part of this section.

5.1 General wPID Definition Suggestion

As described in Table 1, the wPID consists of three main parts. Below, there are provided more details on choices made for their structure and how this relates to existing web standards like the WARC standard (packaging format used for many web archives) [12] and URI scheme standard [4].

- **Web archive**
  Is specified by Sequence of URI Unreserved Characters ('-', '_', '.', '~', alpha: 'a'-'z', 'A'-'Z' or digits: '0'-'9'.
- **Date/time**
  Is specified as a short UTC timestamp with the same definition as the WARC-Date field in the WARC standard, i.e. formatted as YYYY-MM-DDThh:mm:ssZ, conforming to the W3C profile of ISO 8601 [12,22], but omitting ":" in order to conform with the URI standard (as explained later).
- **Identifier**
  Is a URI as defined for the WARC-Target-URI field in the WARC standard. This field is for the archived URI which must have a value written as specified in [4,12]

There are no real restrictions to what a web archive name can be. In the examples used in this paper, the domain name for the archive is used. The reason for this is that the domain names are known today. However, proper names could be used if a register is created (similar to the NAAN registry[34] for ARK) and possibly maintained by the IIPC or a similar body. Such names could be *InternetArchive* for archive.org, DKWebArchive for the Danish web archive etc. In all cases, a register should be made at some stage, since archive domains can change (e.g. archive.today is now named archive.is). Note that such a registry should allow several names for each archive, since archives may be merged or renamed. Thus, old references need to remain persistent and traceable, regardless of use of the old name.

Additionally, we need to be able to avoid the ambiguity of *the parts and the whole*. We can accomplish that by specifying a *contentspec* parameter, which can have the values:

- *harvest*, in case the parts are taken from the archive in the traditional way,
- *part*, in case the wPID is to be interpreted as the single web page part.

We assume that "harvest" is default in case nothing else is specified.

Finally, in order to make it compatible with a URI, it must follow the URI syntax [4] and be defined as a URI scheme[35]. The URI syntax causes some challenges, since the definition will be recursive, as the defined wPID URI contains a URI in its definition:[36]

*wpid-**URI** = scheme ":"*
*<hierarchical part incl. archived-**URI**>*
["?" query] ["#" fragment]

The challenge is that there is no way to distinguish whether queries and fragments belong to the *wpid-**URI*** or the *archived-**URI.*** Thus queries and fragments cannot be given unambiguously to the *wpid-**URI***. The information about the *contentspec* therefore cannot be specified as a query, but instead it needs to be part of the *hierarchical part*. There is already an indirectly proposed solution for dealing with this challenge. Internet Archive specifies the access parameters for the Wayback, as previously explained, by adding a flag to the timestamp portion. Thus, the challenge can be solved by having the suggested *contentspec* as timestamp flag extensions in the same way.

Another challenge with the URI syntax is the limitation on the use of delimiters within the *hierarchical part*. If we define the

[34] wpid:archive.org:2015-09-17T131414Z:http://www.cdlib.org/uc3/naan_table.html

[35] The wPID URI scheme is registered as a provisional URI scheme, see wpid:archive.org:2016-04-17T062512Z:http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml and wpid:archive.org:2016-04-17T062702:http://www.iana.org/assignments/uri-schemes/prov/wpi.

[36] The syntax is defined using the same symbols as other RFC standards, i.e. test in double quotes "" are text to be written, text without double quotes are entities defined later, '/' means 'or', '[' and ']' surrounds optional parts, '(' and ')' surrounds parts that can be repeated where '+' means one or more times, and finally '<' and '>' surrounds explanatory text.

wPID as a URI with *scheme* "wpid" and a *hierarchical part* as a *path* with no authority and without segments, then the best choice of delimiter is ":". However, this collides with the colons used in the UTC timestamp. The suggestion to work around this challenge is to strip the colons in the UTC timestamp.

The resulting wPID definition is consequently the following:

wpid = "wpid:" webarchive ":" archivedate [contentspec] ":" archiveduri

webarchive = +(unreserved)
contentspec = "harvest_" / "part_"
archivedate = <as date/time in table 1 stripped for ":">
archiveduri = <as identifier in table 1>
unreserved = <as defined in RFC 3986 [4]>

A wPID (for an archive context) consisting of the example elements from table 1 would then be:

wpid:archive.org:2016-01-22T112029Zharvest_: http://www.dr.dk

since *harvest* is the default *contentspec* this is the same as

wpid:archive.org:2016-01-22T112029Z:http://www.dr.dk

Note that a wPID cannot leave out any of the syntax components from table 1, since all will be needed in order to make a persistent identifier. Thus the wPID should only be used when the reference is verified to be present in the specified archive.

