Reliable Links *

Yury Andreev
andreev.yurij@gmail.com

John doesn't maintain that file any more, Jane does.
Whatever was that URI doing with John's name in it?
It was in his directory?
I see.
"Cool URIs don't change", w3.org[11]

Introduction

Millions of links appear every day in the Web and millions of links become broken, leading to different query errors (e.g. widespread 404-error). This reflects the dynamic character of the Web, but it becomes a difficulty when we wish to keep the content spread over the Web together.

The author had a favourite collection of videos on video hosting (Youtube). After a year, more than a half of these videos were no longer accessible. The same illustration could be presented using any of the "collection services": it could be collections of bookmarks (in a browser, or in a social bookmarking service), collections of items in social media, and so on.

Let's say that *a link is a* ***reliable link*** *if it referring to content*[1] *which remains the same and remains accessible by the link.*

There exists a more technical definition in terms of link traversal. ***Successful link traversal*** *generally means finding a resource with perfect precision and recall, and retrieving an authentic representation of the resource in a timely fashion, i.e. with sufficiently low latency.* For more details, see [3].

In the last part of the Introduction, we consider the problem of broken links more precisely. Next, we consider examples, and analyze, and rank links by their reliability. Finally, we formulate rules to help us to make links more reliable.

*The current work was presented on Free Society Conference and Nordic Summit, November 6-8, 2015; The first edition was published in [1].

[1]Despite recommendations in [2], we will use the term *content* in the present article. Content is the information, such as text, video, and sound, that constitutes a publication or document, which is ready to transfer or could transfer potentially. It is interesting that if the word "content" is a loan word in a language other than English (e.g. Russian), it has just this special meaning.

Dynamics and statics

The reliability of links plays an important role in software. Projects have many *external dependencies*, represented by links: remote interfaces, web services, web widgets, Content Delivery Networks, Domain Name System, Search Optimization. An increasing number of interaction modes are being used in projects every day. By external dependencies we mean links to material which is not under control or ownership of the project maintainer. Control over these parts of the project should be arranged in software apriori or delivered to the technical support, it is an object of special research.

These technical cases, as well as technical aspects of information transference and retrieval, will not be considered here. *The present article is devoted to the static content.*

Broken links could be considered as a natural filtration mechanism, but it is easy to throw the baby out with the bathwater.

Technical and scientific publications, as an important class of the static content, may contain URL links (more and more each day!). The right to verify scientific results implies access to its background and previous results. (The text of any program was considered as a scientific article in the beginning [4]. And the right to study and modify software, indeed, reflects this scientific approach to software development.) Therefore, the original material mustn't ***lose integrity.*** The most famous example where integrity was lost is in Fermat's Last problem: This theorem was first conjectured by Pierre de Fermat in 1637 in the margin of a copy of Arithmetica where he claimed he had a proof that was too large to fit in the margin. His claim was discovered some 30 years later, after his death, written in the margin of a book, but with no proof provided. The first successful proof was released in 1994. Analogous situations may happen nowadays. Centre de Recerca Matematica (CRM, Barcelona) erased all preprints dated before 2006 from its site. (*The Engelbart's "Library System" requirement for an open hyperdocument system[5]* was not met.)

By a *link* we mean a URL link or a URL link equipped with additional information (like a point in the References section). Both, URL and URI are used synonymously throughout the paper.

Keeping and searching

If a link is broken we have to search corresponding material. *Searching is a contiguous problem.* It's always a good idea to keep searching in mind when you make or choose links (see Links Good and Bad).

Also, sometimes the inverse problem arise: find a link by a bit of *exact* information. A unique string of text could be sufficient.

Historical overview

In addition to a force majeure situation (war, fire, flooding), there is a one additional natural contributer to loss of content, namely lack of interest. A sig-

nificant amount of data is lost each time the ***information carrier is changed***. In ancient times, for example, many texts which were not of great interest to the people of that time were never copied from papyrus to parchment codex [6]. Moreover, because parchment was a costly material, some was reused, and thereby ancient manuscripts which were in a good state of preservation have been overwritten.

This process was on display throughout the 20th century. There have been numerous types of data storage devices: punch cards, magnetic tapes, diskettes, optical disks, and others. Each method replaced the last, and now it is difficult to read even a 3.5" diskette. Also, unlike the situation before the electronic age, we can no longer read information directly from the carrier. We need a mediator, or reading device. The problem of ***physical access*** arises. (There exist enthusiasts, who collect obsolete hardware and keep it in use.) Of course, physical access to the library always has been a problem.

