# Citing Web Addresses That Last

## How to Recognize and Use Persistent Identifiers

by Ian Watson, PhD

Not all web addresses are created alike. Some are designed to last, and these are the most useful for genealogists who want to use web addresses in their citations.

We've all experienced broken web addresses that lead nowhere. Media researchers speak of the problem of "link rot." One study showed that only about half of the web addresses cited in United States Supreme Court decisions still work.[1]

Web gurus have been honing ways to increase web addresses' durability since the 1990s. Today, the most common term for web addresses that are intended to last as long as humanly possible is "persistent identifier" (PI).[2] You'll also see other, basically synonymous, terms like "permanent URL."

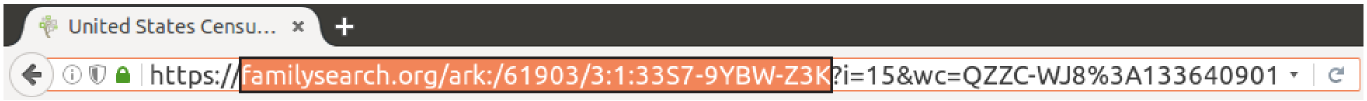
In the past, genealogists have been reluctant to use web addresses in source citations. But as PIs become more widespread and their workings better known, there are good reasons for us to use them. Above all, they take us to the original source of a statement with a single click. Most are short enough to fit passably on a printed page.

Three curated schemes set the gold standard for persistent identification: ARKs (Archival Record Keys), DOIs (Digital Object Identifiers), and the Handle System. Each of these schemes is centrally administered by a responsible institution and the identifiers are designed to last a very long time. You can use ARKs, DOIs, and the Handle System knowing that they're backed by an explicit commitment by the website owner to keeping the document available and the forwarding links updated.

Beyond these three methods is a second tier of PIs which site owners promise will last, but which lack a central authority and a strong guarantee of durability. A third tier includes web addresses with no guarantee of permanence at all, but which are concise enough for a clean citation and have at least certain elements which are likely to persist.

[1] Jonathan Zittrain, Kendra Albert, and Lawrence Lessig, "Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations," *Harvard Law Review* 127, no. 4 (February 2014), accessed 7 November 2017. www.harvardlawreview.org/2014/03/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations.

[2] Several decent introductions to persistent identification are available on the web. Try Juha Hakala, "Persistent identifiers - an overview," *Technology Watch Report: Standards in Metadata and Operability* (2010), accessed 7 November 2017. www.metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview.



United States Censu… ×

https://familysearch.org/ark:/61903/3:1:33S7-9YBW-Z3K?i=15&wc=QZZC-WJ8%3A133640901

**Figure 1**: When citing an ARK, don't include anything after the question mark. The "http://" can also be left out. The highlighting shows the parts you should copy.

## First Tier Persistent Identifiers

### ARKs

Many nonprofit organizations use ARKs to identify their digital collections. For genealogists, the most important of these is likely FamilySearch.org. Here's an example of a 1910 census page:

**familysearch.org/ark:/61903/3:1:33S7-9YBW-Z3K**

The **ark:** identifies it as an ARK, while the **61903** is FamilySearch's ID within the ARK system. The **3:1:** is FamilySearch's internal code for digital artifacts like this US 1910 census page image. Roughly a dozen letters and numbers follow, which identify the exact document. You shouldn't leave off the **familysearch.org** at the beginning, as an ARK has to be sent to a specific website to be retrieved, or "resolved" in more technical terms.

How long will FamilySearch ARKs last? "Hopefully . . . forever," according to a comment by FamilySearch's head information architect.[3] Even if FamilySearch's domain name changes at some point in the future, we should still be able to submit the same ARK to the new resolving site.

FamilySearch ARKs are a handy tool for citing original documents like census and probate records. An advantage is that FamilySearch ARKs refer to specific pages. This makes it easy to pinpoint specific parts of poorly numbered or labeled source documents. Once you get familiar with the structure of a FamilySearch ARK, you can successfully note the location of a record just by writing down the alphanumeric code at the end. FamilySearch often includes the ARK identifier in the suggested citation, both on the indexed page for the entry, and in the information tab on the image page.