The analysis of the quality of traditional web references suggests a need to add additional information about the reference target quality. However, it is not possible to do an analysis that can cover all possible scenarios, and it doesn't add any additional value on how to find the resource, thus this is not a subject for standardization, but could instead be made as a comment along with a wPID reference.

5.2 Why define wPIDs as URIs

It may not seem obvious why the wPID has to be defined as a permanent URI scheme in the form of a Request for Comments (RFC) as part of publication from the Internet Engineering Task Force (IETF)[37]. The claim here is that the benefits are worthwhile in spite of the disadvantages in form of the (not very elegant) workarounds for parameter and delimiters.

The benefits of a new wPID URI schema are first of all that it is a standard for the World Wide Web deployed since the creation of the Web [21]; secondly, it is the next step towards possible creation of some sort of resolving service via a browser, accessing locally or globally. For example, tools like the Memento tool could assist in wPID resolution, or special browser plug-ins recognizing wPID URIs could redirect to current access implementation (or APIs) using the HTTP/HTTPS protocols, and likewise from local browsers to closed archives.

6. DISCUSSION & FUTURE WORK

True persistence of a web reference will always come down to the existence of the archive responsible for the preservation of the reference contents and accessibility. This is true for all archive material, but will probably be a bigger issue for web archives as their existence hinges on the legislation and/or business models, which they are grounded on.

There are still challenges that are not fully addressed concerning data management, including corpus building and annotation. Some of the challenges relate to having unambiguous references web pages that may consist of several web elements that originate from one or more web archives. These challenges will be the basis for further investigation within current research projects[38] based on the suggested wPID standard.

Search for the right web reference has not been the focus of this paper. However, it is needed, and the Memento protocol is well suited for this task at least for the open web archives covered.

Additionally, when choosing a web reference in a web archive, it is important to take into account the possible temporalities and the evaluation of persistency of the archives. It will be the user of the web reference that is responsible for such evaluations, but compilation of guidelines for this task could be useful.

It will also be worth considering whether wPIDs could be applied as persistent references to digital library resources in general, i.e.

wpid:<library domain>:<timestamp>:<UUID for resource>

could be a reference to a library resource registered with a *UUID* and archived at the time specified in the *timestamp*. In this way it would also be possible to distinguish persistent identifiers for original versions and migrated versions of resources.

7. CONCLUSION

We have argued that there is an urgent need for better persistent web referencing practices, in order for researchers to include valid and precise web references in their research.

We proposed a new best practice for web referencing with a supplementary new wPID standard for references to web archives.

The paper has included a selected number of challenges within today's practices and future references and we have made a walkthrough of issues to be aware of when choosing a persistent web reference scheme. In particular, for wPIDs, this includes thorough validation of the web reference by the users of the reference before using it, as well as sustainability of the web archive, its preservation program for web resources and ability to offer access services based on archived URI and harvest time.

In addition, we have argued for the benefits of defining the wPID as an RFC standard by defining it as a URI scheme. This opens up the opportunity for a standard that can be used for technology independent access to web archives in the future.

The paper has included illustrations of the complexity and ambiguity that has become part of today's web referencing practices, especially for references to complex web pages. We argued that the suggested standard can be the basis for further studies of how to cope with these challenges including data management in web research.

8. ACKNOWLEDGMENTS

The authors wish to thank colleagues from the Royal Library of Denmark, especially Jakob Moesgaard, who contributed with web collection expertise. Also thanks to Jane Winters from University of London for her useful comments. Finally, thanks to digitalbevaring.dk for the illustrations.

[37] wpid:archive.org:2016-03-27T010831Z:https://www.ietf.org/

[38] The two co-authors of this paper are each conducting their Digital Humanities web research projects partly financed by the Danish Ministry of Culture: One project documenting online mourning and memorial culture and one project about the archiving of the various and constantly changing digital platforms of the literary field.

9. REFERENCES

[1] Ainsworth, S. G., Nelson, M. L., Van De Sompel, H. 2015. *Evaluating the Temporal Coherence of archived pages*, wpid:webarchive.org.uk:2016-01-14T233144Z:http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_18_Nelson.pptx.

[2] Alam, S., Nelson, M.L., Van de Sompel, H., Balakireva, L.L., Shankar, H., Rosenthal, D.S.H. 2015. *Web Archive Profiling Through CDX Summarization*, wpid:archive.org:2015-12-05T065734Z:http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf#sthash.G2BjVUf4.dpuf.

