Capturing the Web at Large
A Critique of Current Web Referencing Practices

Caroline Nyvang, PhD., researcher at the Royal Danish Library, cany@kb.dk
Thomas Hvid Kromann, PhD., researcher at the Royal Danish Library, thok@kb.dk
Eld Zierau, PhD., IT consultant at the Royal Danish Library, elzi@kb.dk

INTRODUCTION

The Internet and the cultural phenomena that exist online are increasingly attracting academic awareness, and e-materials both supplement and replace physical materials. These new opportunities come with a range of challenges. Websites are connected in new and unfamiliar ways, the amount of data easily surpasses what we have experienced previously, and we do not yet have an infrastructure that can lend prober support to the increased scholarly use of web resources [1-2]. This paper is an attempt to grapple with one of the core challenges, namely our ability to provide precise and persistent references to web material.[1] The paper charts prevailing ideals and practices regarding web references within the Humanities. We highlight the challenges based on an analysis of web references in two case studies – a selection of Danish master's theses from 2015 and academic books on contemporary Danish literature. We propose a new best practice that is consistent with good scientific practice in terms of both precision and persistency, which cannot be obtained following the existing standards.

IDEALS

Citations and bibliographies ensure that research results can be properly reproduced and credited, and each scientific field seems to have a predilection for a particular style. In the final decades of the 19th century and the first half of the 20th century, a number of international standards were established, including the Chicago and the Harvard System as well as APA [3-4]. These are the most widely used among the journals that publish the highest number of peer-reviewed articles from Danish scholars [5, pp. 2-3].

Although there is not one conformal system, some basic principles do apply to all styles. These can be summarized with Skov's minimum requirements to bibliographical references:

- Reference must contain all bibliographical elements needed to univocally identify and retrieve a document.
- References must be immediately understandable in the sense that individual elements appear in a logical order that reflect their level of importance in the retrieval process.
- References must be produced consistently! [6]

[1] This paper builds on data that are partially presented in [5].

However, when we want to make references to online material, we fall short on the above measures. The most widely used citation styles, e.g. Harvard and Chicago, do not request information beyond URL and date, and the recommended styles do not offer the minimum of bibliographical information that researchers are held individually accountable for [7, p. 11]. This suggests that the common citation styles – throughout the centuries fine-tuned to handle analogue materials of different kinds – do no adequately support web references.

APPROACH

Since the propagation of the Internet, researchers have been concerned with the viability of websites, domains and individual web objects [8-10]. In 2002, Koehler described the results of a study that followed 361 URLs during the course of four years [11]. He concluded that during that time span, around 67% of the URLs – due to one reason or another – became inaccessible. Koehler's study was based on randomly selected websites, but inspired a range of subsequent investigations probing the extent of the problem within distinct research fields such as History [12], Information Science [13-15], Law [16-17] and Medicine [18-20]. Results vary according to the types of links investigated and the timespan of the study, but the conclusions are clear: Websites are updated at an increasing speed, and URLs are disturbingly unstable.

Our paper is inspired by these studies, but differs in its focus on both practices and ideals. We combine a qualitative and a quantitative approach, which give an impression of how students and researchers incorporate online materials in their work as well as a snapshot of the stability of the accompanying web references. In two corpora – 35 Danish master's theses from 2015 and academic books on contemporary Danish literature – we have identified all references containing "www" and/or "http/https" in the body text, footnotes, bibliography and appendices. By following these URLs, we have examined whether the individual link still produces a hit in the web at large. Although the material only covers a small share of an entire academic field, it enables us to identify to some overarching tendencies and problems regarding the use of web references.

PRACTICES AMONG MASTER'S STUDENTS

To get an understanding of how students incorporate Internet resources into assignments, we have analysed a selection of Danish master's theses uploaded to DISKURS [21]. DISKURS is a student assignment database managed by The University of Copenhagen. As of early February 2016, it contained around 3500 master's theses that have been voluntarily uploaded under a Creative Commons licence (by-nc-nd). In this paper, we focus on the 35 theses handed in at the Faculty of the Humanities in 2015. This corresponds to 3.3% of the total number of passed theses with the faculty.[2]

[2] Personal correspondences with coordinators Sanne Eyrich (January 16 2017), Annemarie Hede-Andersen (January 13 2017), Jeanette Sporon-Fiedler (January 12 2017), Birgit Hüttmann (January 12 2017), Nina Simkunas (January 13 2017) and Lis Lachtane (January 12 2017).

