Web Archive†

Niels Ole Finnemann

University of Copenhagen, Department of Information Studies,
Njalsgade 76, DK 2300 Copenhagen,
<finnemann@hum.ku.dk>

Niels Ole Finnemann is Professor in digital media, Department of Information Studies, University of Copenhagen. Formerly, he was a professor in internet studies at Aarhus University (2005-2014). He holds a PhD in computer semiotics. He participated in a number of national and international research projects and in the establishment of the national Danish web archive. His publications include articles on the internet and legacy media, mediatization theory and digital media. He is now focusing on the "third wave" in the history of digitization and the complexities of multiple source hypertext systems. He recently published a history on electronic text, "E-text" in Oxford Research Encyclopedia Literature.

Finnemann, Niels Ole. 2019. "Web Archive." *Knowledge Organization* 46(1): 47-70. 99 references. DOI:10.5771/0943-7444-2019-1-47.

Abstract: This article deals with the function of general web archives within the emerging organization of fast-growing digital knowledge resources. It opens with a brief overview of reasons why general web archives are needed. Sections two and three present major, long term web archive initiatives and discuss the purposes and possible functions and unknown future needs, demands and concerns. Section four analyses three main principles for the selection of materials to be preserved in contemporary web archiving strategies, topic-centric, domain-centric and time-centric archiving strategies and how to combine these to provide a broad and rich archive. Section five is concerned with inherent limitations and why web archives are always flawed. The last section deals with the question whether and how web archives may be considered a new type of knowledge organization system (KOS) necessary to preserve web materials, to allow for the development of a range of new methodologies, to analyze these particular corpora in long term and long tail perspectives, and to build a bridge towards the rapidly expanding but fragmented landscape of digital archives, libraries, research infrastructures and other sorts of digital repositories.

Received: 28 May 2018; Accepted 14 June 2018

Keywords: web archives, materials, archives, digital, knowledge organization

† Thanks to the three anonymous reviewers and my editor for very valuable comments to former versions.

1.0 The web—a nearby perfect knowledge organization system

With the rapid spread of the web protocols and the legalization of commercial internet activities in the United States early in the 1990s (Abbate 1999, 213-218; Schelin and Garson 2004, 591), the internet was within a decade transformed from a specialised communication tool for scientists and students to a globally accessible, societal infrastructure to which other media, institutions and corporations—together with individuals—had to accommodate. The WWW-protocols provided easy access to all sorts of information resources. They also opened for a third wave[1] of digitization characterised by exponentially growing amounts of data, by new communicative genres and new formats of knowledge production, dissemination and organisation (Duranti and Thibodeau 2006; Jenkins 2006; Meikle and Young 2011; Hilbert and Lopez 2012; Kitchin 2014; Finnemann 2014a, 2018; Huurdeman et al. 2015).

Today the web has become the most comprehensive knowledge resource ever. This is the result of a vast number of decisions taken by a huge variety of agencies all over the globe each acting due to their own needs and goals. The aggregated result of these efforts has been the development of the peculiar hypertext architecture elaborated on the basis of the TCP/IP and web protocols. The core features of this network architecture are based on the establishment of a uniform, global address system, which can be expanded in both horizontal directions and hierarchical levels and in which any address as well as its content may be accessed from any other address unless specific limitations are imposed. The infrastructure allows editable point-to-point connections between any two machines. It also allows exchanges of all sorts of coded instructions of content, of communicative interactions between people and—the path breaking potentials—of interferences between the functional architectures of the machines in the network.

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

The affordances of this architecture are based on the interconnection of a range of distinct characteristics, such as continuous updating, on-going editing, searching, addition of new sources, calculations and compilations across distance, 24/7, global reach and not least the inclusion of a growing range of multiple source knowledge systems eventually incorporating real time data from any deliberately chosen set of sources. Together this provides an extremely flexible tool, which can be adapted on the fly for knowledge organization performed by public, academic, commercial, or civic service providers as well as for personal use or use within an organisation. Since any relevant source can be added, modified, or deleted and optional selections of resources can be composed any time, the web as a whole seems to qualify as a nearly perfect system for knowledge organization.

Three major obstacles however prevent the web from being a sufficient solution for KO in the twenty-first century. First, it remains too large for any observer, including mechanical crawlers, to overview.[2] Even if a search engine actually covered the whole web, the result of any given search would be incomprehensible because of the number of positives and false positives, negatives and false negatives, intricate language and terminology issues, and the limitations of automated classification.

A second obstacle is the ephemeral character of the accessible materials. According to Brewster Kahle (1997), founder of the first general Internet Archive, archive.org in 1996, the average lifetime of a web page back then was estimated at forty-four days. Later calculations tell similar histories.[3] The ephemeral—or fluctuating—character is a result of the intrinsic characteristics of digital media as they allow any deliberately chosen unit to be connected, disconnected or modified any time. For digital materials the editor position remains open and any deliberately defined sequence of bits or pixels on the screen may be ascribed its own frequency of updating and eventually modification or deletion. Digitization brings with it "the end of an object's stability" (Masanès 2005, 73) because of the constantly on-going updating of addresses, link locations, instructions and content whether the product of automatized routines or on-going human editing of already published materials (Masanès 2006; Brügger and Finnemann 2013; LeFurgy 2015; Huurdeman et al. 2015; Schafer, Musiani and Borelli 2016).

A third obstacle follows from the major advantage of the Internet that it allows everyone to publish. There is no gatekeeping function in the input structure (Costa and Silva 2017, 191).[4] The materials produced are increasingly heterogeneous in purpose, format and interrelation to other materials and subject to changes due to a variety of "editors" at any given time after publication.[5] A further source of heterogeneity is the spread of digitization processes into a still wider array of different types of social processes ranging from scanning of outer space to the interior of our bodies and everything in between. For these reasons, the largest archive of knowledge and information in the world today has itself to be archived and documented insofar as the materials are considered worthy to preserve and to remain accessible in the future.

2.0 General web archives—overview

The attempts to build general web archives based on ongoing "deliberative and purposive preservation of web material" (Brügger, 2010, 349) took off in the mid 1990s only a few years after the spread of the WWW protocols.[6] The approaches differed in respect both to the range of the materials collected and to the criteria for selection. In 1996, the private *Internet Archive*, archive.org in the US took a "generalized philanthropic" approach aiming to cover the whole web (Webster 2017: 181). The same year *Kulturar3w* in Sweden and *Pandora* in Australia (both based within the national libraries) took a national domain perspective. The kind of materials collected also differed as the *Internet Archive* and *Kulturar3w* aimed to collect the widest possible set of materials while the *Pandora* project focused on a selected set of sites considered to be the most valuable or authoritative sites (Webster 2017; Koerbin 2017).

The early initiatives have been followed by a growing range of national initiatives especially in Europe. National libraries are predominant agencies covering national domains, except for the US, where the non-profit Internet Archive aims to provide worldwide coverage and the Library of Congress maintains a huge selective archive. In addition, a range of selective archives is established at major universities. There are only a few web archives, if any, in the Near Middle East, Africa and South America (Costa and Silva 2017, 198-99). Thus, web archiving is mainly established in the northern hemisphere even if "this ever-growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression" (UNESCO Charter on the Preservation of Digital Heritage, Article 1).[7] According to the charter, all sorts of digital heritage, born digital heritage included, should be "protected and preserved for current and future generations" (Charter Article 1). Web archives belong to the category of "born digital cultural heritage" (materials created in digital form); however, they differ from other kinds of born digital materials, because archived web materials may include coded Internet links in the messages. Due to the global reach of the address system of possible destinations from any anchor and the indefinite number of possible instructions to be performed by any link on the live web, web archives have become more complex than any formerly known set of data, except for the live web as a whole. As

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

web archives, no matter how they are built, also include broken links to the surrounding web, they are also always flawed.[8]

A list of web archiving initiatives can be found in Wikipedia.[9] As of 20 April 2018, the list included eighty-five initiatives. Many of these web archive projects are also member of The International Internet Preservation Consortium (IIPC), which was established in 2003. The membership list can be found at http://netpreserve.org/about-us/members.[10]

Webster (2017) distinguishes between generalized "philanthropic" archives, national web archives acting according to a national responsibility for the published record and archiving efforts by organizations (be they governmental institutions, universities, research communities, corporations, churches, activist groups and others) aiming to preserve their own web content. There are, today, two major general and philanthropic archive initiatives, the Internet Archive, established in 1996, and Common Crawl (commoncrawl.org) established in 2007.[11] Since 2006, the Internet Archive also provides a subscription-based archive service, Archive-it (archive-it.org) allowing anybody to establish a tailored web archive, which may also be incorporated in the internet archive. The European Internet Memory Research, a commercial offspring from the Internet Memory Foundation, provides a similar service, archive-the-net, since 2011.[12]

Brügger (2018) distinguishes between transnational archives, national archives, regional and local archives, research-oriented archives driven by universities, university libraries, museums, activist web collections, social media databases, adding also "restored collections" of various sorts of otherwise lost web materials made accessible on the live web by enthusiast, nerds and others. Thus, there is a growing array of agencies, archives and criteria for collection (harvesting) of web-materials. This is partly a result of the still young and decentered history of web archiving. It also reflects a need to rethink the principles and criteria for archiving, organizing and usage of these materials, as former principles of archives and libraries are not sufficient.[13] Neither are the principles of knowledge organization (KO) as argued by Duranti and Thibodeau (2006) and Ibekwe-SanJuan and Bowker (2017) and further discussed in this article.

2.1 Web archives, digital libraries and archives.

So far, web archives develop in unclear relations to each other as well as to other sorts of digital libraries, archives, repositories, collections, research infrastructures, and a variety of curated digital heritage institutions. These and their equivalents (formerly often prefixed as cyber- or e-) all seem to be "evolving too fast for any lasting definition" as concluded by Seadle and Greifender (2007, 169) in a mini survey of definitions of digital libraries. There are patterns in this process, however.

Back in 1992, Buckland identified three major steps in the digitization of libraries. He characterizes (Finnemann 2014b) the first step as an initial process of automation of catalogues to fulfill the same tasks as before but more effectively on a local scale. The second step was the digitization of publications, and finally Buckland identified a third step in the use of the internet as a means for distribution and communication. A similar perspective is presented by Borgmann (1999, 238-9) as the "librarian view" in which the internet is a means for distribution and collaboration, and emphasizing that it cannot in any way be considered a digital library itself (Jones, Andrew and MacColl 2006, 4-5). A main reason given is that the WWW or the internet is not an institution and the materials are not selected or documented in any standardized form. This does not exclude that a specific website can function as a digital library insofar as any specific site follows the standardized practices within library and information science. Thus, the internet is recognized as important for digital libraries, but only as a means of distribution and communication—or what could be characterized as a platform perspective external to the digital library.

In the late 1990s, the platform perspective also appears in a different and inclusive form in the US National Science Foundation's (NSF) short definition of digital libraries, which "basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digitals libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass manipulate the information on the Net."[13]

The positions differ in their conceptualization of the role of the internet and WWW. In the librarian view, the internet is external to the digital library, while it is precisely a digital library in the NSF perspective (1999), aiming to "effectively mass manipulate the information on the Net." They also differ in their professional perspectives. In the NSF perspective, the materials in the-internet-is-a-library can be manipulated ("effectively") as all other types of digital materials. The internet is a platform for huge amounts of information accessible for analyses and the "librarians" are substituted for software tools used "to manipulate effectively," reflecting a tension between computer science and library science ideas of "digital libraries" (Jones, Andrew and MacColl 2006: 5). A single website may serve as a library, but the Internet is either fully outside the digital library or it is itself such a library.

Borgmann (1999, 227, 239) furthermore identifies a tension between notions of digital libraries in different disciplines, though not similar to the NSF versus the librar-

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

ian's concepts: on the one hand, she describes a "researcher community view" focusing on usage of the content and on the other hand, "a practicing librarian's view of digital libraries as institutions delivering 'information services in digital forms." The tension between the perspectives of research communities and library professionals can also be found today and have also been articulated within the area of web archiving (Jones 2006; Dougherty et al. 2010; Meyer and Ralph Schroeder 2015; Webster 2017; Huurdeman and Kamp 2018).