In the digital world it is easy to save information by making copies[2], but issues of ***accessability and persistency*** begin to play an important role.

Indeed, even in the early beginning of the electronic age the Xanadu concept was described [7]. Particularly, it was an attempt to get accessability and persistency[3], because "the World Wide Web allows nothing more than dead links to other dead pages". Nowadays, new concepts regarding how to transform the URL Scheme appear. A list of the links to these ideas can be found in [8], but... every link is broken in the list at the moment.

Control over permissions. Examples

An important contributer to reliability of the link is a control over permissions. Let's consider several examples.

- Any "collection service" as mentioned in the Introduction is a good example. An owner may delete material or change permissions at will and we will not be able to obtain access to it anymore. Material may be blocked by service as well.

  Note only, that owner may set permissions incorrectly or may miss policy updates. This can be a cause of restriction reinforcement or leaks.

  Accessability depends on availability of the entire service, server or network.

- The author had a nice mailbox at `london.com` (served by `mail.com`). This mailbox does not exist anymore, because mail server was closed by company. In fact a paid service had been offered there! Persons who had given this address in their contact information became unreachable.

[2] However, information still may be lost in part (name of the source, the author's name, dates, etc.).

[3] One could find an analogy between hypertext and HTML. The first allows incorrect links, the second is not strict and allows invalid documents.

- Even a big IT corporations do not provide reliability. About discontinuous projects see, for example, [9].

  There were a few funny stories about Microsoft. In December 1999, Microsoft forgot to renew the domain name `passport.com`, and so rendered its Hotmail service partially crippled. As if that wasn't bad enough, in 2003 again Microsoft forgot to renew an important domain `hotmail.co.uk`. The new owner proceeded immediately to contact the giant, however, Microsoft only took notice when The Register contacted the company to enquire why its site was registered to a private individual.

  Video hosting from Yandex was closed; cloud service Ubuntu One was closed, however they notified users a long time before.

Government control

- Github was totally inaccessible in Russia for a couple of days last year. It was totally blocked[4] by government organization because one of the users uploaded inappropriate content.

Undesirable information and reliability

Before we continue, let's look at some cases where reliability is not desirable.

The social service `snapchat.org` became popular because it provided a mechanism which worked with undesirable information. Using the application, users can take photos, record videos, add text and drawings, and send them to a controlled list of recipients. These sent photographs and videos are known as "Snaps". Users set a time limit for how long recipients can view their Snaps (as of September 2015, the range is from 1 to 10 seconds), after which Snapchat claims they will be deleted from the company's servers. Of course there is no warranty that this content will be deleted from the servers. Indeed, on May 9, 2013, Forbes reported that Snapchat photos do not actually disappear, and that the images can still be retrieved with minimal technical knowledge after the time limit expires. And of course images may be recorded by screenshots as well.

In general, detractive information could be considered under the Right to be Forgotten, and it has become possible to request removal from a search engine.

On the other hand, services delete some kinds of information because of a violation of terms of service without asking for permission. *Note*, that some of this information is better to keep saved for the future, to let future generations know our mistakes.

Micro-reliable links. Another problem may happens with one-time links. These links, pointing to some sensitive information, can not be valid after first use. For example, see [10]. But links should be valid (and thus reliable) before

[4] it was not possible to block a part of it because of *https* protocol

first use by proper recipient! There is a chance, that software used for communicating goes ahead, or some web crawler may come first. Initialization lists of links on distributed nets should be valid until first use as well.

Link's classification: Triangles of links

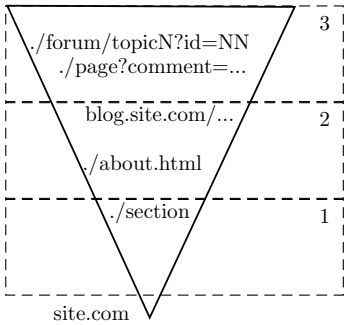
Keeping in mind "Controls over Permissions", let's try to compare reliability of the links approximately.

The most reliable links are links referring to the same material.