[3] AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L. 2014. *Who and what links to the Internet Archive*. In Research and Advanced Technology for Digital Libraries, Volume 8092, Lecture Notes in Computer Science, Springer pp. 346-357.

[4] Berners-Lee, T., Fielding, R., Masinter, L. 2005. *Uniform Resource Identifier (URI): Generic Syntax* (RFC 3986), wpid:archive.org:2016-03-26T121040Z:http://www.ietf.org/rfc/rfc3986.txt.

[5] Blaney, J. 2013. The Problem of Citation in the Digital Humanities, wpid:archive.org:2015-04-29T220653Z:http://w ww.hrionline.ac.uk/openbook/chapter/dhc2012-blaney.

[6] Brügger, N., Finnemann, N.O. 2013. *The Web and Digital Humanities: Theoretical and Methodological Concerns*. Journal of Broadcasting & Electronic Media 57, nr. 1, pp. 66–80. doi:10.1080/08838151.2012.761699.

[7] Burrows, R., M. Savage. 2014. *After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology*. Big Data & Society 1, no. 1, doi:10.1177/2053951714540280.

[8] Clavert, F., Serge, N. 2013. *Digital Humanities and History. a New Field for Historians in the Digital Age*. In L'histoire contemporaine à l'ère numérique / Contemporary History in the Digital Age, Clavert F., Serge N (eds.), pp. 15–26.

[9] Davis, R. M. 2010, *Moving Targets: Web Preservation and Reference Management*. Ariadne Web Magazine, issue 62, wpid:archive.org:2016-03-22T034748Z:http://www.ariadne.ac.uk/issue62/davis/.

[10] Eysenbach, G. 2005. *Going, Going, Still There: Using the WebCite Service to Permanently Archive Cited Web Pages*. JMIR publications, Vol 7, No 5.

[11] Hakala, J. 2010, *Persistent identifiers – an overview*, wpid:archive-it.org: 2013-10-10T181152Z:http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/.

[12] ISO 28500:2009. 2009, *WARC (Web ARChive) file format*.

[13] Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å., Kruger, A., Giles, C. L. 2001. *Persistence of Web References in Scientific Research*, wpid:archive.org:2016-03-05T222400Z:http://clgiles.ist.psu.edu/papers/Computer-2001-web-references.pdf.

[14] Netarkivet. Brugermanual til Netarkivet (User Manual for the Netarkivet). 2015. wpid:archive.org:2016-03-10T143320Z:http://netarkivet.dk/wp-content/uploads/2015/10/Brugervejledning_v_2_okt2015.pdf.

[15] Rosenthal, D.S.H., Taylor, N., Bailey, J.: *DRAFT: Interoperation Among Web Archiving Technologies*, wpid:archive-it.org: 2016-04-13T114517Z:http://www.lockss.org/tmp/Interoperation2016.pdf.

[16] Van de Sompel, H., Klein, M., Sanderson R, Nelson, ML. 2014. *Thoughts on Referencing, Linking, Reference Rot*, wpid:archive.org:2016-03-03T190515Z:http://mementoweb.org/missing-link/.

[17] Wikipedia. 2016. *Help: Using the Wayback Machine*, wpid:archive.org:2016-03-12T041354Z:https://en.wikipedia.org/wiki/Help:Using_the_Wayback_Machine.

[18] Wikipedia. 2016. *Link rot* wpid:archive.org:2016-03-19T113853Z:https://en.wikipedia.org/wiki/Wikipedia:Link_rot.

[19] The Memento Project, *About the Time Travel Service*, wpid:archive.org:2016-03-15T080039Z:http://timetravel.mementoweb.org/about/.

[20] Wikipedia. 2016. *Uniform Resource Locator*, wpid:archive.org:2016-04-05T134446Z:https://en.wikipedia.org/wiki/Uniform_Resource_Locator.

[21] World Wide Web Consortium (W3C). 2004. *Architecture of the World Wide Web, Volume One*. wpid:archive.org:2016-04-04T213830Z:https://www.w3.org/TR/webarch.

[22] World Wide Web Consortium (W3C). 1997. *Date and Time Formats*. (W3C profile of ISO 8601). wpid:archive-it.org:2016-03-31T232655Z:http://www.w3.org/TR/NOTE-datetime.

[23] Zittrain, J., Albert, K., Lessig, L. 2014. *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*, doi:10.1017/S1472669614000255.