While the librarians and the NSF disagree concerning the question of whether the internet is a digital library, they both ignore the questions of whether and eventually why web materials should be archived. Similar discussions can be found in archival science. In 2015, Theimer distinguishes between four "commonly used" notions of digital archives: collections of born digital records, websites that provide access to collections of digitized materials, websites featuring different types of digitized information around one topic, and, finally, web-based participatory collections. In this perspective, the WWW and the internet is mainly, as for Buckland, a public platform for websites, some of which are used by libraries and archival institutions as entrance to their collections of digital records and professionally produced collections of selected sets of digitized materials (e.g., digitized cultural heritage materials). However, certain kinds of genuine web materials (web-based participatory collections) now appear as possible objects for archiving efforts. In this entry, web archives, however, does not belong to the class of "digital archives."

The internet-as-platform perspective external to the digital library is further elaborated in Michetti's (2015, 104) entry "Archives and the web" in the *Encyclopedia of Archival Science* on arguing that the so-called "web 2.0" represents a change from a more autonomous institutional position to a more participatory position, a platform for interaction with stakeholders, and still considering the web as an external environment, which also may pose serious threats to the authority of the archival institutions.

However, in the end, in the last entry in the archival science encyclopedia, "Web Archives" finally appears introducing some of the unique characteristics of web materials, that make such archives valuable as well as extremely difficult to archive: "In many cases, however Web archiving activities deal with content that is interlinked at different levels and is spread across many different sites" (LeFurgy 2015, 414). This also implicitly explains why these archives have their own separate history. The complexity of these materials raises questions to established library and archive principles as manifested in the series of cautious conceptual steps taken to the final inclusion of "web archives" in the fields of library and archival sciences.

Since the web is a means of distribution, a platform for interconnections and interactions between all sorts of agencies as well as a medium with its own distinct types of content, web archives may also take on multiple functions. They can be dealt with as archives of web pages and linked relations between these, as a resource from which a variety of corpora can be extracted for a variety of analytical approaches, as well as a historical index to a wider set of information resources, digital archives, libraries and other repositories included, independently of their own definitions and delimitations. One might consider whether digital archives, libraries and other repositories should also prepare their own sites for future recognition via web archives.

2.2 Archives with broad scopes and open-ended time perspectives

A major distinction between current initiatives is the question whether the "deliberative and purposive preservation of web material" (Brügger 2010, 349) is predefined with a limited timespan or aims to be on-going with open-ended time perspective. A second major distinction is between archives based on thematic limitations and archives based on broader social and cultural criteria. In the following, the focus will be on general web archives dedicated to on-going collection with an open-ended time perspective and oriented towards a broad set of social and cultural criteria. The three main reasons for this are: 1) general web archives cover a much broader range of social and cultural practices than special collections; 2) general web archives will include more complex sets of data materials and codes and thus also reflect the complexity of social and cultural relations more fully than special collections; and, 3) general web archives raise without doubt the most challenging archiving issues ever thus providing the richest resource for the understanding of the development of digital materials. It is generally accepted that digital materials "constitute complex research objects that may include a variety of formats and content types such as images, data and publications" (OCLC 2018: Vol 1, 8; Duranti and Thibodeau 2006). These kinds of complexities apply to many kinds of digital materials, but for web archives comes a radically new type of complexity due to the hypertext nature of the web, which manifests itself in a "complex array of links to external sites." [15] However, the complexity is not simply a matter of the array of links, but even more related both to the array of coded instructions that may be attached to any link and to ever evolving utilizations of new kinds of editable time sensitivity. The links may include instructions for the creation (calculation, manipulation, aggregation, modification, deletion) of content and of functions performed on the site linked from, linked to or on any other

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

destination—if only somebody wants.[16] In an evolutionary, theoretical perspective, the more complex set of data should also form the basis for characterizing less complex datasets while there is no way from the description of less complex set of data materials to the description of a more complex set. Special collections of web materials are less complex than general web archives. In so far web archives belong to the most complex types of digital materials, their description may be considered paradigmatic for the more elaborate notion of all sorts of digital materials.

The notion "archive" usually refers to the collection and preservation of materials produced within an institution or corporation or as private collections of materials. Web archives are, in most cases, concerned with materials published on and captured from the live web. The web itself, however, is not delimited to public materials only, as the web protocols are used also for internal purposes in most institutions and organisations. The delimitation raises both technical as well as legal issues, because the border between public and private is editable. Materials made public can be made private and vice versa. Site owners may also protect their pages against web crawlers by including a robot.txt instruction in the top directory of the site.[17]

The delimitation of general web archives from specialised social media archives is also unclear. Social media like Twitter and Facebook are both available on the WWW and via apps on mobile platforms. For Twitter, which is based on public and distinct messages (with text, tags, links and images), a full archive is possible. In 2010, the US Library of Congress was allowed to keep a full Twitter archive, but in 2017, the Library of Congress moved from a full archive strategy to a selected strategy leaving access to the full archive or to a selected set of tweets to commercial vendors.[18] The case of Facebook is more complicated, first because of ever on-going user modifications of privacy settings, second because Facebook operates both as moderator and to some extent as editor, and third because the communication patterns are highly dependent on user behaviour including references to sources outside Facebook. In the case of Facebook, Twitter and similar services that are driven by large corporations or even monopolies, it might be worth considering whether agreements of access to their own archives could be made or enforced. A third option might be to establish specialised archives dealing with specialised multiple source real time information and knowledge systems, which are either not only web based or does not fit into the general web archive strategies.

The legal issues concerning harvesting, preservation and access are also dealt with in different ways, not least depending on national legislation on privacy protection and copyright. Some archives build on legal depository laws, which may allow them to trespass robot-txt limitations, other archives respect robot.txt while others again allow materials to be deleted from the archive on request by the owner. Copyright and privacy issues are not dealt with in the following as they depend on national legislations.[19] Materials published on the web are subject to archiving efforts on a par with materials published in other formats be they non-digital, digitized or digitally produced but published on non-web platforms and media. Thus, web archives should be considered as part of the wider issue of preservation of the published record and global cultural heritage.

3.0 Web archives—purposes and functions.

One fundamental reason for archiving is easily at hand. Most researchers studying one or another kind of web activity are familiar with the need to ensure copies, archives of the materials they study, as they can never know whether the materials are still there in the same unmodified form tomorrow. Thus, a web archive, however small it may be, is needed to ensure that "the use as a trusted citation in the future" is possible.[20] The need for trusted citation also implies a need for institutionalized solutions both to guarantee the collection, the validity and the preservation and accessibility of the sources. Each of these issues gives rise to many questions beyond the scope of this article. One aspect, however, needs to be addressed since web archives are confronted with issues of trust, which differ from other sorts of born digital materials. While authorship has played a major role in establishing trust in the modern libraries and archives, authorship relations in the web landscape are often difficult or even impossible to establish, due to use of anonymous profiles, remix, on-going modifications and updating as digital materials remain editable (Dougherty and Meyer 2012). Even if this applies to all digital materials, the issue of establishing authorship and trust becomes critical in the networked landscape of web materials in which modifications can be imposed across distance, as is the case in many multiple source knowledge systems.[21] The question why web archives will always be flawed due both to intrinsic characteristics of web materials and to selections methods will be further elaborated in section five.

Trusted citation forms the basis for documentation and the establishing of the validity of knowledge including not least the distinction between past and present. Thus, archives, libraries, museums and other sorts of collections play a very fundamental if not always highly appreciated role in modern societies.[22] The appreciation of web archives is also still lacking, as they are still not used that much except for consultation of individual webpages. According to Meyer and Schroeder (2015, 191-2), web archives are at risk of ending up as "dusty archives," because scientists prefer to use the live web in spite of the missing

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

materials, which are outweighed by the even faster growth of live web data. This is maybe the case for internet researchers of today that are strongly oriented towards the new developments and shows only a marginal historical interest. A history of digitization, the inscription of nature, culture and society into the binary alphabet, is still to be written. This notwithstanding, there are strong reasons to believe that these archives will become increasingly useful. First of all because the live web cannot replace web archives in a long-term perspective. The live web and the archived web will develop as increasingly different types of archives and serve as resources for different kinds of studies, even though such studies in some cases may be combined. At the same time, web archives are likely to become a still more unique source, sometimes even the only source available for a growing range of historical studies.[23] Though, to remove the dust, the archives could actually take a more active role as suggested in Winters (2017), eventually also by providing explorative facilities to scholars, scientists, students and the wider public.

As society increasingly articulates itself on networked digital media platforms, web archives become still more significant primary sources for the documentation of cultural and societal processes, which web materials either refer to or are the product of. The web today has become a main resource for externalised human memory whether as individual memories or as an array of shared memories in which the individuals take part, be it on local, regional, national, or transnational scales. Thus, the history of the twenty-first century cannot be written without these archives. They are also a main source for the documentation of the history of the web and the growing range of web-genres even if some parts of the history can also be documented in other media-formats.

To foresee any sort of future use, the ideal solution would be to preserve all of it. Since this is not possible for a variety of reasons, which will be discussed in the following, the criteria for selection of materials come into the fore.[24] What should be preserved and why? Such questions of course have been given an answer in each and any existing archive, but the answers are strikingly different and seldom discussed in the literature.

In a long-term perspective, web archives are legitimized by the value of their use. Again, the ideal solution, to select the materials most relevant for future needs and concerns, is not an option, as "the interest[s] of future users are poorly represented in selecting materials to preserve" (The Blue Ribbon Task Force 2010, 2). This is not least an issue because "one doesn't know what information future generations will consider important" (Arvidson, Persson,and Mannerheim 2000). The future needs and concerns remain unknown at the time of archiving. Future usages presuppose the existence of the archives, which have to build on expectations of future value for yet unknown demands and purposes.

The issue of unknown future demands has been addressed from an economical point of view in the Blue Ribbon Task Force Report on sustainable preservation of digital materials. The report considers long-term preservation of digital materials as a "societal challenge on a par with climate change and sustainable energy" (81) and focuses on digital "materials that are of long-term public interest" (1) while the market does not fulfill the need for long-term solutions. The report identifies four content domains "with diverse preservation profiles" in respect to economical sustainability:

> Scholarly discourse: the published output of scholarly inquiry; Research data: the primary inputs into research, as well as the first-order results of that research; Commercially owned cultural content: culturally significant digital content that is owned by a private entity and is under copyright protection; and Collectively produced Web content: Web content that is created interactively, the result of collaboration and contributions by consumers.

According to the report, the insufficiencies of the market apply to all four domains as a result of structural challenges in respect to: 1) long time horizons; 2) diffused stakeholders; 3) misaligned or weak incentives; and, 4) lack of clarity about roles and responsibilities among stakeholders. The report suggests that "trusted" public institutions like libraries and archives step in when required acting as proxies for future needs possibly in public private partnerships (2).[25]

The four domains each with their own economical preservation profiles do not fit to contemporary web archiving strategies. General web archives will include some materials of all these types, but also a much wider set of digital materials. Some of these materials are better cared for in specialised institutions be they data repositories, research infrastructures or special collections of various kinds. The distinction between commercially owned cultural content and collectively produced web content also seems to reflect an early—pre-commercial—period in the history of social media. Today, most digital materials whether scholarly discourse, research data or collectively produced web content belong to the category commercially owned cultural content, at least if they are publicly available.

A further limitation is that the economical approach taken cannot respond to "the dynamism and uncertainty of long-term value of digital content on the web environment" for which the conclusion is, that it has to be left to interested parties to "model and test preservation strate-

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

gies, and to provide clarification about long-term value and selection criteria" (Blue Ribbon 2010, 4).

It is probably no coincidence that the report is most vague when it comes to dynamic and interactive hypertext materials, which happens also to be those that are unique for networked digital media and constitute the fundamental architecture of the web, the kernel in contemporary societal infrastructure and which cannot be properly documented in any former medium (Jenkins 2006; Kitchin 2014; Finnemann 2001, 2017, 2018).

While the report is insufficient in the structuring of materials and issues to be considered, it brings into focus that the longstanding preservation strategies for scholarly discourse across the four domains considered "have been disrupted by digital technologies" (Blue Ribbon 2010, 49). The notion of disruption, however, is rather unclear. Two of the four domains, "scholarly discourse" and "commercially owned cultural content," are digitized transformations of existing domains. Digital "research data" represent a fast-growing amount of data generated in the "cooking" of the data captured in a research project. These data are increasingly considered to be valuable resources also for other research groups as they allow new usages. Finally, "collectively produced web content" is a genuinely new domain even if the notion of "collectively produced" covers a wide range of different types of coproduction and collaboration.