The 35 examined theses have a total of 707 unique references, which contain ”www” and/or ”http/https”.[3] These appeared in the body texts, footnotes, appendices and bibliographies. On an average, each thesis contained 20.8 web references, but there was a marked variance. One thesis stood out, as it did not contain a single reference to web resources, whereas three other theses all had 75 unique web references. The theses cover a wide array of empirical topics – ranging from Shakespeare to social media – but the amount of web references does not seem to correlate with particular topics. This confirms a widely accepted hypothesis: Namely, that the web provides resources to a rather large portion of Danish students, even when their field of study is not the Internet.

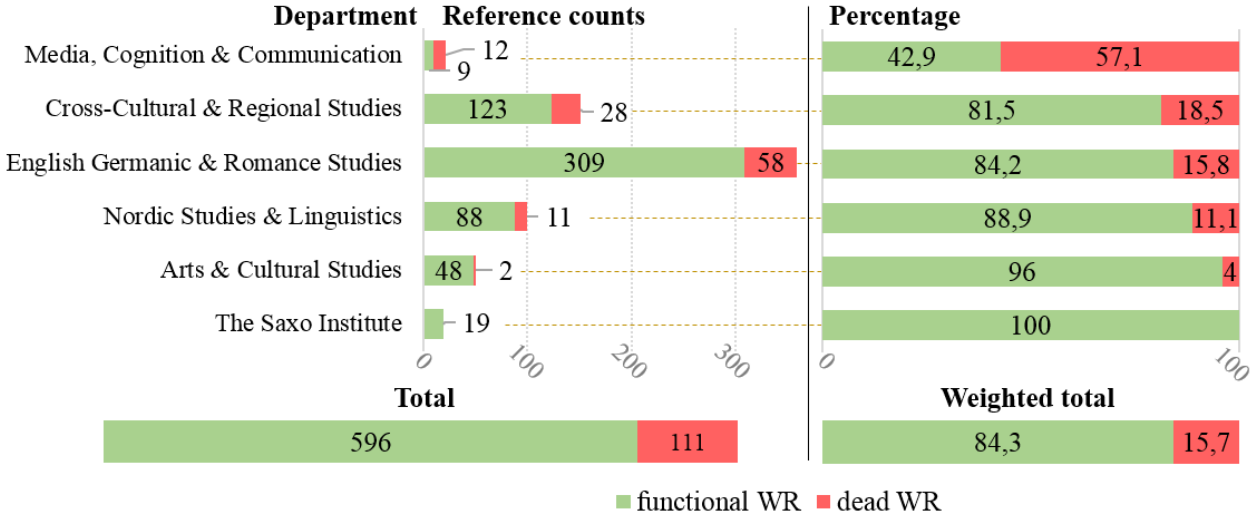

Figure 1: Web references (WR) in master's theses by department. Source: https://diskurs.kb.dk, theses handed in at the Faculty of the Humanities, Univ. of Copenhagen, 2015

Dead links affected ¾ of all examined theses. Out of 707 unique web references 111 resulted in an error message that the requested content or server could not be found (Figure 1). This means that the reader would be unable to check nearly 16% of all cited web resources. However, there are some rather large fluctuations. In some theses, all URLs were functional, whereas in others every fourth link was dead.

A 16% error rate is not remarkable compared to references to analogue material [22-23]. However, diachronic link analysis has shown that the mortality rate increases as links age [19, p. 787; 11-12]. In this respect, web references are markedly different from references to other types of material. Whereas references to analogue materials are stable, the number of erroneous web references will likely increase significantly with the passing of time.

The theses also offer insights into the ways students actually refer to web materials. As mentioned, the commonly used citation styles recommend that web references are made using an

[3] One master's thesis had to be excluded from the survey because the author did not provide URLs for any of the included web references.

URL + date (either of the last update or the last visit to the site). However, since publicly available information on a webpage's last update is no longer customary, it is hardly surprising that none of the authors added this to any of their web references. Of the 33 theses that had references to specific URLs, 7 didn't give a date for any link. A comparable amount of theses (21%) gave a date for each link. Yet, most authors – 58% – noted the date a site was "visited", "seen", "located" or "accessed" for some, but not all web references. Furthermore, many web references lack precision. Several web references point to the front page of a blog or simply consists of a link to a Google search. None of these will enable a reader to assess what is actually being referenced to. In conclusion, the theses illustrate that students are neither consistent nor very precise when they have to refer to material found online. However, our investigation also shows that problems are bound occur, even when students follow recognized citation templates.