When copying a FamilySearch ARK out of the address bar of your browser, look for the question mark and copy only what comes *before* it (highlighted in figure 1). The part of the web address after the question mark contains information related to your session that isn't necessary for a record citation.

The ARK system is administered by the University of California's California Digital Library (www.cdlib.org).

### DOIs

DOIs are most commonly used in the for-profit publishing sector. Publishers assign them to academic journal articles, books, and book chapters. Few, if any, genealogy publishers use them. Typically, the DOI leads to a sort of cover page, and another click leads to the often-paywalled full text. Here's a DOI for a chapter in an edited collection:

**dx.doi.org/10.1057/9781137367310.0014**

The DOI's first set of numbers (here, **10.1057**) identifies the publisher. What follows is a code that identifies the item among everything that publisher has produced. This publisher chose to assign each chapter of the book a code starting with the book's ISBN. To check whether there is a DOI for a given article or book title, use the search box provided at crossref.org, the largest DOI registrar. The system is managed by the International DOI Federation (www.doi.org).

### The Handle System

HathiTrust, the digital library run by a consortium of universities, uses the Handle System to identify scanned book volumes. Handle System addresses look like this:

**hdl.handle.net/2027/wu.89062938204**

Here, the **2027** identifies HathiTrust, while **wu.89062938204** identifies the volume. When your browser displays a page from a volume at HathiTrust, the Handle System address for that volume is shown in a box to the left of the page. It's labeled "Permanent link to this book."

Many university and government repositories also use the Handle System. As your browser loads a page using a Handle System address, the address will disappear and be replaced by the actual "target" address of the resource. This is also true of DOIs, but not always of ARKs.

The Handle System is managed by a Swiss organization called the DONA Foundation (www.dona.net).

## Second Tier Persistent Identifiers

A second tier of persistent identifiers are generated by websites as durable labels for their own content. These PIs are not centrally managed, so their longevity depends entirely on the health and policies of the website involved. They probably won't last a

[3] Randy Wilson, comment on FamilySearch support, "PALs that aren't PALs for long?" circa 2014, accessed 7 November 2016. www.getsatisfaction.com/familysearch/topics/pals_that_arent_pals_for_long_a_problem_with_some_links_in_massachusetts_land_records#reply_14086360

hundred years, but for near-term citation purposes, they're fine. These are sometimes called "permalinks" to distinguish them from centrally managed PIs. Here are some contexts where permalinks can be useful for genealogists:

**Book titles at WorldCat.** For books that haven't been scanned, WorldCat's permalinks are a useful way of pinpointing a specific title or edition. From the details page of any book at worldcat.org, you can click on "Permalink" on the upper right-hand corner of the screen, to display the permanent address. The link's "active ingredient" is the serial number from the OCLC bibliographic database:

**www.worldcat.org/oclc/15603049**

**Historical newspapers.** Some historical newspaper websites give concise, durable web addresses for each scanned newspaper page. Here's a front-page report of the Apollo moon landing from the New York State Historic Newspapers site. Look closely to see information that can, in a pinch, identify the page traditionally, including the Library of Congress control number (**lccn/sn84031165**), the date (**1969-07-21**), the edition number (**ed-1**) and the page number (**seq-1**):

**www.nyshistoricnewspapers.org/lccn/**
**sn84031165/1969-07-21/ed-1/seq-1/**

**Blog posts.** The most popular blog platform, WordPress, assigns a permalink to each blog post on a site. Of course, if the blog owner ever changes domain name or blog software, or goes offline altogether, these links will go dark. Here's an example from a popular genealogy blog:

**www.geneamusings.com/2014/04/should-we-put-digital-image-urls-in.html**

**FamilySearch PALs.** Before FamilySearch settled on ARKs for their records, they used homemade permalinks that start with **pal**. You still see PALs around and they still resolve, but they are now deprecated, which means that the system still supports that scheme, but a newer, preferred convention (ARKs) has been introduced. So avoid using a PAL in favor of the corresponding ARK. They look like this:

**familysearch.org/pal:/MM9.3.1/TH-1951-25146-9109-9**

## Third Tier Persistent Identifiers

Web addresses in this category contain no explicit promise of permanence from the website owner. However, they are concise and easily citable, and contain elements which are likely to persist.