In any case, the amounts and ephemeral character of web materials imply that archiving has to take place on the fly as things are published, before they are modified or removed and before a validation whether they are worth to be preserved. This is at odds with principles of selection due to claimed value and quality but is in accordance with widely used legal deposit principles for printed materials. It is also at odds with the use of acknowledged content providers (e.g., publishing houses or media corporations) as proxies guaranteeing the quality due to the overwhelming number of digital content—and service providers and the transnational reach. Anyway, the here and now condition of web archiving introduces timescale-dependencies unusual to traditional archiving strategies, as it will be further discussed below.

Since it is not possible to predict future needs and concerns, selection should rather aim to cover a wide range of materials in order to document the variety of agencies, platforms, genres, and topics, interfaces as well as network patterns and so forth. The range of possible purposes are more insecure but still important. This is an argument for diversity as a fundamental principle of general web archiving.

To remedy the lacking insight in future needs and concerns it might help to set up a range of generic purposes. In his presentation of The Internet Archive, Brewster Kahle suggested that such an archive might "prove to be a vital record for historians, business and governments" (Kahle 1997, 1). If elaborated a bit, it might include preservation of cultural heritage, future commercial purposes and future research purposes. A "public service" for civil society and citizens might also be added. Even if these generic purposes overlap, they remain relevant as distinct criteria for ensuring diversity. This is very much in continuation of well-known criteria for archiving.

Two more criteria, which relate to the specific characteristics of digital media, need to be considered. First, insofar as diversity is used as a main criterion for selection, web archives may serve as a time-sensitive index not simply to the web history (e.g., web resources, agencies, link relations, genres and all sorts of online activities) but to a wide range of social and cultural practices, relations and agencies by preserving the website and the link relations.

The web of today is not solely the most comprehensive and uniformly addressable knowledge resource. It also hosts a range of knowledge portals each organised due to a set of specialised criteria and somehow fenced off from the flow of interactions to protect and ensure the stability, reliability and validity of the materials.

Many special archives are not included in general web archives, but even so their existence can often be traced.[27] In this way, general archives may also serve as index for existing special collections at any given point in time. This would also include documentation of and eventually access to the expanding array of special collections of web materials as well as other sorts of digital data materials, including research data and eventually social media data.

Second, since the web at any given point in time provides access to a hitherto unknown broad range of societal practices, an on-going, cumulative, archiving strategy will provide a fast-growing set of data allowing for a huge variety of analyses of a growing range of patterns not otherwise recognizable, mainly restricted by the development of adequate methodological tools. This may be true both in respect to patterns manifested in materials from the same period (long tail) and in respect to diachronic patterns in materials collected over the years (long term).[27]

Diversity, however, remains a loose category and should be further elaborated in respect to a wide array of dimensions, such as e.g., authorship, cultural and social practices, communicative genres, visual and auditory characteristics, search facilities, interfaces and web design, link and network relations, themes and issues, time sensitivity of the materials and the facilitation of both synchronic and diachronic perspectives to be selected. On top of the array of dimensions there is also an array of future purposes ranging from cultural heritage, historical documentation and testimonies, to possibly future commercial purposes and the documentation of civic society as well as individual in-

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

terests and personal concerns. Finally, there is also a need to reflect the range of scales of analysis from micro studies of single cases to regional and global scales.

4.0 Strategies for value?

The principles for web archiving are partly derived from the principles developed in the long history of archiving and the building of libraries for books and other materials, but the material characteristics of web materials make it inevitable to transform these principles. This is the case for the methods of collection and preservation, for making the materials available, and for the array of possible usages. At the same time, these material characteristics allow for an array of usages and purposes that were not feasible in archives of former types of materials.

Since the launch of the first major initiatives for on-going archiving, the establishing of general or national web archives as indicated above has been accompanied by a fast growing range of special collections whether created by scholars, researchers, archival institutions, universities and other agencies concerned with collection of materials within a limited time span of a specific project or special collections concerned with a particular set of themes including also a range of new (digital) research infrastructures, which are either e-archives, repositories or functions as portals such as the Holocaust Research Infrastructure.[28]

The distinctions between special collections, research infrastructures and general archives are not clear cut, but they still make sense, because each of these purposes has implications for the array of methods used for selection. Thus, the "perfect" system for knowledge organization is transformed into an ever-growing bricolage of web materials harvested and archived due to a variety of criteria.

4.1 Canon and topic-centric selection

One set of criteria for selecting the materials to be archived relates to the established idea of a canon based on quality (of the content of the source) or authority (of the author, publisher or editor). Such strategies can focus on a specific area, as for instance governmental sites, a discipline or a domain, e.g., literature, art or other areas where canonization plays a significant role. Such archives—eventually supported by focused crawlers—may be targeting any particular theme, topic or purpose either for a limited period of time or as an on-going activity. In accordance with Masanès (2006), they are referred to as "topic centric."[29] A topic centric collection of web materials, for instance covering a political election campaign with in a limited period of time, is also described as a web sphere delimited by theme, time, stakeholders etc. (Schneider and Foot 2005). All such efforts, however, will only include a tiny fragment of web materials. They cannot serve as documentation of the development of the web or a larger part of society.

The difficulties facing attempts to establish some sort of a canon within any field also apply to similar efforts to establish archives based on quality, societal significance or relevance or in short to establish web archives based on a validated canonical hierarchy, expertise or state defined authority. The criteria of selection of such validated special collections may be more or less the same as the criteria for non-digital archives, libraries and collections, but the conditions for collection differ.

Even if the purpose is clear and well defined, the question remains where to find the materials relevant for the canon or topic in question. These materials may appear at many different web addresses embedded in networked relations on a blog, on Facebook, YouTube or any other public site located in one or another national domain or in other domains or subdomains. The question where to find relevant materials on any given topic may have very different answers from day to day. Over the years, migration of archives adds the question of how a given set of materials are embedded in changing archive histories.[30]

Topic-centric archiving includes the harvesting of materials related to a particular domain, understood as an area of knowledge. This is quite different from the notion of a "web domain," understood as a particular set of web-addresses and which constitutes "domain centric" harvesting (Masanès, 2006, 41-3).

The distinction between value- and quality-based, topic centric, archiving and "broad and rich," domain-centric, bulk archiving is not simply a matter of choice, as the former strategy presupposes that materials remain available during the process of quality validation and collection. It also presupposes intellectual validation and selection of a relatively small subset of materials produced. Thus, quality-based archiving is no longer sufficient due to the huge amounts and the ephemeral character of web materials.

4.2 Domain-centric selection

A second set of criteria for which there is no non-digital equivalent relates to "domain-centric strategies," departing from a specified list of web domain-addresses and looking for whatever content stored at those addresses and eventually at all the locations linked to from the URLs listed in an initial seed list. This strategy provides a "snapshot" of all websites present within the specified domain list at the time of harvesting. Such strategies play a significant role in a growing number of general web archives departing from a national domain. Web domain addresses are necessary in all strategies; you cannot get the content if your machine

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

does not have the domain address. The use of domain addresses as a main criterion for selection is particularly relevant for national archiving strategies. Archiving based on domain addresses have several advantages, not least that they can be automatized to a very high degree, because the harvesting of materials can be done with crawlers, who simply follow the links from an initial site (or a seed list of initial sites) to the pages on a specified number of the subdomain levels. The automated procedures are of course also much cheaper than selection based on intellectual resources.[31]

Archives and collections defined by a particular issue or purpose will base their strategies on the issue or purpose in question. They will ask where the materials are concerning a given issue, X. They will search for the domain addresses where the content is stored. In these cases, materials are selected to be preserved, because they relate to the subject in question, while general web-archives tend to ensure a broad and rich representation of what was there (within a given range of web addresses) at the time of collection. They will search for any content stored at a given set of addresses, be it the whole web or a selected set of web domains. General or broad web archives are not that general though, as they most often are centred on a particular set of web domains, as, for instance, national domains. This particular delimitation is relevant since the web is most often closely integrated into the public sphere within a nation.[32]

Answers to the question of what to preserve are highly dependent on national, cultural and eventually linguistic scopes. At the same time, the delimitation is difficult since most web domains include sites from agencies in many countries and since people are still free to use sites on most domains. Thus, domain centric archiving of a national domain is not the sole source of relevance for a national web archive. Attempts to collect materials of national interest from other domains remain necessary at least until the establishing of archives, based on equivalent archiving selection principles related to all top-level domains.

The amounts and ephemeral character of web materials call for the use of mechanised and automatized archiving methods also favouring mechanized methods for providing metadata. While this leaves the materials insufficiently described for many purposes it also allows for new analytical strategies to be further developed as the metadata collected may serve as a kind of mark-up allowing for instance the analysis of—changing balances between—file formats, inter site link relations and other possible indicators for relationships and usages.

Mechanized archiving methods ensure a richer and more varied set of archived materials than otherwise obtainable. Thus, it is possible, for instance, to document and further analyse the long tail of web link relations within a given period, as well as a broad range of long-term developments in the communicational practice as the archives develop over the years. Fake news will be there as well and some of traces of their history may be revealed. Web archives, furthermore, contain traces of link connections, thus serving as kind of index to the social, cultural and political agencies whether civic or professional and their interrelations at a given time. They may also be designed to serve as an index to specialised types of KO in the form of links to special collections, research infrastructures, and—time sensitive—multiple source knowledge systems.[33] They may, furthermore, include traces of the emergence of new genres before such genres are recognized as such.

While topic-centric archiving requires a relatively high amount of human curating to find and validate the materials and resulting in very limited set of materials, domain-centric "bulk" archiving ("snapshots") takes place without preceding validation. Whether it is worth it to preserve all these materials of questionable quality is of course a highly controversial issue and the discussions are still on-going.[34] National domain-centric strategies are used in a huge number of national web archives, which might indicate that there are advantages and values making it worthwhile to do.[35]

Such values can be identified on six dimensions: 1) all sorts of individuals, groups, organisations and institutions today produce web materials. For this reason, the materials give a much broader and richer documentation of human life than have been recorded ever before in human history. Thus, they also enter into the debates concerning narrow, high quality meritocratic notions of "valuable culture" versus broad notions of "low"—culture and society as a whole; 2) when stored in web archives, these materials form a unique type of source materials for studies in many areas not feasible without these materials; 3) the collection in digital form of these materials, furthermore, allows for an ever-growing range of new methods to exploit the networked connections of the materials independently of any higher order imposed on these materials; 4) bulk harvesting of a national domain will also include materials that might belong to topic-centred archives but are not found via topic-oriented harvesting methods. Such materials would include, for instance, traces of new genres, tendencies and agencies not yet identified and their future role not yet recognized at the time of harvesting; 6) bulk harvested snapshots also fill some of the inevitable gaps between all sorts of special collections, including also materials, which are only recognized as valuable at a later point in time; and, 6) bulk harvesting of snapshots can also be supported by the "big data" argument that the inclusion of all possible materials (N = "all") allows the detection of more outliers and thus more nuanced analyses than the

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

use of representative samples (e.g., Halevy et al. 2013).[36] Thus, the values stretches far beyond the fundamental need for trusted citation of any given website. General web archives allow for much wider array of documentation of social and cultural practices and they include the aforementioned function as index as well as an emerging array of new methodologies to be used in the analysis of archived web corpora. This is the case on scales ranging from small scale to the overall corpus within and archive. Broad strategies neither exclude deletion or augmentation of materials in the future.

4.3 Time-centric selection

A third set of criteria for selection relate to the complexity of the variety of time scales, which may be coded into web materials in a deliberately chosen granularity of screen pixels. Like the second set, these are unique for archiving of digital materials. A main trajectory in the development of web genres is the on-going developments of new ways to exploit time variations. A few examples showing the increase in use of variable and editable timescales will do. Web archiving history is at least to some extent rooted in the fear of or the experience of the sudden disappearance of websites overnight.[37] This aptly explains that time-sensitive archiving strategies in some cases need to be real time archiving on the fly and more generally that web archiving need to reflect the updating frequency of a site, a page, a link or even any single element on a page.