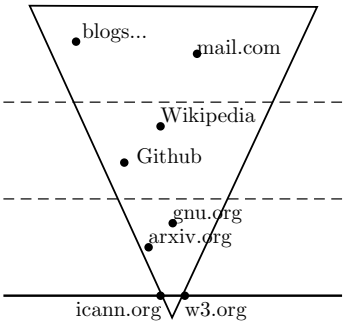
However, if the document consists of many pages, all of the pages should be on their places.

We can place links within one site in the triangle diagram (pic. 1).

- Let's put the root of the site at the bottom of the triangle as the most reliable within this site.
- Let's put links pointing to the permanent information of the site at the first level. This corresponds to the materials editable only by an administrator, or other person granted special access.
- Links to user generated content belongs to the upper (least reliable) deck. Users may delete or modify this content at will.



pic. 1



pic. 2

If we find center of mass of each triangle, we can draw another triangle diagram reflecting dependencies between sites (pic. 2).

- The most reliable links are the links to organizations, which manage and maintain the Web, and develop standards: `icann.org, w3.org`, `example.com` ...
- The next level consists of huge societies, `gnu.org, ctan.org`, huge integrators (archives) like `arxiv.org`, sites with special top level domains, `*.edu, *.gov`, *some* services of big corporations, `gmail.com`.

- Sites which allow collaborative modification of their content (like Wikipedia), have a "version problem"; publications on news site might be moved to an archive. We have to be accurate with these situations (additional information about a link may help). On the other hand, modification of content is not a problem for a Github. Usually, we just need to know about existence of a project. *Moreover,* modifications indicate a project's popularity (and thus reliability). We say that those links suppose a low ***reliability level***.

- Personal blogs and links to the publications on social media will be at the top of the diagram. However, some materials, on the personal site especially, might be of value to author, thus it might be more reliable (if he don't forgot to continue domain or other service at in time). *Also,* reliability of the main (root) link to the personal account on social media seems equal to the reliability of the email address.

Links good and bad

Let's describe how to make links more reliable. Also, there is an interesting article "Cool URIs don't change"[11][5], describing how to make links more reliable *on a server*.

Readable and meaningful links

The most reliable links have the property that even if access is lost we have a good chance to retrieve the content we desire. Let's describe a search process if the typical URL link

`http://<site-name>.tld/<section>/<page-name>?<query>`

is broken. If other information is not provided, or redirection is not given, the following possibilities remain.

1. Search information on the accessible pages on site (moving toward the top of triangle, pic.1)

- `http://<site-name>.tld/<section>/<page-name>`
- `http://<site-name>.tld/<section>/`
- `http://<site-name>.tld`

2. Use "search"

- search `<section>, <page-name>` on the web site if possible

[5] A reader might notice that corresponding link should be placed almost at the top of the right triangle (pic.2). Thanks to Alexei Khlebnikov for mentioning this article.

- search `<site-name>, <section>, <page-name>`; also, search by author name or any additional information if available
- general search the web

3. Searching traces of information:
   - wayback machine `archive.org`
   - caches of search systems.

However, Caching depends on searching engine, frequency of indexing, setting in `robots.txt`; Time of snapshot doesn't depend on actual changes at the web site. This caution is written on `archive.org`. Also, it doesn't save all kinds of content (video, for example).

Thus, it is better to give readable and meaningful names to `<site-name>, <section>, <page-name>`. This is a rule of SEO as well.

As an example, if you follow the link on Karl Fogel's site `http://www.red-bean.com/kfogel/948.html` you can read:

"You actually followed a link named 948? What did you expect to find? Its prime factors are 2, 3 and 79, now go home."

Integrators

Information collected, stored or distributed by libraries seems more reliable in comparison with all of the rest of the digital world, except in the case of force majeure situations. Once, the author read a journal (Bulletin of the Russian Academy of Sciences) dating to the XIX century, which had survived both fire (year 1824) and flood (year 1988).

There is a practice, that central libraries are collecting a sample or several samples, of every book which has been published within country. Also, libraries are making digital copies. (This is one more example of changing of carrier.) Moreover, there exist organizations devoted to *coordination of digitisation, digital preservation and digital access to cultural heritage* [12].

Big integrators, such as digital libraries and archives, arise. Some of them use the same logic that PGP public key servers have used. The *HAL* archive notifies: *"Any deposit is definitive, no withdrawals will be made after the on-line posting of the publication."*[13] Articles that have been announced and made public cannot be completely removed from *arXiv*. *"You may submit a withdrawal notification for your article... You must provide a specific reason for the withdrawal... arXiv makes all previous versions of submissions publicly available... even though the current version of a paper may be marked as withdrawn, previous versions can still be retrieved."*[14]

Therefore, integrators like these above have been found by society to be sufficiently reliable.