### Persistent Identifier Lingo

- *ARK*: A centralized persistent identifier system, used especially by nonprofit and public-sector archives, managed by the California Digital Library.
- *DOI*: A centralized persistent identifier system, used especially by for-profit publishers and scholarly journals, run by the International DOI Foundation.
- *Handle System*: A centralized persistent identifier system, used by HathiTrust and others, run by the DONA Foundation.
- *PAL*: An internal permalink system formerly used by FamilySearch, but now deprecated in favor of ARK.
- *Permalink*: Typically, a persistent identifier that is generated by a website for its own pages and isn't centrally managed.
- *Persistent identifier* (PI): A string of symbols that refers to a resource, allows users to locate that resource, and is designed to last a long time. These days, PIs most often take the form of a web address.
- *Web address*: Also known as a URL. This is what appears in your browser's address bar, typically with an http:// prefix.

**Book scans from the Internet Archive.** Here's the Internet Archive's address for a particular scan of a book on eighteenth-century Pennsylvania:

**archive.org/details/fortsonpennsylva00hunt**

It's natural to be a little skeptical about whether the word **details** will continue to be part of this address for decades to come. However, at least the resource ID (**fortsonpennsylva00hunt**) is likely to endure. Note that you could also replace the word **details** with **stream** or **download**, which gets you to a different interface for the same resource.

Actually, the Internet Archive has assigned ARKs to all its holdings, but doesn't advertise them in any way. They're buried in a metadata file and have to be resolved through a separate website, n2t.net. For this book, the ARK is **n2t.net/ark:/13960/t5k948z0x**. Most people prefer using the archive.org/details web address.

If you do want to find an Internet Archive ARK, start from the details page, locate the "Download Options" box, click "Show All," open the file ending in **_meta.xml**, and search for **identifier-ark**. Alternatively, call up **archive.org/metadata/resource-ID** and search for **identifier-ark**.

**FamilySearch Memories.** For family photographs and other user-uploaded media, FamilySearch doesn't use ARKs. Here's a link to a user-submitted photo:

**familysearch.org/photos/artifacts/26293118**

FamilySearch doesn't guarantee that the structure of these links will stay the same, although we might reasonably guess that the ID number at the end will persist.

## Other Ways to Cite Web Addresses

### Make Your Own PI

The site Perma.cc allows anyone to have a particular web page archived and permanently identified at no charge for the purpose of citing it in a legal or scholarly work. For example, I created a Perma.cc copy of the comment that I cited above about the longevity of FamilySearch ARKs:

**perma.cc/VJ66-HDLS**

The catch is that Perma.cc can't directly archive web pages that are behind login barriers (like most of those at Ancestry.com). Perma users can work around this by uploading their own snapshot of a page, but doing so may run afoul of Perma's (and the original site's) terms of service. Also, unaffiliated users are limited to ten pages a month. Perma.cc was originally created at Harvard University for courts and legal scholars and is now managed by a large group of university law libraries.

### Use Traditional Citations, Too

The authors of the ARK specification explain that "persistence is purely a matter of service." They mean that persistent identifiers are only persistent insofar as someone commits to supporting them.[4] PIs' durability depends on social institutions more than on technology. There is no way to ensure that a PI will be followable a century from now. We can only maximize the probability.

That's why it's always preferable, when writing for posterity, to describe your source in words (with author, title, publication info, and date) as well as pointing to it with a web address. Giving both types of information allows users a backup if the web address fails. It's also more informative, because traditional citations say more than web addresses about the nature and quality of a source.

Of course, there are also many contexts where it is completely appropriate to use just a PI with no further explanation, such as notes to ourselves or short-lived emails.