Web pages and websites are not only short lived; they are often also interactive and include scripts eventually embedded in dynamic link instructions that may use materials and other scripts from other sites. From the point of view of archival record theory, this gives rise to a double reformulation of the notion of digital records (Duranti and Thibodeau 2006). First, these records are described as distinct to electronic and paper-based records as "the stored components of digital records enable reproduction of the record, but are not the record" (51). This fundamental distinction between the stored and invisible sequences of bits and the sensible manifestations on a screen or another output device apply to all sorts of digital materials. The distinction is crucial, they argue, because of possible errors in the processing of the manifested record. It is maybe even more crucial, because the codes organizing the reproduction of the manifested report remain editable and also depend on the specific interface used to initiate the reproduction. The relation between the stored content and the interface is always an editable hypertext relation. This editable space is not always used for semiotic purposes, but it is possible to do.

Duranti and Thibodeau also identify a need for an even more far reaching reformulation of the notion of a record due to the interactive, experiental and dynamic properties of digital media and most radically due to networked digital media in so far "the first manifestation cannot be reproduced with the same content and in the same form" (51, 66).

Among the interactive documents, they distinguish between documents with variable content, where the rules for enabling the record do not vary and documents for which also the rules may vary. The former group includes frequently updated materials in which the updates are not cumulated and existing materials not overwritten. The latter group include documents created according to user inputs or depending on the sources of content data (e.g., personalization). The most difficult cases in their perspective finally relate "to the use of adaptive or evolutionary computing applications where the software can change autonomously" (45-46).

As a conclusion, they distinguish between digital materials that can be archived as records, materials that can be partly archived as records and finally some materials that cannot be archived due to the lack of significant fixed features.

An even greater complication is that links and hypertext relations are not simply connections (as a reference system or footnote system); they also always include a set of instructions of what to do at a specified destination somewhere on the network. The content can be deleted, modified, moved elsewhere, new content added, or remixed, old content overwritten or downloaded, images can be redrawn, figures can be recalculated, new rules for calculation and other types of transformation can be implemented. Take a Google search that involves the execution of hundreds, if not thousands of instructions for collecting, sorting and presenting the results of any single search as an easy illustration of the complexity of scripted instructions performed by activating a link. These operative instructions are often, but falsely ignored as integral part of hypertext relations even if they might trigger modifications according to an editable timescale of any element specified on a page at any location.

As a result, any webpage or a part of it can be made dependent on new inputs via the interface or via instructions from external sources (e.g., personalized services) wherever they are located if only connected to the internet. Thus, web materials can be modified any time by the provider or owner or by coded instructions built into the site, possibly triggered by a visitor, or built into another site from which the materials are accessed, or the action is triggered during a page-request. (Duranti and Thibodeau 2006; Masanès 2006, 13-17; Taylor 2012; Brügger and Finnemann 2013).

This facility is increasingly used in contemporary network-based knowledge organization systems or multiple

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

source knowledge systems. An example is Cetina's (2009) analysis of a software system in which six to eight screens are used to configure a huge number of cells, each linked to its own specific source with its own timescale and updating frequency. The system is used in (or constitutes) the Foreign Exchange Market and the screen cells include real time information from all sorts of financial markets worldwide, as well as journalistic news sources, real time algorithmic trades and deals performed by the human traders. Cetina introduces a concept of synthetic situation defined by a particular scope for the collection of multiple sources into one system. She also describes the time dependent demands for response presence, which in this case is specified within a fractional part of second defined by the purpose and the updating frequencies of the sources. The response presence, however, can be specified and implemented otherwise and is itself an editable feature, which allow for specifying a variety of "windows of interaction" in networked knowledge systems.[38]

Network-based multiple source knowledge systems are used in a growing range of areas far beyond the financial sector, in climate research and real time monitoring of all sorts of processes both on a local, regional and global scale. They represent a fast-developing new kind of knowledge organization, which however often requires their own archiving strategies because of the use of multiple timescales, variable updating frequencies and also depending on principles for selection of materials in the collection. Since editable timescales can be inserted in between any two elements, they constitute a reservoir for development of new genres while they at the same time create a number of complications for archiving.

Time sensitivity, finally, also takes on a new form due to the archiving process, which adds its own set of time dimensions. The complexity of updating frequencies relate to the editable timescales inherent in the materials, while the archiving process add a set of external timescales imposed in the archiving process as a result of decisions taken in this process. Some of these are deliberately chosen, such as the criteria of selection and the time span covered, while others are implicit and may not be known of in advance—or ever—as they are the result of the disconnection of links and scripts and of changes in the materials taking place during the harvest of the materials in question. As a result, the archive may include materials in the same harvest that never existed together on the live web, as well as materials that did actually coexist may be missing.[39] Since web materials are always restored or "replayed" (Duranti and Thibodeau 2006; Taylor 2012) from a server when called for, the call may generate transformations and cannot take into account former appearances of the materials.[40] Thus, web archives are composites of a variety of time horizons: the time horizons of what is told, which may be modified during later additions to the story; the timeline of telling—the on-going editing and the sequence of modifications; and the possibly disturbing timescales of archiving, which both brings closures of open relations (such as interactivity and response presence) and break down link connections, which may lead to disturbances in performance as well as lacking content.

Most web archives can be described as multiple source knowledge systems. They are created by cutting off some of the link relations and time scales related to the surrounding web and thus characterised by a set of closures built in to the archived materials as part of the archiving process itself.

4.4 Strategies combined

In modern society, a complete collection of printed materials was feasible at least in the imagination; for web materials, it is simply not possible. There is no way to archive a complete collection of web materials and the question is, how to combine different archiving methods to ensure the most valuable result?

There is no final answer to this.

The *Internet Archive*, the mother of all web archives, today uses a broad range of harvesting strategies, including harvesting on the level of national domain, regional domain, bulk, selective, event and thematic.[41] The archive also facilitates suggestions of websites from the public. Way back in 1996, they used bulk harvesting collecting simply as much as possible.

The Swedish project, *Kulturar3w* was initiated to create a comprehensive national, domain centric web archive based on bulk harvesting of a few snapshots of the Swedish domain per year (Arvidson, Persson and Mannerheim 2000).[42] Contrary to these and more in accordance with established library traditions, a topic centered project, *Pandora*, aimed to archive a limited number of selected sites due to authority and quality (Koerbin 2017). Event based harvesting was introduced by the *Internet Archive* during the 9.11 terror act in 2001 to collect materials related to unexpected or predictable events resulting in the creation of new pages or the appearance of materials related to the event on unexpected sites somewhere on the web (Schneider and Foot 2003; Webster 2017.)

The strategies emerged as conceptually very different approaches, but they did share a very fundamental limitation as they only "preserve our archiving of the internet in static terms" (Duranti and Thibodeau 2006; Finnemann 2001, 40). As a result, many types of materials would be missing in the archives. This included for instance frequently updated sites (news, web portals, many personal webpages (homepages) chat foray, materials documenting new genres, the development of link structures, digital art

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

forms, and other sorts of frequently updated or dynamic web materials, which would not be included in a canon-based archive at all and often disappear in between two snapshots.[43]

A few years later the national Danish Web Archive, *netarkivet.dk*, developed a more elaborate strategy, which combined domain centered, topic centered archiving, with event-based harvesting and with a stronger focus on the dynamic and time sensitive character of the materials.[44]

Thus, time sensitive selective harvesting due to updating frequencies, rather than canon, was introduced though in very limited scale to include harvesting of non-cumulative, frequently-updated sites within three major areas: news sites, a limited number of other types of popular sites and a limited number of creative end explorative sites whether in respect to social and political communication or artistic creativity and originality. Since such strategies are expensive, as they depend on a high volume of mental labor while bulk harvesting depends on a high volume of machine labor, only a very limited number of sites were actually included. The limitations were imposed for economic reasons. At the time, it was assumed that official sites and canonic sites would appear in the snapshots as they were supposed to be cumulative or had not yet really utilized the dynamic features.

Time sensitivity is crucial in respect to the frequency of updating. It is also crucial in respect to events, which may appear in between two snapshots and may also generate new websites or bring materials on unexpected sites. In the Danish strategy, selections based on various time-dependencies have complemented more traditional criteria of selection of high quality and authoritative sites as a main criterion for selection. The time sensitivities of web materials in the early twenty-first century were far from fully exploited nor fully understood. New forms emerge and the incorporation of multiple timescales in computer games, in multiple source knowledge systems and platforms, eventually exploiting real time data both on local and global scales, forms a major trajectory in the development of new web genres.[45]

Today the most widespread web archiving strategies represent a variety of combinations of mechanized bulk snapshots, selection based on various types of time sensitivity, selection based on criteria for quality and authority of the sources and a growing range of special collections either related to a theme, to specific research projects or to cultural heritage projects. Crowdsourcing and donation of archived sites are also often included.

In spite of the fast-growing range of archiving projects experimenting with a variety of archiving strategies built on a variety of epistemological principles, we do not have studies comparing the different archiving strategies and their coverage and there is as of today no way to monitor (not to speak of curating) the full array of archived web materials.[46] Thus, we cannot tell whether the materials preserved are those worth being preserved. There is also lack of criteria for deciding which materials should be considered worth preserving. This is also the case for the preservation of digital materials more generally. In both cases, society is today confronted with commercial digital information monopolies, the relation to which may pose one of the most vital challenges in the years to come.

5.0 Web archives are always flawed

As already explained, the growing array of archiving strategies cannot hide the fact that web archives are always flawed. Some flaws, as those addressed in Section 1.0, relate to the nature of the web. These flaws also occur as a result of the variety of editable timescales, which can be ascribed to any part of any message. This is not least the case for materials that include real time data.

Some flaws are the result of the very process of archiving as this process will always include broken links. Web materials come as interconnected and interfering materials and have to be carved out by cutting the links to the surrounding part of the web. In so far these links include scripted materials to get content or functionality (images, calculations, quotes etc.) from other sites, these materials will be missing in the archive. This is also the case for scripts activated by individual users, for interactive materials, streaming and other formats, which cannot be archived at the time the materials are published. The archived materials may also be flawed due to the modifications of web materials during the archiving process as materials are deleted or moved to another address taking place in the timespan between the collections of different parts of the materials.[47] Other flaws again are the result of the specific criteria used for selection of materials to be archived as discussed in Section 4.0. Something will always be missing.

Web archives also pose problems with metadata, because the greater part of the materials needs to be harvested automatically on the fly (domain based rather than topic based). At the time of harvesting metadata is mainly limited to include specifications of the materials that are generated automatically during harvest (time stamps, amounts, file types, and similar types of metadata even if the URL's in some cases can serve as metadata too). Monitoring, detailed selection and curating of materials have to take place afterwards, which allow huge amounts of informational trash to be meshed into the archive. Since there is yet no secure method for automatic generation of metadata for the content of the materials, such metadata has to be provided "manually" (i.e., by humans), which is only possible for very small sets of archived web materials.[48] Finally, they are also flawed due to interface issues, as

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

we have no access to the interfaces used on the live web and of course physical problems resulting in informational noise.[49]

Web archives can never be a copy of what was once online. The very act of archiving imply that the archived materials are disconnected from the surrounding web replacing connections on the web by imposing distinctions in the archive defined by the criteria of selection during the archiving process. Rather than collections of copies of the past web, archives should be considered as a particular kind of a "multiple source knowledge system" in its own right, composed to ensure a wide array of traces left of the activities performed on the web and to provide a rich if not complete set of source materials for future studies incorporating a diachronic perspective that cannot be traced on the live web. The issue of trust will always remain, but it will be reduced insofar as the materials are archived and fenced off from the ever oscillating live web, if not in real time then with a minimal delay.

6.0 Alternatives and supplementary strategies?

The establishing of general, often national web archives is not the only method for preservation and organization of the knowledge resources on the web. The development of the web and of web archives has been accompanied by the development of other strategies aiming to optimize and preserve the use of web materials as a knowledge resource. The overarching challenges relate to the constitutional role of hypertext, which is increasingly utilized in ways that turns upside down the original ideas of computational processes.

The notion of hypertext was originally coined by the philosopher Ted Holm Nelson and conceptualized as a means to establish mechanized but relevant semantic connections between all sorts of texts and units of text and other media forms as well (Nelson 1965 and 1993). For Nelson, hypertext was always extrinsic to the text and not part of it, but he also assumed that the relation between an anchor and the content referred to would be fixed. With the idea of a global, interlinked "docuverse," he seemingly took the classical ideal of knowledge organization into the digital realm. If nothing else, the exponential growth of the amounts of web materials would prevent this kind of approach. Ironically, the production of these amounts is not least made possible precisely because of the hypertext architecture of the TCP/IP and web protocols. The "docuverse" is here, in the form of the web as a whole, where everything is interlinked and connected to the same flexible address system—and thus independent of the content. This again allows hypertext to serve both extrinsic and intrinsic relations to a text or any part of it stored randomly at any address. The links reflect an array of different relations between elements among which consistent semantic connections are only a tiny fraction. The complexity is made possible precisely because the links are not simply go-to commands but also may include all sorts of instructions of what to do at the destination.