PRACTICES AMONG DANISH LITERARY SCHOLARS

Just like other aspects of our culture, the literary field was transformed by the advancement of digitization and advent of the Internet. A range of new outlets has supplemented printed books as well as the literary criticism in papers, journals and monographs. As literature is increasingly being produced, distributed and criticized online, the Internet is an integral part of this literary food chain.

The material for this part of our examination consists of monographs published during the years 2011–16 by academic publishing houses and written by Danish researchers that – partially or fully – treat contemporary Danish literature.

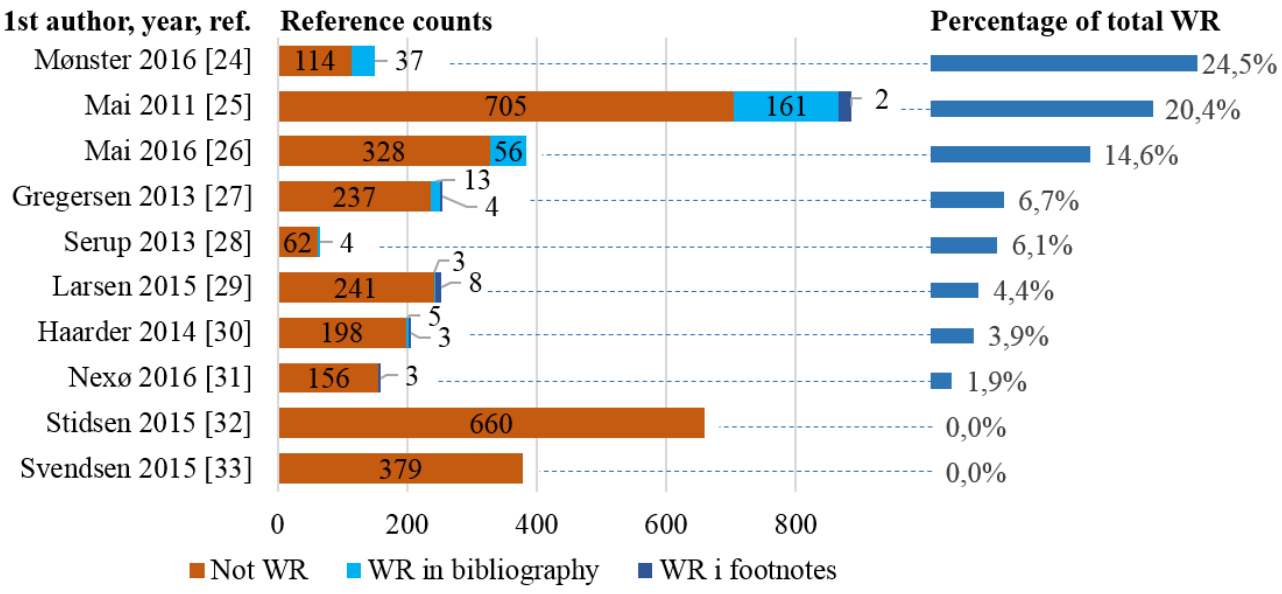

Figure 2: References and web references (WR) in monographs on contemporary Danish literature, 2011–16.

As can be seen from Figure 2, the number of web references in the investigated works is scattered across a broad spectrum. In the ten studied works, there are 3.397 references, 634 of which are unique web references. Among the academic publications, in which web references play an important role, one will find Mai's *Hvor litteraturen finder sted* (2011) and *Galleri 66: en historie om nyere dansk litteratur* (2016) with 20.4 and 14.5% web references respectively, only surpassed by Mønster's 24.5%. At the other end of the scale, web references play a miniscule or no role. Tue Andersen Nexø's *Vidnesbyrd fra velfærdsstaten* (2016) only has three web references (hidden in footnotes) and Stidsen's two-bind dissertation contains a total of 660 literature references, but not a single web reference.