### Volumes versus Pages

When reading a scanned book at HathiTrust or the Internet Archive, the web address you see in the browser's address bar works as a link to the exact page you're viewing. It's handy to be able to zero in on a particular page. However, the addresses for pages are less reliably persistent than the addresses for the work as a whole. To stay absolutely on the safe side, especially in a print publication, cite the PI for the entire work and give the page number separately, even if the reader will need more time to find the page. I do use the precise page address if I'm just noting it down for myself or a colleague, though.

These days, I like to give both a concise traditional citation and a persistent web link in my footnotes. For example:

1910 census (roll 1662), Ward 12, Seattle, King Co., Washington, ED 204, sheet 8B (**familysearch.org/ark:/61903/3:1:33S7-9YBW-Z3K**).

Including PIs in a formal report can obviate the need to attach full copies of original documents, saving time and space.

In blogs and other online publications in HTML format, using a PI as a link target lets readers reliably reach your source with one click. You can write "Mary Jones appears in this 1910 census record" and link a PI to the underlined words so that the reader can click on them to bring up the source. If you don't want to clutter the text with a footnote or parenthetical citation, but would still like to fully describe the source, you can have a box with a traditional citation appear when the reader hovers over the anchor. To do this, put the citation in the link's "title" attribute. Here's an example of how the HTML code might look:

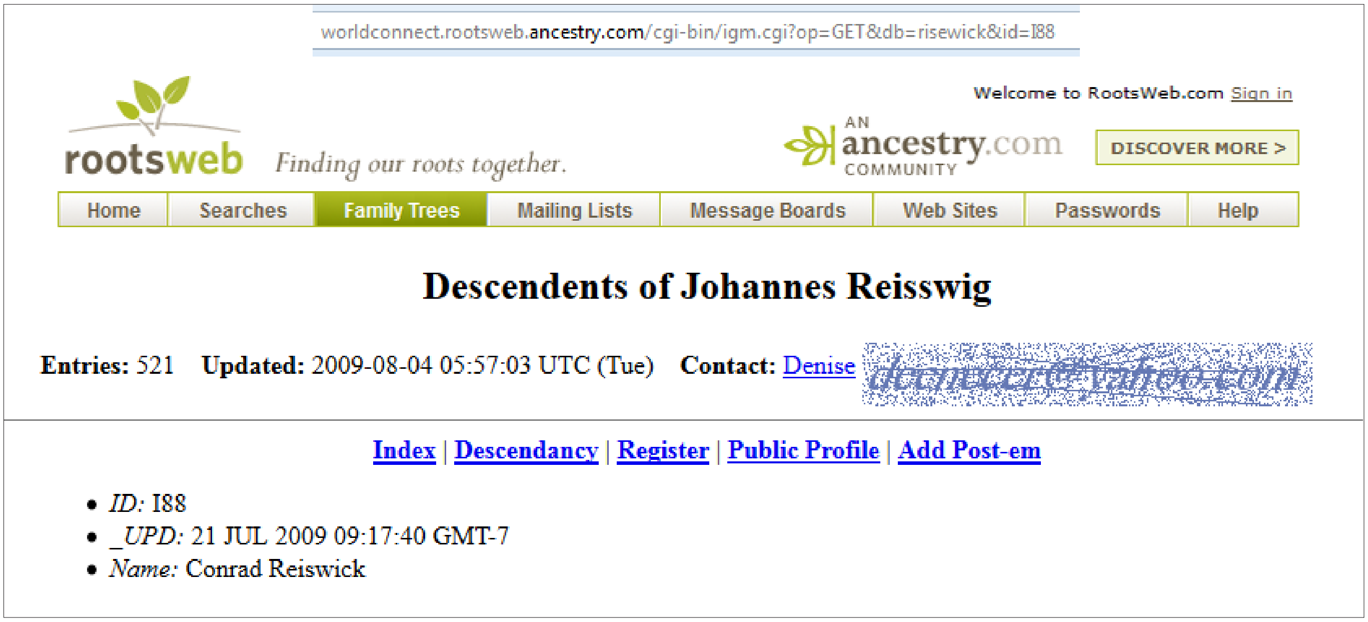
```
<a href="https://familysearch.org/
ark:/61903/3:1:33S7-9YBW-Z3K" title="1910 cen-
sus (roll 1662), Ward 12, Seattle, King Co., Wash-
ington, ED 204, sheet 8B">
```