A related project is the semantic web, initiated by Tim Berners-Lee, the creator of the web protocols, aiming to "bring structure to the meaningful content of web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (Berners-Lee, Hendler and Lassila 2001, 3; Berners-Lee, Shadbolt and Hall 2006). The project is built on the claim that it is possible to automatize semantic analyses of materials to create coherent semantic metadata, which can be used by the machine either by help of an AI inference system or as automatic creation of linked data. Whether this is possible beyond controlled vocabularies within in a formalized semantic universe remains to be seen. In a linguistic perspective, it is difficult to perceive such systems remain stable in a long-term perspective.

The semantic web project relates directly to the online web. In the ARCOMEM project, the focus is on archiving social media sites and the aim is to build content selection mechanisms into crawlers ensuring quality and relevance of a topic archive or an event archive (Risse et al. 2014, 2). It is assumed that social media represent "the wisdom of crowds" and that tools can be built, extracting this knowledge to help archivists in selecting materials for inclusion in an archive. The project apparently is based on the idea that the web is primarily relevant due to social media and community-based archiving. Thus, these archives will reflect only what the social media populations prioritize today. The larger societal perspectives and long-term values are not taken into account.

A third alternative to consider is the suggestion that it is only necessary to preserve the source codes of the webpages. The source code includes valuable information, which tells much about the webpage and its link relations, and it may provide a very useful supplement, but it cannot stand for the page and provide a valid basis for reconstructing old webpages, as they are interpreted and made sensible to humans by help of browsers and editable interfaces.[50]

A fourth strategy is to rely only on topic-centric special collections either for a specific project and limited in time or for a specific theme and eventually on-going time sensitive archiving. This would be archives leaving out materials documenting the on-going developments and not allowing the use of the archives as documentation of the major and broader part of web activities.

A fifth, though supplementary-only, type of strategy are recovery strategies aiming to recover an archive by collecting "evidence of uncrawled pages" from pages that are part of the archive (Huurdeman et al. 2015, 247). The

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

study shows that it is not only possible to uncover the existence of unarchived pages but also to recover significant parts by "reconstructing representations of these pages from the links and anchor text" in the archived pages.

Responding to the overwhelming amounts and the "ruinous" character of general web archives, it has also been discussed whether archiving in the form of print outs of source codes, filming of screens and other non-digital storage formats might be more useful and eventually also could be made for a lower cost. Such efforts alone would somehow reinforce the limitations of web archives while the values would be missing. General web archives based on domain harvesting will never fully replace special, curated collections, research infrastructures and other repositories for digital materials. On the other hand, curated collections on their side cannot replace broad domain-based archiving.

It might be argued that web archives after the spread of mobile media and the advent of a range of non-web based digital media platforms (mobile apps) are not as sufficient (nor as central) as before, and efforts to connect web archives with non-web collections are needed. General web archives, however, remain a unique source in respect to a range of purposes. They will, to a high degree, deliver as a trusted source for documentation of past events and activities not simply as a source for the history of the web but for the history of society and the cultural practices, which are increasingly enacted on web based or web related platforms. Their value will grow as the materials are accumulated and the archives will increasingly also be the only source left.

Considered as a special type of knowledge organization, they may be useful also for a range of new kinds of analyses, as the materials may document both long tail and long-term patterns in the archive as a whole or in any sort of delimited, frozen web sphere within the archive. They also have the advantage that they can be used to document the emergence of new genres and practices before they are fully recognised and included in topic centric archives. Thus, they may also fill out many gaps and empty spaces between special collections, research infrastructures and other kinds of curated repositories. They may, furthermore, serve as a new, unique kind of index to history and culture of the societies, and as index to other knowledge organization sources as they develop in the future. Since most general archives are national archives, there is a need to facilitate interoperability in between these and the internet archive and other on-going archiving initiatives.

7.0 Web archives are multiple source knowledge organization systems

The principles of general web archives are seldom discussed from the perspective of knowledge organization. There are reasons for this. First, if KO is primarily focusing on the systematic documentation of resources with a strong focus on metadata, general web archives will remain in the margin, because most of the materials have to be collected and preserved automatically. The size alone makes traditional methods for cataloguing "too time consuming and expensive" (Costa, Gomes and Silva 2017, 193). They argue for automatic indexing.

The most widely used format for storing the materials is the WARC format, which was designed for this purpose and established as an ISO standard in 2009.[51] The crawler used to harvest will normally also collect a set of metadata in the same automatic process. These metadata, however, will always be insufficient and mainly related to architectural relations between the stored objects.[52]

The WARC format only includes minimal information on the nature of the distinct object types and content and does not include sufficient information on the provenance of the materials, the principles for the selection of the initial set of URLs and other kinds of contextual information including known limitations, errors and so forth. Human curation may add information on the level of the corpus harvested but is not feasible on the level of a website or webpage. The needs for metadata furthermore depend on and vary with particular research questions and methods to be used.

For general web archives, a main task is to collect and preserve the heterogeneous nature of these materials in respect to the variety social, and cultural and political practices. Such archives include primarily materials, which are not yet analysed or established as knowledge, and can only be considered a possible source for future knowledge production. However, they are collected and organized for this purpose. Each archive is built according to a specific set of principles (though changing over time) for selection, preservation, presentation (knowledge visualization) and search facilities provided by the particular archive. These principles represent a particular type of knowledge organization, which organize source materials for a huge variety of possible research projects.[53] Each research project will generate an array of results based on a specific selection of primary sources within the archive eventually combined with other sources. Such projects can be anchored in different epistemological principles, methodologies and possibly related to a range of different domains whether these overlap or not. The knowledge produced on the basis of these materials may belong to many different topic-domains and enter into other KOSs. They may also—if facilitated—deliver valuable metadata back to the archive. Web-archives may serve other purposes as well, but a main role is to preserve primary source materials for trusted citation, historical documentation and future research.

Second, general web archives contain some of the most complex types of digital materials hitherto known and can-

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

not be appropriately described within the vocabularies of previously developed KOSs. The reasons for this are the hypertext character of networked digital media and the complexities added in the archiving process. Thus, there is no way to describe web materials and archived web materials within a conceptual framework, which does not bring hypertext, rules and codes as part of individual messages, interactivity, time sensitivity, windows of interaction and many other—coded—dynamic features of electronic texts into the fore. All elements in these materials can be remixed or coded as time sensitive, and they may include coded links and scripts, thus also disturbing any permanent distinction between program and data. Programmes are produced, circulated, treated and executed as data and the processes are always initiated by humans.[54] This is the case even if such processes are performed via long chains of automated and responsive sequences as in "self-driving" cars.

The relation between data materials and analytical tools is closer than between print materials and methods applied to the analyses of these, because digital materials can only be accessed via some sort of search facility, which will also be a point of departure for the methodologies applied. On the other hand, digital materials always also allow for the application of new search entrances representing epistemological principles different from those applied in the first instance. The materials used for one type of knowledge production may later be used for other types. Thus, general web archives do not belong to one particular domain.

In a discussion of the implications of big data for knowledge organization, Ibekwe-SanJuan and Bowker (2017) argue that big data create a need to rethink the standpoint from which the KOSs are designed. As indicated in the title, the source of the requirements to rethink the principles of KOSs is the spread of "big data," which is conceived of as complex and always imperfect and often lacking adequate metadata. If so, web archives qualify to be included, and the question is whether their suggestions to rethink the principles of KO also apply to web archives.[55]

First, they suggest a move from apodictic to faceted, flexible schemas in order to take into account the fast-growing amounts and huge variety of new, often more complex kinds of data produced. It is not clear yet, however, whether faceted and domain-oriented schemas are sufficient to take into account the complexities of timescales, links and scripts as they appear on the web and in the archived web materials.

Second, they argue there is a need to take into account the changing nature of data output. This is in accordance with the preceding analysis, though big data sources if they include real time data with updating frequencies measured in seconds or less will have to be made subject of a specialised archiving strategy, reflecting these particular time frequencies. Thus, there is a need for a more elaborate conceptualization of the data captured in respect to metadata and whether and how it can be archived at all. The questions include how data are captured and processed until the archiving, itself a kind of recapturing, takes place, how they are composed in respect to links, scripts, updating frequencies, interfaces—in short to their hypertext configuration (Finnemann 2017)—how they are harvested due to what sort of archiving strategy and how they are made accessible and searchable in the archive. The question what the data are about apply of course to the archiving strategy of topic centric archives. For domain centric archiving, this question is left to later research.

Third, they argue for turning around from "purely universalist and top down approaches to more descriptive bottom up approaches" that can include a variety of perspectives. This suggestion is closely connected to the fourth element in their rethinking as they see a methodological need for combining automated techniques on the one hand and amateur crowdsourcing methods on the other. Both approaches are bottom up. This is maybe the most problematic issue in their rethinking, as a bottom-up approach to the internet seems to be nearby impossible due to the dynamic, interlinked and systemic architecture. The history of web archiving is of course—as many older global knowledge systems—generated by a series, more or less coordinated "local" initiatives, but insofar as they collect information from globally distributed sources they transcend the situated character (Edwards 2017). If the bottom up strategies for collection are limited to automated collection (snapshots eventually combined with pattern analyses tools, counting of incoming links, and tags etc.) and crowdsourcing based on for instance social media, the archives will be idiosyncratic reflecting primarily activist minorities and the "zeitgeist" of today. Such strategies may be helpful, but they are neither capable to deal with the complexities and time sensitivity of the materials, nor with the global and long-term perspectives of the future in which they are to be used.

Their argument is to a high degree built on Hjørland's (2012, 2013) critique of universal bibliographical classification schemes, the neglect of subject knowledge and the reluctance within the KO community to include data analysis techniques "as an alternative to manually constructed KOS's" (Ibekwe-SanJuan and Bowker 2017, 189).

As it has been shown in the preceding analysis of one particular set of big data, general web archives, this rethinking will not only need to include the role of human expertise in the production of "good metadata" and inclusion of amateurs in crowdsourcing, it also requires a more elaborate conceptualization of the data materials reaching

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

far beyond the notion of data, whether raw or not, given or captured. While a universalist perspective is not available there is a need for a general perspective beyond the "local" and situated bottom-up-perspectives. One might even argue that situated perspectives are becoming increasingly inappropriate precisely because of the spread of internet-based communication, which is characterised by the constantly on-going connections mixing multiple and fluctuating situations into each other across the globe. Since the links are part of the electronic text, any two or more situations may be conflated in time while remaining distant in space. This is why national web archives and all kinds of archives should be designed to collaborate and thought into a globalized system of all sorts of KOSs. The global perspective is itself a local perspective within the biosphere, which forms a tiny part of the cosmos, but is transcendental to personal, situated human experience. The very act of web archiving and the building of general web archives at the same time also undermine the notion of "the situation" as an epistemological platform as they cannot but refer to a global context—Facebook and many other agencies are globally present agencies taking part in the on-going interactive communication processes all over—and to an unknown future if we are to make sense of these archives. In spite of the deconstruction of the archive in postmodern philosophy (Derrida and Prenowitz 1995), written during the transition from printed to digital archives, at the time of the creation of the first web archives and other digital archives and KOSs—not least those needed for dealing with global issues—archives and collections seem to survive or even transcend the limitations of postmodern social constructions.

The multiplicity of interconnected and conflated situations on the internet should rather lead to condense scientific and scholarly thinking into globalized, non-universal and general perspectives. There should be no single paradigm for KO. Rather, they should stretch from clearly specified and closed KOs to ever evolving general web archives, which may both serve as a KO in itself and as an index to an otherwise incomprehensible set of KOs and to all sorts of societal cultural practices. Consistency in the organization of human knowledge, even if limited to scholarly and scientific knowledge, may remain the ideal, but it is not an option, and it is not necessary, since anything can be incorporated and made searchable in a networked system of hypertexts.