In conclusion, there is no clear trend in the use of web references in the monographs, and the location of the web references is also quite ambiguous: In some of the works the web references are in the bibliography, in others only in the footnotes. In a third variant some have select web references in the bibliography and others in the footnotes alone. Overall, the numbers indicate that literary scholars still focus on printed literary works. However, it cannot be ruled out that the instability of web references plays a role as monographs are often presumed to have a longer shelf life than articles.

None of the literary monographs gives an access date except Mai 2016. We have not conducted a systematic review of the all web references, as the works are so new that the majority of the links used will – statistically – work. Since Mai 2011 is the oldest surveyed work, it was selected for in-depth investigation. The bullets below describe the status of the 181 web references in the book, five years after its publication. For practical reasons, the red, yellow and green of a traffic light are used to illustrate how easily a reader can access the cited web material.

- *Red: dead links or another webpage's acquisition of the content*
  34.3% (62) of the links used were red. Either they result in a bug report, or an entirely different web page – typically of a commercial nature.
- *Yellow: critical links that either refer to a general page or redirect you to an overall website*
  21% (38) were yellow: You get to a webpage that you have reason to believe is the right one, from where you are able to search for (and easily find) the intended content.
- *Green: links that work and – as far as one can tell – lead you to the intended content*
  44.7%, almost half of the links, are green. The link works and you are directed to content that matches the described.

Paradoxically, the most stable links are those that refer to a general front page (e.g. emmagad.dk) with an elaborating text specifying to which article the author refers. Long URLs have a greater risk of disappearing, whereas a short URL exists longer despite the reorganization of content. What is lost in precision is gained in persistence!

CONCLUSIONS AND RECOMMENDATIONS

In our dual study we have documented that references to online resources often fail to meet the standards of good scientific practice. The study furthermore shows that both researchers and

students deliver inaccurate and unstable web references even when they follow renowned citations styles and general recommendations.

This suggests that the recommended and widely used URL + date standard is in need of an update. Alternative options do exist. Among them are PIDs such as Crossref's DOI, which lets users register digital material so that unique references can be generated, and commercial citation services where one can archive snapshots of webpages. Certain reference management programs such as Zotero has added the same functionality, and a number of applications have been developed to archive specific pages and on-screen action. However, both PIDs and citations services depend entirely on the continuous existence of the provider, and a range of services have already gone down, e.g. Mummify.it, which was launched in 2013, but was discontinued the following year [34, p. 58; 35, p. 247]. In addition, it should be noted that few (if any) citation services allow the user to make data accessible. Therefore, these services cannot ensure that research results based on web resources can be verified and reproduced.

Web archives – and especially national web archives that collect material as part of national legal deposit laws – are by far our best opportunity for creating both accurate and persistent references to web resources, but a transparent persistent web material identification should be established before this potential of web archives can be fully exploited. Informed by our study of citations behaviour among Danish scholars and based on research on identification of sources in web archives, we propose that a minimum of four ingredients are necessary in order to make a prober reference to a resource in any web archive [35]:

1) *Web archive*
   The unique identification of the particular web archive in which the reference source can be found (e.g. Netarkivet or Internet Archive (open part) or by their domains netarkivet.dk and archive.org)
2) *Time of archiving*
   The registered time when the web archive has collected and archived a particular resource from the web at large specified in UTC (Coordinated Universal Time) to avoid any time zone confusion.
3) *Archived URL*
   web resource location (in a web archive it works as a so-called URI, since it is no longer points to a location on the web)
4) *Content coverage specification*
   denotes the part of the web resource covered by the reference. This is particularly important for web resources because these may contain multiple web elements (such as images, blocks of text, etc.)

With these four components, a precise and persistent web reference can be written in text form as follows:

> Web archive: archive.org, archiving time: 2016-04-20 18:21:47 UTC, archived URI: http://resaw.eu/, content coverage: webpage.

Our proposal also includes a so-called PWID URI (persistent web identifier), which can eventually[4] be used to open an archive resource directly in a browser. The PWID URI contains the same information as listed above, but transformed into a form that ensure unambiguity and respects the formal URI syntax for browser addresses [36]. In the form of a PWID URI, the corresponding reference would appear as the black text in Figure 3.