### When There Is No PI

Some genealogy websites, including Ancestry.com, unfortunately don't provide their users with persistent identifiers. In these cases, lay out a so-called "breadcrumb trail," which allows a user to recreate the process of navigating to a record. A greater-than sign separates each point along the trail. Thomas W. Jones wrote about these types of citations in the June 2016 issue of the *APGQ*.[5]

[4] John A. Kunze and R. P. C. Rodgers, "The ARK Identifier Scheme" (2008), section 1, www.cdlib.org/services/uc3/arkspec.pdf.

[5] Thomas W. Jones, "Genealogy Waypoints: An Option for Locating and Citing Unindexed Numbered Online Images," *APGQ* 31, no. 2 (June 2016), 71–79. An interesting designer-side resource on breadcrumb trails is Jacob Gube, "Breadcrumbs in Web Design: Examples and Best Practices," *Smashing Magazine*, 17 March 2009, www.smashingmagazine.com/2009/03/breadcrumbs-in-web-design-examples-and-best-practices.


worldconnect.rootsweb.ancestry.com/cgi-bin/igm.cgi?op=GET&db=risewick&id=I88
rootsweb
Finding our roots together.
Welcome to RootsWeb.com Sign in
AN ancestry.com COMMUNITY
DISCOVER MORE >
Home
Searches
Family Trees
Mailing Lists
Message Boards
Web Sites
Passwords
Help

**Descendents of Johannes Reisswig**

**Entries:** 521 **Updated:** 2009-08-04 05:57:03 UTC (Tue) **Contact:** Denise [illegible]@yahoo.com

Index | Descendancy | Register | Public Profile | Add Post-em

- *ID:* I88
- *_UPD:* 21 JUL 2009 09:17:40 GMT-7
- *Name:* Conrad Reiswick

**Figure 2**. The url for this RootsWeb page is long and complicated. Using a waypoint citation and describing the website location within the narrative might be more practical.

Many websites include session info and database queries at the ends of the web addresses displayed by your browser. They usually include characters such as ?, =, and &. Session info and database queries make the web address quite long. Whether or not the website provides a PI, resist the temptation to paste the entire length of web addresses like these into your citations. They are too long to copy by hand and look awkward on a printed page. Web addresses which rely on session and database details are anything but persistent. They sometimes break when website software is updated. In the worst case, they will stop working as soon as you log out from the website or close your browser. Truncate the web address down to its persistent core and, if needed, describe in words how you get from there to the resource you're citing.

For example, instead of citing this web address,

**worldconnect.rootsweb.ancestry.com/cgi-bin/igm.cgi?op=GET&db=risewick&id=I88**

a more durable approach would be something like:

RootsWeb WorldConnect Project Family Trees at rootsweb.ancestry.com, Risewick database, entry for Conrad Reiswick, born between 1745 and 1752 (number I88).

Even more durable, and more concise, would be to archive this page at Perma.cc.

## Key Points to Remember

- Direct one-click web links to cited material are especially appropriate in online publications, in blogs, in notes to ourselves during our own research, and in material that is designed to be read and used relatively soon. They can be useful in paper publications, too, as long as we minimize their length and maximize their persistence.
- Using an ARK, DOI, or Handle System identifier is the most durable, and therefore best, way to use a direct link to cite a web-based resource. Homegrown permalinks are useful but are likely to be less long-lasting.
- Practice recognizing and stripping off unneeded elements from the end of a web address. Avoid putting session info or database queries in a citation.
- Unless there's a reason not to, balance the user convenience of a one-click path to a record with the durability of a more traditional citation.

Technology continues to change. As the world of information becomes ever more digitized and web-based, newer and more efficient ways of source citation are sure to emerge.

*Ian Watson, PhD (www.ianwatson.org) has been involved in genealogy since his teens and teaches information architecture at the Norwegian University of Science and Technology.*