In the twenty-first century, exponentially growing amounts of digital materials are immersed in a globalized multi-leveled and hierarchized hypertext landscape—and there is a need for further analyzing the implications for the development of KOSs, not simply the multiple source and partly real time-based systems but the whole array of new formats for the range of possible KOSs. The internet and particularly the WWW and related networks is not simply a means of distribution or a platform for interaction. It is increasingly significant as the "docuverse" within which culture and society takes place, as a growing range of agencies articulate a growing range of their activities in a growing range of genres by help of a growing range of digital media.

If knowledge organizations are used to model our knowledge of the world, they need to be capable also to monitor and to track changes both globally and over longer periods of time. The time sensitivity of the web as a whole and of web archives may be seen as a paradigm or prototype for future KOSs.

Notes

1. The distinctions between first, second and third wave of digitization refer to predominant ideas related to the development of mainframe computers, desktop computers and networked digital media respectively. Today they form three significant paradigms of digital materials, the first characterised by the distinction programme-data and the automated execution of rules; the second characterised by man-machine interaction (HCI, CSCW) and the third characterised by networked digital materials including both interaction between networked machines, HCI and between connected humans (Finnemann 2014a).
2. If big data methods are applied to large fractions of the web they will need to build on statistical analyses based on a limited number of predefined indicators across a huge variety of semiotic regimes (e.g., math, images, diagrams, many different spoken and typed languages). For limitations of big data analyses see e.g., Boyd and Crawford (2011); Moretti (2013); Kitchin (2014); Gatto (2014); Ibekwe-SanJuan and Bowker (2017).
3. The size of the web and the fluctuations of web materials make it very complicated to measure the lifetime of web materials. However, the various methods used all lead to the same general conclusion that most web materials are either modified moved or erased within a year or even a shorter period. See among others Mannerheim (2000); Lyman and Varian (2003); Masanès (2006, 2); Hilbert and Lopez (2012); Pennock (2013); Brügger (2018, 55).
4. Many advantages are well known. Today a pertinent question is whether there are also too many or strong disadvantages related, e.g., to hacking, and other forms of subversive economic, political and cultural activities.
5. Masanès (2005, 72-74) identifies changes in "authorship form," "content shaping," "convergence" and "technique" as four major factors making web archiving more

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

complex than archiving manuscripts and printed documents.

6. The array of specialised web archives, which focus on a particular topic, a single purpose, and eventually for a limited period of time, is only marginally touched upon.
7. *UNESCO Charter on the Preservation of Digital Heritage*: http://portal.unesco.org/en/ev.php-URL_ID=17721%26URL_DO=DO_PRINTPAGE%26URL_SECTION=201.html. The charter also gives a hint on what should be kept in article 7: "As with all documentary heritage, selection principles may vary between countries, although the main criteria for deciding what digital materials to keep would be their significance and lasting cultural, scientific, evidential or other value. 'Born digital' materials should clearly be given priority. Selection decisions and any subsequent reviews need to be carried out in an accountable manner, and be based on defined principles, policies, procedures and standards."
8. Thus, networked digital media turns upside down the character of digital materials as defined within a single file or single machine perspective. See Kirschenbaum, Ovenden and Redwine (2010) for an analysis of issues pertinent to the archiving of born digital files stored on a computer hard disk, including issues concerning the particular physical devices used in the production and eventually in the circulation. The issues considered relate to texts, video and audio files, but does not include issues related to interferences between internet connected machines that forms the basis for interactivity, multiple source systems and the configuration of multiple time scales within a given webpage. In the single-machine-and-closed-file world, complete archiving is feasible and may even include hidden information stored in the machine or in the browser history (Kirschenbaum et al, 33).
9. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives. Last updated 21 December 2018.
10. http://netpreserve.org/about-us/members/
11. The Common Crawl Foundation is a non-profit organization founded in 2007. Commoncrawl's data are located on Amazon S3 as part of the Amazon Public Datasets program from which anyone can download the files entirely free. https://aws.amazon.com/public-datasets/
12. Archive-the-net: http://archivethe.net/en/index.php . Internet Memory Foundation, established 2004, http://internetmemory.org/en/ and European Internet Memory Research, https://internetmemory.net/en/, established in 2011.
13. In a recent survey focusing on metadata for web archives, it is suggested to create a hybrid type of metadata that combines archival and bibliographic metadata practices "as new types of digital content permeates our collections" OCLC 2018, Vol 1: 8. See also note 52.
14. Seadle and Greifender (2007,169) quoting National Science Foundation (1999) *Digital Libraries Initiative: Available research*. US Federal Government. Source given, but not found: http://dli2.nsf.gov/dlione/ The same quote is also found in Richard E. Jones, Theo Andrew, John MacColl (2006). The Institutional Repository. Oxford: Chandoras Publishing (2006, 5). Source referred to as "the NSF website." This is reference rot, but in this case the Source can be found in archive.org: https://Web.archive.org/Web/19991007203722/http://www.dli2.nsf.gov:80/dlione/
15. To solve issues related to the complexity of external links, the report stresses the need for contextual metadata. The issue of external links will be further discussed below.
16. The basic hypertext function is the go-to relation between an anchor and a destination. Since the go-to is mechanized, it will always include an operation, a "to do" instruction that make hypertexts different from, for instance, a foot-note reference or an index, which can be described as proto hypertext formats (cf., Hjørland 2018; Finnemann 2017). Since the instructions of what to do at any set of destinations can be deliberately composed, they specify the degree of complexity of the data materials in question. The degree of complexity also depends on whether hypertext is limited to function within a single file, or on a standalone machine that includes the possibility to modify the functional architecture of that machine, or whether it is applied to networked machines with a shared address system, which in principle allow any user to interfere with any element on any other machine.
17. Robot.txt is a *de facto* standard based on consensus within the WWW developer community in the early 1990s and with no juridical back up, see http://www.robotstxt.org/orig.html . Since 2013, there has been an ISO standard for web archiving, which "defines statistics, terms and quality criteria for Web archiving. It considers the needs and practices across a wide range of organisations such as libraries, archives, museums, research centres and heritage foundation" quoted from https://www.iso.org/standard/55211.html
18. *Library of Congress Update on the Twitter Archive at the Library of Congress. December 2017*. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf The update includes links to relevant documents on the archive principles and practices.
19. A list of countries with/without legal deposit laws for web archives can be found at The International Internet Preservation Consortium (IIPC) web pages:

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

http://netpreserve.org/Web-archiving/legal-deposit/ The list also gives information on the—very different—conditions for accessing the archives. For an analysis of copyright issues and legal deposit web archives using Singapore as case, see Cadavid (2014). For summaries of the development of different national web archives, see, e.g., Koerbin (2017) for *Pandora*, Australia. For Denmark, see Schostag and Fønss-Jørgensen (2012). For Croatia, Holub and Rudomino (2015). The Internet Memory Foundation performed a survey on web archiving in 2011, which gave an overview of the state of art concerning legislation, access, methods of harvesting etc. See: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf

20. Quote from Internet archive, front page, https://Web.archive.org/
21. The issue of "trust" of born digital content or heritage circulated as closed files on the internet is discussed in Kirschenbaum et al. (2010), emphasizing the relation between authorship and trust.
22. The postmodern dissolution of history and critical analyses of the power structures inherited in museums and libraries and archives, (e.g., Derrida and Prenowitz 1995) may have weakened the position of these institutions as authoritative knowledge organizations and facilitated their opening towards broader audiences.
23. There is a growing awareness of the need for web archives among historians. According to Milligan (2016, 80) "This is not an abstract concern: the history of the 1990s will be written soon" and also identify one the unique characteristics, that broad web archives "represent a massive collection of non-elite speech" (Milligan 2016, 78).
24. The limitations relate, among other things, to "the ever increasing size and rapidly changing content" (Huurdemann et al. 2015, 248) as well as the intrinsic characteristics of web materials, of archiving and preservation methods, and of the archive interface to the materials, cf. Schafer, Musiani et Borelli (2016). The issue is also dealt with in Section 5.0.
25. A short list of relevant market failures are mentioned in Blue Ribbon Appendix 3 "When Markets Do Not Work," 91-92. The role of Proxies is mentioned in appendix 5 "The Role of stakeholder Interests," 96-98.
26. Some web archives often limit the number of harvested site levels to the top levels. This, of course, reduces the value, but still allow the archive to function as a historical index of websites and implicitly of the agencies and an array of societal interrelations.
27. Among a growing range of strategies for statistical analyses of large cultural datasets see for instance Christakis and Fowler (2009); *Moretti (2013*); Aiden and Michel (2013); Kitchin (2014). So far, the methods are still on the bench to be further validated, but they are far from being dismissed. A major issue is whether traditional samplings are less valuable as the "all data available" approach that allow for the inclusion of outliers, which would be dismissed in sampling and thus provide richer and more nuanced results. For a study of linguistic web corpora see Gatto (Ed.) (2014). A second issue is whether it is possible to move beyond the indexical or indicative coding schemes to semantic and meaning full interpretations.
28. The European Holocaust Research Infrastructure EHRI: https://www.ehri-project.eu/
29. Masanès (2006) distinguishes between site centric, topic centric and (web-) domain centric archives (Masanès 2006, 41-43; Brügger 2018, 73-85).
30. For an analysis of such intricacies in Google's Usenet-web archive, see Paloque-Berges (2017, 229-51).
31. National domain addresses, however, are insufficient, because materials of relevance for any society can be found on many sites outside a particular national domain. Domain centric archives, therefore, also need supplementing strategies, which have to be less systematic and to be topic centric, dependent both on the conceptualizations of national relevance and of resources to identify relevant materials on other domains. One would expect leading agencies in the field to develop a more comprehensive general strategy by coordinating domain centric harvesting of national domains and other domains.
32. The relationship is manifested in web- and social media activities of politicians and legacy media. Recent studies, furthermore, shows that younger generations ("millenials") increasingly get news from a variety of sources, legacy media included, via Facebook. See The Media Insight Project (2015). *How Millennials Get News: Inside the Habits of America's First Digital Generation.* The Media Insight Project is a collaboration between the American Press Institute and the AP-NORC Center.
33. In Finnemann (2017) and (2018), internet based multiple source knowledge systems (MSKS) are described in respect to a variety of parameters. The notion networked knowledge organization (NKO) is not available, as it is currently used for the utilization of the web-based internet as an environment for digital libraries. In that perspective, hypertext is conceived of as a navigational tool for facilitating multiple access forms to established KOs extrinsic to the materials, while hypertext and scripts may also be intrinsic in web archive materials (cf. Gail Hodge 2000). See also the NKO homepage at http://nkos.slis.kent.edu/.
34. Domain-centric bulk harvesting is sometimes practiced as a broad and surface-oriented method delimited only to capture one or two top-levels of a site and opposed

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

to topic centric in depth harvesting of full sites. In other cases, more levels are included to ensure that more sites are harvested in full depth. A second aspect of depth relates to the so-called "deep" and "dark" web. The deep web includes websites, which are not indexed or made inaccessible for search engines. The dark web is a grey zone within the deep web, which is made more difficult to enter by requiring specific software, specific configurations or other kinds of filters. It's a grey zone, because some of the activities performed may be legitimate but private, while others are illegitimate by law or considered illegitimate for political reasons.