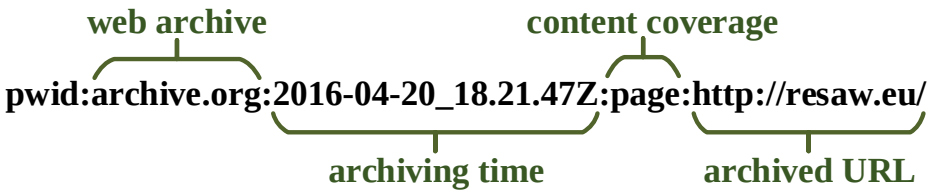

Figure 3: PWID URI. Explanatory notes in green.

In conclusion, the suggested four components give us the best conditions for combining precision and persistency and thus to be in keeping with good scientific standards. Other important aspects such as accessibility and transparency can be provided by full implementation of a PWID URI.

LITERATURE

[1] N. Brügger and N. O. Finnemann, 'The Web and Digital Humanities: Theoretical and Methodological Concerns', *J. Broadcast. Electron. Media*, vol. 57, no. 1, pp. 66–80, Jan. 2013. doi:10.1080/08838151.2012.761699

[2] B. Thomee B., Shamma, D. A., Ni, K., Friedland, G., Poland, D., Li, L.-J., 'YFCC100M: the new data in multimedia research', *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016. doi:10.1145/2812802

[3] R. B. Evans, 'Growing pains: The American Psychological Association from 1903 to 1920', in *The American Psychological Association: A historical perspective*, R. B. Evans, V. S. Sexton, and T. C. Cadwallader, Eds. Washington, DC: American Psychological Association, xvi, pp. 73–90.

[4] E. Chernin, 'The "Harvard system": a mystery dispelled', *BMJ: British Medical Journal*, vol. 297, no. 6655, pp. 1062–63, 1988

[5] C. Nyvang, T. H. Kromann and E. Zierau, 'Det levende web og de døde links', *Nordisk Tidsskrift for Informationsvidenskab og Kulturformidling*, no. 1-2, in press.

[4] The PWID URI is proposed as an URI standard. Furthermore, browser plugins will need to be developed in order to make it resolvable.

[6] Skov, A. *Referér korrekt!: Om udarbejdelse af bibliografiske referencer*. 2. ed. by Bo Gerner Nielsen. Det Informationsvidenskabelige Akademi, Københavns Universitet. Web archive ref.: pwid:archive.org:2016-05-15_23.39.48Z:page:http://iva.ku.dk/refererkorrekt

[7] *The Danish Code of Conduct for Research Integrity*. Danish Ministry of Higher Education and Science, 2014. Web archive ref.: pwid:nerarkivet.dk:2016-02-11_03.02.04Z:part:http://ufm.dk/publikationer/2014/filer-2014/the-danish-code-of-conduct-for-research-integrity.pdf

[8] S. M. P. Benbow, 'File Not Found: The Problems of Changing Urls for the World Wide Web', *Internet Research*, vol. 8, no. 3, pp. 247–250, Aug. 1998. doi:10.1108/10662249810217867

[9] A. Chankhunthod, P. B. Danzig, C. Neerdaels, M. F. Schwartz, and K. J. Worrell, 'A Hierarchical Internet Object Cache; CU-CS-766-9', *Computer Science Technical Reports. Paper 720*, 1995.

[10] W. Koehler, 'An Analysis of Web Page and Web Site Constancy and Permanence', *Journal of the Association for Information Science and Technology*, vol. 50, no. 2, p. 162, 1999.

[11] W. Koehler, 'Web Page Change and Persistence — a Four-Year Longitudinal Study', *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 162–171, 2002.

[12] E. Russell and J. Kane, 'The Missing Link: Assessing the Reliability of Internet Citations in History Journals', *Technol. Cult.*, vol. 49, no. 2, pp. 420–429, 2008.

[13] D. V. Dimitrova and M. Bugeja, 'The half-life of internet references cited in communication journals', *New Media & Society*, vol. 9, no. 5, pp. 811–826, Oct. 2007. doi:10.1177/1461444807081226

[14] D. H.-L. Goh and P. K. Ng, 'Link Decay in Leading Information Science Journals', *Journal of the American Society for Information Science and Technology*, vol. 58, no. 1, pp. 15–24, Jan. 2007. doi:10.1002/asi.20513

[15] B.T. Sampath Kumar and K.S. Manoj Kumar, 'Persistence and half-life of URL citations cited in LIS open access journals', *Aslib Proceedings*, vol. 64, no. 4, pp. 405–422, Jun. 2012. doi:10.1108/00012531211244752

[16] R. Liebler and J. Liebert, 'Something Rotten in the State of Legal Citation: The Life Span of a United States Supreme Court Citation Containing an Internet Link (1996-2010)', *Yale JL Tech*, vol. 15, p. 273, 2012.