Additional information

Additional information about a link, such as full name of the material, version, last viewed date, could be helpful. *However*, some of the information may be the cause of an irrelevant search. For example, ISBN, and DOI may lead to a publisher rather than to original material.

Alternatively, putting any information (excepting the date) in the URI itself is asking for trouble one way or another [11].

Shortened links

Shortened links is an option to make links better. Some of the link shorteners, like `http://tinyurl.com/`, allows one to give a readable name to the part of the shortened link. Big projects usually provide their own link service for several reasons. Debian has it at domain `deb.li`. There is an open source project to make your own service [15].

However, a shortened link is just a link to a link, and it cannot make the original link more reliable. Also, it depends on service reliability itself.

Future insight

The format of URI is not static in general, it changes and extends. This should be kept in mind. Recently, new Top-Level Domains were introduced. The Internationalized Domain Names (IDN) concept is developing. Particularly, cirillic URI, like `http://xn--j1ail.xn--p1ai/eng/` (looks meaningless, doesn't it?) appeared. The *New gTLD* program was opened. [16]

Reference rules. Summary

- Choose more readable and meaningful links if possible
- Choose more reliable source if possible (integrators)
- Provide additional information
- Provide both ordinary and e-version if each exists
- Provide ***structure links*** if necessary:

  site name: section name – subsection name – ...

  Sometimes, the original link looks unreadable or time-dependent. The logic structure of the site is more reliable.
- Make shortened links if necessary
- Links should correspond to the receiver. For example, links lead to English-speaking sources in the present References section, because this article is devoted to international audience. Also, parts of URI, like `/en/`, should

be included/excluded. For the same reason, links should not depend on software. Some of the installed programs could change links. A browser extension Wikiwand is currently doing so.

- Links must to be precise. Otherwise an additional search will be needed.
- Sometimes it's easier to include material itself rather than include the proper link referring to it. (Material is widespread, it is easy to search.) The author did so in the present article.
- Use common sense and ask a question: whether it is too important to worry about? *However*, importance of the referred material may be underestimated by author.

The author hopes this article will help to make links more reliable.

Acknowledgements

The author would like to thank to Dmitry Kostyk, Vladimir Molokov, Trevor Richards, and participants of the Linux Vacation Eastern Europe conference (`lvee.org`), and Free Society Conference and Nordic Summit (`fscons.org`) for meaningful discussions and comments.

References

[1] Yury Andreev, "Nadjozhnye Ssylki" ["Reliable Links"] (in Russian), Otkrytye tehnologii, pub."Al'ternativa", Brest, Belarus, 2015

[2] Richard Stallman, "Words to Avoid (or Use with Care) Because They Are Loaded or Confusing", `http://www.gnu.org/philosophy/words-to-avoid.html`

[3] "Link Reliability - Why URNs are Not the Answer", `https://www.w3.org/Propagation/reliable-links.html`

[4] Karl Fogel, "What Is Free Software", `http://www.onlamp.com/pub/a/onlamp/2005/09/29/what-is-free-software.html`

[5] "An Evaluation of the World Wide Web with Respect to Engelbart's Requirements", `https://www.w3.org/Architecture/NOTE-ioh-arch#library-system`

[6] Wikipedia, "Manuscript", `https://en.wikipedia.org/wiki/Manuscript`

[7] Project Xanadu, `https://en.wikipedia.org/wiki/Project_Xanadu`, `http://xanadu.com`

[8] Catalog Numbers: URNs, `https://www.w3.org/Addressing/citations`

[9] Discontinued Google Services, `https://en.wikipedia.org/wiki/Category:Discontinued_Google_services`

[10] One Time Self Destructing Links for Sharing Sensitive Information, `https://1ty.me/`

[11] "Cool URIs don't change", `http://www.w3.org/Provider/Style/URI`

[12] Digisam, `http://www.digisam.se/index.php/en/`

[13] HAL archive, `https://hal.archives-ouvertes.fr`

[14] To Withdraw an Article at arXiv, `https://arxiv.org/help/withdraw`

[15] Open URL Shortener Source, `http://rod.gs/about/sc`

[16] Non-Latin Domains, `https://newgtlds.icann.org/about/idns`