35. For an overview of combinations of archiving strategies in general web archive practices today, see Gomes et al. (2012) and the website of IIPC (The International Internet Preservation Consortium).
36. See, for instance, Halevy, Nordvig and Pereira (2013).
37. Anecdotic evidence, for instance, in Kahle (1997).
38. This and other examples of multiple source systems based on networked digital media are discussed in Finnemann (2017; 2018).
39. The relation between the ephemeral and persistent character of web materials and the implications for web archiving is also discussed in Schneider and Foot (2005), Masanès (2006), and Brügger (2005). Masanès (2006, 13) describe how the cardinality of books "at least were unified from creation to access" while web materials located at a server even if they "have a unique identifier, … can be generated virtually infinitely and undergoes some degree of variation for each of its instantiations."
40. Taylor (2012). The newly published OCLC-report on *Descriptive Metadata for Web Archiving* (OCLC 2018) describe the archiving process as "highly transformative," because the process "changes the very nature of the resource: each crawled version becomes a fixed object, preserved for the future in a particular location and associated with any other versions that have been captured" (vol. 1: 9).
41. As reported at the IIPC member site: http://netpreserve.org/about-us/members/internet-archive/
42. The article presents their delimitation of the Swedish web (the domain .se + generic top-level domains with a Swedish address or phone number). They also introduce time sensitive harvesting of newspapers and identify the existence of materials collected in the same harvest, which did not ever exist at the same time on the web.
43. The examples are drawn from Finnemann (2001, 33-39).
44. The Danish case is documented in Christensen-Dalsgaard et al. (2003). The report includes (46) an internet related definition of materials of relevance for a national Danish web-archive located outside the national domain (so called "Danica"). The strategy suggested was carried forward into the Danish Legal Deposit Law of 2004 (http://pligtaflevering.dk/loven/) in which it was also stated that the archived materials should be considered cultural heritage. See also *Udredning om Bevarelsen af Kulturarven* (Report on the Preservation of Cultural Heritage) requested by the Danish parliament (http://www.kulturarv.dk/fileadmin/user_upload/kulturarv/museer/Bevaring_af_Kulturarven_1_.pdf). See also Schostag and Fønss-Jørgensen (2012), Finnemann (2001) and Brügger (2001). For a detailed discussion of how to delimit a "national" web domain, see Brügger (2017c). For an English version of *netarkivet.dk*, see http://netarkivet.dk/in-english/
45. Knorr Cetina, 2009; UN Sustainable development goals (2015); Steffen et al. 2015; Finnemann 2017; Edwards 2017;
46. Brügger (2005); Masanès (2006); Jinfang Niu (2012); Pennock (2013); Gomes, Miranda and Costa (2011); Laursen and Møldrup-Dalum (2017); Gorsky (2015); O'Carroll et al. (2013); Risse et al. (2014); Plachouras et al. (2014); Saad (2009).
47. Arvidson, Persson and Mannerheim (2000); Masanès (2006); Brügger and Finnemann (2013); Brügger (2017c). See also Koehler (2004), Day (2006), Klein et al. (2013); Liepler and June (2013); and Massicotte and Botter (2017) for more detailed studies of "linkrot" and "reference rot" in web materials and web archives.
48. The Semantic web project (Berners-Lee 2001 and 2006) is probably the most well-known project aiming to remedy this limitation but focusing mainly on formalized and thus closed semantic spaces. The ARCOMEM project (Risse 2014; Plachouras 2014) initiated within the "Future Internet Initiative" aims explicitly to automatize the collection of semantic information during the crawling process.
49. For analyses of the lack of archivability, see, e.g., Duranti and Thibodeau (2006), Zierau (2011) Kelly et al. (2013).
50. For a more elaborate discussion, see Brügger (2017a). For archiving of websites using source code, see Helmond (2017).
51. The WARC format was developed within the IIPC community and stores the harvested data in an aggregate file, a container format that can include a wide array of data object types and also include metadata related to the harvest, eliminate duplicates and manage some forms of data transformations and to a high degree ensure the reproducibility of the webpages. The ISO standard was last revised in 2017, available at https://www.iso.org/standard/68004.html. For a brief overview, see Li-

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

brary of Congress website. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

52. A recently published OCLC 2018 report on descriptive metadata for web archiving describes current metadata practices as characterised by a range of inconsistencies. The report identifies three patterns: "1: Existing descriptive standards generally do not address the unique characteristics of either live or archived Websites. 2: Institutional metadata guidelines vary widely in both the elements included and in the choice of content within those elements. 3: Some metadata practitioners follow bibliographic traditions, others take an archival approach (such as describing a collection of sites in a single metadata record), and hybrid approaches combining characteristics of both are common.' OCLC 2018 Vol. 1: 13. The report aims to provide a metadata standard for Web archives build on a combination of librarian (typically single title oriented) and archival (typical collection oriented) principles.
53. The knowledge organization of Web archives also include the collections principles and strategies as well as the visualization facilities and the organization of search facilities, which on the other hand are connected to the on-going development of research methods and related analytical software tools. These dimensions are not further addressed in this article.
54. See Brügger and Finnemann (2013); Sim and Gallardo-Valencia (2013); Brügger (2018); Finnemann (2017) and (2018).
55. There is no precise definition of big data. Web archives, however, fit to most characteristics such as high volume, variety, messiness and volatility except for real time (velocity). Big data are not necessarily real time systems. Such systems, however, will require a different kind of archiving and preservation strategy (Duranti and Thibodeau 2006; Boyd and Crawford 2011; Kitchin 2014).

References

Abbate, Janet. 1999. *Inventing the Internet*. Cambridge, MA: MIT Press.

Aiden, Erez and Jean-Baptiste Michel. 2013. *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books.

Arvidson, Allan, Krister Persson, and Johan Mannerheim. 2000. "The Kulturarw3 Project: The Royal Swedish Web Archiw3e: An Example Of 'Complete' Collection of Web Pages." Paper read at *66th* IFLA Conference. https://archive.ifla.org/IV/ifla66/papers/154-157e.htm

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web; A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities." *Scientific American* 284, no. 5: 34-43.

Berners-Lee, Tim, Nigel Shadbolt, and Wendy Hall. 2006. "The Semantic Web Revisited." *IEEE Intelligent Systems* 21, no. 3: 96-101. doi:10.1109/MIS.2006.62

Blue Ribbon Task Force. 2010. "Sustainable Economics for a Digital Planet: Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access Ensuring Long-Term Access to Digital Information." http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Borgmann, Christine L. 1999. "What are Digital Libraries? Competing Visions." *Information Processing and Management* 35: 227-43. doi:10.1016/S0306-4573(98)00059-4

boyd, danah and Crawford, Kate. 2011. "Six Provocations for Big Data." Paper read at A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011. doi:10.2139/ssrn.1926431

Brügger, Niels and Niels Ole Finnemann. 2013. "The Web and Digital Humanities: Theoretical and Methodological Concerns." *Journal of Broadcasting & Electronic Media* 57, no.1: 66-80. doi:10.1080/08838151.2012.761699

Brügger, Niels, ed. 2017a. *Web 25: Histories from the First 25 Years of the World Wide Web*. Digital Formations 112. New York: Peter Lang.

Brügger, Niels. 2001. "The Last Page on the Internet?" In *Preserving the Present for the Future: Proceedings Conference on Strategies for the Internet, 18-19 of June, 2001*. Copenhagen: Danmarks Elektroniske Forskningsbibliotek, 43-54.

Brügger, Niels. 2005. *Archiving Websites: General Considerations and Strategies*, trans. Stacy Cozart and Patricia Lunddahl. Aarhus: The Centre for Internet Research.

Brügger, Niels. 2010, "The Future of Web history." In *Web History*, ed. Niels Brügger. New York: Peter Lang, 349-53.

Brügger, Niels. 2017b. "Connecting Textual Segments: A Brief History of the Hyperlink." In *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 3-28.

Brügger, Niels. 2017c. "Probing a Nation's Web Domain: A new Approach to Web History and a New Kind of Historical Source." In *The Routledge Companion to Global Internet Histories*, ed. Gerard Goggin and Mark McLelland. New York: Routledge, 61-73.

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, MA: MIT Press.

Buckland, Michael K. 1992. *Redesigning Library Service: A Manifesto*. Chicago, IL: American Library Association.

Bush, Vannevar. 1945. "As we may Think." *Atlantic* 176, no. 1: 101-8.

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Cadavid, Jhonny Antonio Pabón. 2014. "Copyright Challenges of Legal Deposit and Web Archiving in the National Library of Singapore." *Alexandria* 25 nos. 1/2:1-19. doi:10.7227/ALX.0017

Cetina, Karina Knorr. 2009. "The Synthetic Situation: Interactionism for a Global World." *Symbolic Interaction* 32: 61-87. doi:10.1525/si.2009.32.1.61

Christakis, N.A. and James H. Fowler. 2009. *The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York: Little, Brown & Company.

Christensen-Dalsgaard, Birte, Eva Fønss-Jørgensen, Harald von Hielmcrone, Niels Ole Finnemann, Niels Brügger, Birgit Henriksen, and Søren Vejrup Carlsen. 2003. *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001: Final Report for The Pilot Project "netarkivet.dk."* Copenhagen: netarkivet.dk. http://netarkivet.dk/wp-content/uploads/webark-final-rapport-2003.pdf

Costa, Miquel, Daniel Gomes, and Mário J. Silva. 2017. "The Evolution of Web Archiving." *International Journal of Digital Libraries* 18: 191-205. doi10.1007/s00799-016-0171-9

Day, Michael. 2006. "The Long-Term Preservation of Web Content." In *Web Archiving*, ed. Julien Masanès. Berlin: Springer, 177-99.

Derrida, Jacques and Eric Prenowitz. 1995. "Archive Fever: A Freudian Impression." *Diacritics* 25, no. 2: 9-63.

Dooley, Jackie and Kate Bowers 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Dublin, OH: OCLC Research.

Dooley, Jackie and Mary Samouelian. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools*. Dublin, OH: OCLC Research.

Dougerthy, Meghan and Eric T. Meyer. 2014." Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs." *Journal of the Association for Information Science and Technology* 65: 2195-209.

Dougherty, Megan, Eric T. Meyer, Christine M. Madsen, Charles van den Heuvel, Arthur Thomas, and Sally Wyatt. 2010. *Researcher Engagement with Web Archives: State of the Art; Joint Information Systems Committee Report, August 2010*. https://ssrn.com/abstract=1714997.

Duranti, Luciana and Kenneth Thibodeau. 2006. "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES." *Archival Science* 6: 13-68.

Duranti, Luciana and Patricia C. Franks. 2015. *Encyclopedia of Archival Science*. Lanham, MD: Rowman & Littlefield.

Edwards, Paul N. 2017. "Knowledge infrastructures for the Anthropocene." *The Anthropocene Review* 4: 34-43. doi:10.1177/2053019616679854

Finnemann, Niels Ole. 2001. "Internet: A Cultural Heritage of Our Time." In *Proceedings from The Conference Preserving the Present for the Future, Strategies for the Internet, the Royal Library, Copenhagen, 18th-19th of June 2001*. Copenhagen: Danmarks Elektroniske Forskningsbibliotek, 31-42.

Finnemann, Niels Ole. 2014a. "Digital Humanities and Networked Digital Media." *MedieKultur* 30, no. 57: 94-114.

Finnemann, Niels Ole. 2014b. "Research Libraries and The Internet: On the Transformative Dynamic Between Institutions and Digital Media." *Journal of Documentation* 70: 202-20. doi:10.1108/JD-05-2013-0059

Finnemann, Niels Ole. 2017. "Hypertext Configurations: Genres in Networked Digital Media." *Journal of the Association for Information Science and Technology* 68: 845-54. doi:10.1002/asi.23709/full

Finnemann, Niels Ole. 2018. "E-text." In *Oxford Research Encyclopedias Literature*. Oxford: Oxford University Press. doi:10.1093/acrefore/9780190201098.013.272

Gatto, Maristella, ed. 2014. *Web as Corpus: Theory and Practice*. Corpus and Discourse. London: Bloomsbury.

Gomes, Daniel, João Miranda, and Miguel Costa. 2011. "A Survey on Web Archiving Initiatives." In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011 Proceedings, ed. Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt.*. Lecture Notes in Computer Science 6966. Berlin: Springer, 408-20. doi:10.1007/978-3-642-24469-8_41

Gorsky, Martin. 2015. "Sources and Resources: Into the Dark Domain: The UK Web Archive as a Source for the Contemporary History of Public Health." *Social History of Medicine* 28: 596-616.

Halevy, Alon, Peter Nordvig, and Fernando Pereira. 2009." The unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 no. 2: 8-12. doi:10.1109/MIS.2009.36

Helmond, Anne. 2017. "Historical Website Ecology: Analyzing Past States of The Web Using Archived Source Code." In *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 139-55.

Hilbert, M., and Lopez, P. 2012. "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information? " *International Journal of Communication* 6: 936-79.

Hjørland, Birger. 2012. "Is Classification Necessary After Google?" *Journal of Documentation* 68: 299-317. doi:10.1108/00220411211225557

Hjørland, Birger. 2013. "Theories of Knowledge Organization-Theories of Knowledge." *Knowledge Organization* 40: 169-81.

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Hjørland, Birger. 2016. "Knowledge Organization (KO)." *Knowledge Organization* 43: 475-84. doi:10.5771/0943-7444-2016-6-475

Hjørland, Birger. 2018. "Indexing: Concepts and Theory." *Knowledge Organization* 45: 609-39. doi:10.5771/0943-7444-2018-7-609

Hodge, Gail M. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: Digital Library Federation.