[17] J. Zittrain, K. Albert, and L. Lessig, 'Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations', *Legal Information Management*, vol. 14, no. 02, pp. 88–99, Jun. 2014. doi:10.1017/S1472669614000255

[18] D. Aronsky, S. Madani, R. J. Carnevale, S. Duda, and M. T. Feyder, 'The Prevalence and Inaccessibility of Internet References in the Biomedical Literature at the Time of Publication', *Journal of the American Medical Informatics Association*, vol. 14, no. 2, pp. 232–234, Mar. 2007. doi:10.1197/jamia.M2243

[19] R. P. Dellavalle *et al.*, 'Going, going, gone: Lost Internet references', *Science*, vol. 302, no. 5646, pp. 787–788, 2003.

[20] J. D. Wren, ‘URL decay in MEDLINE - a 4-year follow-up study’, *Bioinformatics*, vol. 24, no. 11, pp. 1381–1385, Jun. 2008. doi:10.1093/bioinformatics/btn127

[21] ‘DISKURS’, 16-Mar-2017. [Online]. Available at https://diskurs.kb.dk. Web archive ref.: pwid:archive.org:2016-08-11_23.38.27Z:page:https://diskurs.kb.dk/

[22] N. N. Pope, ‘Accuracy of References in Ten Library Science Journals’, *RQ*, pp. 240–243, 1992.

[23] S. P. Benning and S. C. Speer, ‘Incorrect Citations: A Comparison of Library Literature with Medical Literature’, *Bulletin of the Medical Library Association*, vol. 81, no. 1, p. 56, 1993.

[24] L. Mønster, *Ny nordisk: lyrik i det 21. århundrede*. Aalborg: Aalborg Universitetsforlag, 2016.

[25] A.-M. Mai, *Hvor litteraturen finder sted: bidrag til dansk litteraturs historie*. København: Gyldendal, 2011.

[26] A.-M. Mai, *Galleri 66: en historie om nyere dansk litteratur*. København: Gyldendal, 2016.

[27] M. Gregersen and T. Skiveren, *Det åbne redskabsskur: hovedstrømninger i det nye årtusindes danske forfatterskolelitteratur*. Aalborg: Aalborg Universitetsforlag, 2013.

[28] M. G. Serup and Syddansk universitet, *Relationel poesi*. Odense: Syddansk Universitetsforlag, 2013.

[29] P. S. Larsen, *Poesiens ekspansion: om nordisk samtidsdigtning*. Spring, 2015.

[30] J. H. Haarder, *Performativ biografisme: en hovedstrømning i det senmodernes skandinaviske litteratur*. København: Gyldendal, 2014.

[31] T. A. Nexø, *Vidnesbyrd fra velfærdsstaten: den sociale vending i ny dansk litteratur*. København: Arena, 2016.

[32] M. Stidsen, ‘Den ny mimesis. virkelighedstolkningen i dansk og nordisk litteratur efter Anden Verdenskrig, bd. 1 & 2’, U Press, København, 2015.

[33] E. Svendsen, *Kampe om virkeligheden: tendenser i dansk prosa 1990-2010*. Samfundslitteratur, 2015.

[34] H. V. de Sompel and S. Davis, ‘From a System of Journals to a Web of Objects’, *Serials Librarian*, vol. 68, no. 1–4, pp. 51–63, Maj 2015. doi:10.1080/0361526X.2015.1026748

[35] E. Zierau, C. Nyvang, and T. H. Kromann, ‘Persistent Web References – Best Practices and New Suggestions’, in *Proceedings of the 13th International Conference on Preservation of Digital Objects (iPres)*, 2016, pp. 237–46.

[36] T. Berners-Lee, R. Fielding and L. Masinter, ‘Uniform Resource Identifier (URI): Generic Syntax’, RFC 3986, 2005. Web archive ref.: pwid:archive.org:2016-03-26_12.10.40Z:part:http://www.ietf.org/rfc/rfc3986.txt