Holub, Karina and Ingeborg Rudomino. 2015. "A Decade of Web Archiving in The National and University Library in Zagreb." Paper read at IFLA WLIC 2015. http://library.ifla.org/1092/1/090-holub-en.pdf

Huurdeman, Hugo C., Jaap Kamps, Thaer Samar, Arjen P. de Vries, Anat Ben-David, and Richard A. Rogers. 2015. "Lost but Not Forgotten: Finding Pages on The Unarchived Web." *International Journal on Digital Libraries* 16: 247-65. doi:10.1007/s00799-015-0153-3

Huurdeman, Hugo C., Jaap Kamps. 2018. "A Collaborate Approach to Research Data Management in a Web Archive Context." In *Research Data Management: A European Perspective,* ed. Jesper Boserup and Fillip Kruse Thestrup. Berlin: de Gruyter, 55-78.

Ibekwe-SanJuan, Fidelia and Geoffrey C. Bowker. 2017. "Implications of Big Data for Knowledge Organization" *Knowledge Organization* 44: 187-98. doi:10.5771/0943-7444-2017-3-187

Jenkins, Henry. 2006. *Convergence Culture: Where Old and New Media Collide*. New York: New York University Press.

Jinfang Niu. 2012. "An Overview of Web Archiving." *D-Lib Magazine*18, no 3/4. doi:10.145/march2012-niu1

Jones, Richard E., Theo Andrew, and John MacColl. 2006. *The Institutional Repository*. Oxford: Chandos Publishing.

Kahle, Brewster. 1997. "Preserving the Internet: An Archive of The Internet May Prove to Be a Vital Record for Historians, Businesses and Governments." *Scientific American* 276, no. 3: 82-3.

Kelly, Mat, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson. 2013." On the Change in Archivability of Websites Over Time." In *Research and Advanced Technology for Digital libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013 Proceedings,* ed.Trond Aalberg, Christos Papatheodoro, Milena Dobreva, Giannis Tsakonas, and Charles J. Farrugia. Lecture Notes in Computer Science 8092. Berlin: Springer. doi:10.1007/978-3-642-40501-3_5

Kirschenbaum, M., Richard Ovenden, and Gabriela Redwine. 2010. *Digital Forensics and Born-digital Content in Cultural Heritage*. Washington, DC: Council on Library and Information Resources.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, CA: Sage.

Klein, Martin, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2013. "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot." *PLoS ONE* 9, no.12: e115253-53. http://www.doc88.com/p-7764456934471.html

Koehler, Wallace. 2004. "A Longitudinal Study of Web Pages Continued: A Consideration of Document Persistence." *Information Research* 9, no. 2.: paper 174. http://InformationR.net/ir/9-2/paper174.html

Koerbin, Paul. 2017. "Revisiting the World Wide Web as Artefact: Case Studies in Archiving Small Data for the National Archive of Australia's Pandora Archive." In *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 191-206.

Laursen, Ditte and Per Møldrup-Dalum. 2017. "Looking Back, Looking Forward. 10 Years of Development to Collect, Preserve, and Access the Danish Web."In *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 207-29.

LeFurgy, William. 2015. "Web Archiving." In *Encyclopedia of Archival Science,* ed. Luciana Duranti and Patricia C. Franks. Lanham, MD: Rowman & Littlefield, 413-6.

Liebler, Raizel and Liebert June. 2013. "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 Journal of Law and Technology* 15, no. 2: article 2. http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1085&context=yjolt

Lyman, Peter and Hal R. Varian, eds. 2003. "How Much Information? 2003." http://groups.ischool.berkeley.edu/archive/how-much-info-2003/printable_report.pdf

Mannerheim, Johan. 2006. "The WWW and Our Digital Heritage." Paper read at 66th IFLA Conference. https://archive.ifla.org/IV/ifla66/papers/158-157e.htm

Masanès, Julien. 2005. "Web archiving Methods and Approaches: A Comparative Study." *Library Trends* 54, no. 1: 72-90.

Masanès, Julien, ed. 2006. *Web Archiving*. Berlin: Springer.

Massicotte, Mia and Kathleen Botter. 2017. "Reference Rot in the Repository: A Case Study of Electronic Theses and Dissertations (ETDs) in an Academic Library." *Information Technology and Libraries* 36, no.1: 11-28.

Media Insight Project. 2015. "How Millennials Get News: Inside the Habits of America's First Digital Generation." http://www.mediainsight.org/Pages/how-millennials

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

-get-news-inside-the-habits-of-americas-first-digital-generation.aspx

Meikle, Graham and Sherman Young. 2011. *Media Convergence: Networked Digital Media in Everyday Life*. Basingstoke: Palgrave Macmillan.

Meyer, Eric T. and Ralph Schroeder. 2015. *Knowledge Machines: Digital Transformations of the Sciences and the Humanities*. Cambridge, MA: MIT Press.

Michetti, Giovanni. 2015. "Archives and the Web." In *Encyclopedia of Archival Science*, ed. Luciana Duranti and Patricia C. Franks. Lanham, MD: Rowman & Littlefield, 102-5.

Miligan, Ian. 2016. "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives." *International Journal of Humanities and Arts Computing* 10, no. 1: 78-94. doi: 10.3366/ijhac.2016.0161

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Nelson, Theodor Holm. 1965. "Complex Information Processing: A File Structure for the Complex, the Changing, and the Indeterminate." In *ACM '65 Proceedings of the 1965 20th National Conference Cleveland, Ohio, USA-August 24-26*. New York: ACM, 84-100. doi:10.1145/800197.806036

Nelson, Theodor Holm. 1993. *Literary Machines*. Sausalito, CA: Mindful Press.

O'Carroll, Aileen Sandra Collins, Damien Gallagher, Jimmy Tang, and Sharon Webb. 2013. *Caring for Digital Content, Mapping International Approaches*. Digital Repository of Ireland Series 2. Dublin: Digital Repository of Ireland. https://repository.dri.ie/catalog/5t356044x

OCLC. 2018. *Descriptive Metadata for Web Archiving*, eds. Jackie Dooley and Kate Bowers. OCLC Research Report. Vol. 1-3 2018. Vol. I: Jackie Dooley, Kate Bowers 2018. *Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Vol. 2: Jessica Ventlet, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell, Jackie Dooley. *Literature Review of User Needs*. Vol. 3: Jackie Dooley, Mary Samouelian. *Review of Harvesting Tools*. Dublin, OH: OCLC Research.

Paloque-Berges, Camille. 2017. "Usenet as Web Archive: Multi-Layered Archives of Computer-Mediated Communication." In *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 229-51.

Pennock, Maureen. 2013. *Web-Archiving*: *DPC Technology Watch Report 13-01 March 2013*. DPC Technology Watch Series. [London?]: The Digital Preservation Coalition. doi:10.7207/twr13-01

Plachouras. Vassilis, Florent Carpentier, Muhammad Faheem, Julien Masanès, Thomas Risse, Pierre Senellart, Patrick Siehndel, and Yannis Stavrakas. 2014. "ARCOMEM Crawling Architecture." *Future Internet* 6: 518-41.

Risse, Thomas, Elena Demidova, Stefan Dietze, Wim Peters, Nikolaos Papailiou, Katerina Doka, Yannis Stavrakas, Vassilis Plachouras, Pierre Senellart, Florent Carpentier, Amin Mantrach, Bogdan Cautis, Patrick Siehndel, and Dimitris Spiliotopoulos. 2014. "The ARCOMEM Architecture for Social- and Semantic-Driven Web Archiving." *Future Internet* 6: 688-716.

Saad, Myriam Ben, Stéphane Gançarski, and Zeynep Pehlivan. 2009. "A Novel Web Archiving Approach based on Visual Pages Analysis." Paper presented at IWAW 2009 International Web Archiving Workshop. https://core.ac.uk/download/pdf/38300970.pdf

Schafer, Valérie, Francesca Musiani, and Marguerite Borelli. 2016. "Negotiating the Web of the Past." *French Journal for Media Research* 6. http://frenchjournalformediaresearch.com/index.php?id=952

Schelin, Shannon and G. David Garson. 2004. "E-Government Adoption in the United States." In *The Internet Encyclopedia*, ed. Hossein Bidgoli. Hoboken, NJ: John Wiley & Sons., 1.

Schneider, Steven M. and Kirsten A. Foot. 2005. "Web Sphere Analysis. An Approach to studying online actions." In *Virtual Methods: Issues in Social Science Research on the Internet*, ed. Christine Hine. Oxford: Berg, 157-71.

Schneider, Steven M., Kirsten A. Foot, Michele Kimpton and Gina Jones. 2003. "Building Thematic Web Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive." In *Third Workshop on Web Archives Trondheim, Norway, August 21st, 2003 Proceedings,* ed. Julian Masanés, Andreas Rauber, and Gregory Robena, 77-94. http://bibnum.bnf.fr/ecdl/2003/proceedings.php?f=schneider

Schostag, Sabine and Eva Fønss-Jørgensen. 2012. "Web archiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective." *Microform & Digitization Review* 41: 110-20. doi:10.1515/mir-2012-0018

Seadle, Michael and Elke Greifender. 2007. "Defining a Digital Library." *Library Hi Tech* 25: 169-73. doi:10.1108/07378830710754938

Sim, Susan Elliott and Rosalva E. Gallardo-Valencia, eds. 2013. *Finding Source Code on the Web for Remix and Reuse*. New York, NY: Springer.

Steffen, Will, Wendy Broadgate, Lisa Deutsch, Owen Gaffney, and Cornelia Ludwig. 2015. "The Trajectory of the Anthropocene: The Great Acceleration." *Anthropocene Review* 2: 81-98. doi:10.1177/2053019614564785

Taylor, Nicholas. 2012. "Using Wayback Machine for Research" *The Signal* (blog), Oct. 28. http://blogs.loc.gov/digitalpreservation/2012/10/10950/

Theimer, Kate. 2015. "Digital Archives." In *Encyclopedia of Archival Science*, eds. Luciana Duranti and Patricia C. Franks. Lanham, MD: Rowman & Littlefield, 157-60.

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Ventlet, Jessica, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Literature Review of User Needs*. Dublin, OH: OCLC Research.

Webster, Peter. 2017. "Users, Technologies, Organizations: Towards Cultural History of World Wide Web Archiving." *Web 25: Histories from the First 25 Years of the World Wide Web,* ed. Niels Brügger. Digital Formations 112. New York: Peter Lang, 175-90.

Winters, Jane. 2017. "Breaking in to the Mainstream: Demonstrating the Value of Internet (and Web) Histories." *Internet Histories* 1, no 1-2: 173-9. doi:10.1080/24701475.2017.1305713

Zierau, Eld. 2011. "A Holistic Approach to Bit Preservation." PhD diss., University of Copenhagen. http://www.diku.dk/forskning/phd-studiet/phd/thesis_20111215.pdf

Appendix: Websites and portals

Archive.org. NSF Website, frontpage 7. October 1999. https://Web.archive.org/Web/19991007203722/http://www.dli2.nsf.gov:80/dlione/

Archive-IT: https://archive-it.org/

Archivethenet. (AtN): http://internetmemory.org/en/index.php/projects/at

Commoncrawl.org: http://commoncrawl.org/

The European Holocaust Research Infrastructure EHRI: https://www.ehri-project.eu/

The International Internet Preservation Consortium (IIPC): http://netpreserve.org/

Internet archive: https://Web.archive.org/

Internet Memory Foundation: http://internetmemory.org/en/

Internet Memory Research: https://internetmemory.net/en/

Library of Congress, Websites: https://www.loc.gov/Websites/

Library of Congress, Webpages: https://www.loc.gov/search/?fa=original-format:Web+page

Netarkivet.dk, English version see http://netarkivet.dk/in-english/

Networked Knowledge organization (NKO): http://nkos.slis.kent.edu/

Online Computer and Library Center. 2007. "Trusted Repositories Audit and Certification: Criteria and Checklist" 2007. Online Computer and Library Center and the Center for Research Libraries. www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Pandora Archive – National Web Archive of Australia: https://pandora.nla.gov.au/

UNESCO Charter on the Preservation of Digital Heritage. http://portal.unesco.org/en/ev.php-URL_ID=17721%26URL_DO=DO_PRINTPAGE%26URL_SECTION=201.html

UN Sustainable development goals: https://sustainabledevelopment.un.org/

Warc ISO standard 2017. https://www.iso.org/standard/68004.html

All Links verified December 2017 - April 2018.

https://doi.org/10.5771/0943-7444-2019-1-47 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY