Web archives

Finnemann, Niels Ole

Published in:
Knowledge Organization

Publication date:
2018

Document Version
Peer reviewed version

Citation for published version (APA):
Web archives and Knowledge organisation

To be published in Birger Hjørland (Ed.) ISKO Encyclopedia of Knowledge Organization & Knowledge Organization xx, no. yy.

Abstract.
This article deals with general web archives and the principles for selection of materials to be preserved. It opens with a brief overview of reasons why general web archives are needed. Section two and three present major, long termed web archive initiatives and discuss the purposes and possible values of web archives and asks how to meet unknown future needs, demands and concerns. Section four analyses three main principles in contemporary web archiving strategies, topic centric, domain centric and time-centric archiving strategies and section five discuss how to combine these to provide a broad and rich archive. Section six is concerned with inherent limitations and why web archives are always flawed. The last sections deal with the question how web archives may fit into the rapidly expanding, but fragmented landscape of digital repositories taking care of various parts of the exponentially growing amounts of still more heterogeneous data materials.

1. The web – a nearby perfect knowledge organisation system?
With the rapid spread of the Web protocols and the legalization of commercial internet activities in US early in the 1990's (Abbate 1999, 213-218 and Schelin and Garson 2004, 591) the internet was within a decade transformed form a specialised communication tool for scientists and students to a globally accessible, societal infrastructure to which other media, institutions and corporations had to accommodate. The web-protocols provided easy access to all sorts of information resources. They also opened for a third wave of digitization characterised by exponentially growing amounts of data, by new communicative genres and new formats of knowledge production, dissemination and organisation (Jenkins 2006, Meikle and Young, 2011, Kitchin 2014, Finnemann 2014, 2017b, Huurdeman et al. 2015).¹

Today the web has become the most comprehensive knowledge resource ever. This is the result of a vast amount of decisions taken by a huge variety of agencies all over the globe acting due to
each their own needs and goals. The aggregated result of these efforts has been the development of the peculiar hypertext architecture implemented in 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.

The affordances of this architecture is 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 and global reach.

Together this provides an extremely flexible tool, which can be adapted on the fly for knowledge organisation 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 nearby perfect system for knowledge organisation.

Three major obstacles however prevent the web from being a sufficient solution for KO in the 21th century.

First, it remains too large for any observer, including mechanical crawlers, to overview. 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 webpage was back then estimated to 44 days. Later calculations tell a similar history. 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
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.\(^4\) 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 in so far the materials are considered worth to preserve and to be accessible in the future.

2. 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 1990’s only a few years after the spread of the web 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 US took a ‘generalized philanthropic’ approach aiming to cover the whole web (Webster 2017: 181). The same year Kulturär3w 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 Kulturärw3 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 US where the non-profit Internet Archive aims to provide worldwide coverage and Library of Congress maintains a huge selective archive. In addition, a range of selective archives is established at major universities. There are only few web archives, if any in Near Middle East, Africa and South America. 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). 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) but they differ from other kinds of born digital materials, because archived web materials may include coded Internet links in the messages.

A list of web archiving initiatives (Gomes et al. 2012) can be found at Wikipedia. As off 31 May 2017 the list includes 79 initiatives. A range of the initiatives listed is also member of The International Internet Preservation Consortium (IIPC) established in 2003 that keeps an updated membership list.

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. Since 2006 the Internet Archive also provide 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.

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, but also reflecting the need to rethink the principles and criteria for archiving these materials, as former principles of archives and libraries are not sufficient. There are multiple reasons for this ranging from the global and fast expanding character of the web, the point-to-point architecture allowing everybody to moderate a domain, and not least the hypertext-based network architecture allowing editable coded links and interference between the machines and messages in the network.

Web archives at the same time enters yet unclear relations to each other as well as to other sorts of digital repositories, research infrastructures, and a variety of curated digital heritage institutions.

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.

While the notion ‘archive’ usually refer 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 the live web. Accordingly, national libraries rather than national archives play a major role in web archiving.

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. The delimitation of 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 Library of Congress in US was allowed to keep a full twitter archive, but in 2017 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. 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 of user behaviour including references to sources outside Facebook. In the case of Facebook, Twitter and similar services that is driven by large corporations or even monopolies it might be worth to consider 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.

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. Web archives - are they worth it?
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 if 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. 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. On aspect however need 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.\textsuperscript{18} The question why web archives will always be flawed due to intrinsic characteristics of web materials will be further elaborated in section 6.

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.\textsuperscript{19} The appreciation of web archives is also still lacking, as they are not used that much except for consultation of individual webpages. According to Meyer and Schroeder (2015, 191-192) web archives are in risk to end up as ‘dusty archives’ because scientists prefer to use the live web in spite of the missing materials, which are out weighted 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. Still 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 still more unique source, sometimes even the only source available for a growing range of historical studies.\textsuperscript{20} 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 public in general.
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 21st 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. 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 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 Blue Ribbon Task Force Report on sustainable preservation of digital materials. The Report consider long-term preservation of digital materials as a ‘societal challenge on a par with climate change and sustainable energy’ (Blue Ribbon, 81) and focus on digital ‘materials that are of long-term public interest’ (Blue Ribbon, 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.’ (Blue Ribbon, 1).

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 (Blue Ribbon, 2).

The four domains with each their own economical preservation profile does 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 taken more care of 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 strategies, and to provide clarification about long-term value and selection criteria’ (Blue Ribbon, 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

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, 49). The notion of disruption is 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 allowing new usages, which imply some sort institutional transformation as researchers seldom make such materials accessible for others. 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 odd with well-established principles of selection due to claimed value and quality, which fits well to printed media-materials. Such materials remain stable when published and they allow curating due to value and quality to take place later and eventually utilizing public assessments of value and quality. It is also at odd with the use of acknowledged content providers (e.g. publishing houses or media corporations) as proxies guaranteeing the quality due to the overwhelming amount of digital content- and service providers and the transnational reach. 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 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, in so far 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. 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).23

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 interests 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. 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 have 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.24

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 organisation is transformed into an ever-growing bricolage of web materials harvested and archived due to a variety of criteria.

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. Topic centred archives – eventually supported by focused crawlers – may be targeting any particular theme 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’. 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 less the same as the criteria for non-digital archives, libraries and collections, but the conditions for collection are quite different.

Topic centric archiving includes the harvesting of materials related to a particular domain, understood as an area of knowledge. This is quite different from to the notion of a ‘web domain’, understood as a particular set of web-addresses and which constitutes ‘domain centric’ harvesting (Masanès, 2006: 41-43).

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, as quality-based archiving is no longer sufficient due to the huge amounts of and the ephemeral character of web materials.

Even if the purpose is clear and well defined the question is still 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 add the question how a given set of materials are embedded in changing archive histories.26

**Domain centric Selection.** A second set of criteria for which there is no non-digital equivalent relate 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. 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 doesn’t 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 follows the links from an initial site (or a seed list of initial sites) to the pages on the subdomain levels. The automated procedures are of course also much cheaper than selection based on intellectual resources.27

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 the 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.28

Answers to the question of what to preserve are highly dependent of 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 the users – still – are free to use sites on most domains.
The amounts and ephemeral character of web materials call for the use of mechanised and automatized archiving methods favouring also 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. 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’ of a specified set of web addresses) takes place without preceding validation. Whether it is worth to preserve all these materials of questionable quality is of course a highly controversial issue and the discussions are still on-going. 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.

Such values can be identified on six dimensions: First, 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. Second, when
stored in web archives these materials form a unique type of source materials for studies in many areas not feasible without these materials. Third, 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. Fourth, 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. Fifth, 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. Sixth, and finally, 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 use of representative samples (e.g. Halevy et al. 2013).32 Broad strategies neither exclude deletion or augmentation of materials in the future.

Time Centric Selection. A third set of criteria for selection relate to the complexity of the variety of time scales, which are 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: First, Web archiving history is at least to some extent rooted in the fear of or the experience of the sudden disappearance of websites overnight.33 This explains well 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 web sites are not only short lived; they are often also interactive and include scripts eventually embedded in dynamic link instructions. Links and hypertext relations are not simply connections (as a reference system or footnote system) they also always include an instruction of what to do at a specified destination. These dynamic instructions – often, but falsely ignored as integral part of hypertext relations34 – 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 build into the site, possibly triggered by a visitor, or build into another site from which the materials are accessed or the action is triggered during a page-request. (Masanès 2006, 13-17, Taylor 2012, Brügger and Finnemann 2013).

This facility is increasingly used in contemporary network-based knowledge organization systems or multiple source knowledge systems. An example is Knorr Cetina’s (2009) analysis of a software system in which 6-8 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. Knorr 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.35

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, and represent a fast-developing new kind of knowledge organisation, which however often requires their own archiving strategies because of their use of multiple timescales and updating frequencies.

Since editable timescales can be inserted in between any two elements they constitute a reservoir for development of new genres while the editable timescales 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 add 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. Since web materials are always restored or ‘replayed’ (Taylor 2012) from a server when called for, the call may generate transformations and cannot take into account former appearances of the materials. Thus web archives are composites of a variety of time horizons: The time horizon 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 timescale 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 closing of some of the link relations and time scales related to the surrounding web and thus characterised by a set of closures build in to the archived materials.

5. 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 use a broad range of harvesting strategies, including harvesting on the level of National Domain, Regional Domain, Bulk, Selective, Event, and Thematic. Way back in 1996 they used bulk harvesting collecting simply as much as possible.
The Swedish project, Kulturarw3 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).³⁹ Contrary to these and more in accordance with established library traditions a topic centered project, the PANDORA project 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 (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’ (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 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.⁴⁰

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, but also added a new focus on the dynamic and time sensitive character of the materials.⁴¹

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 (many sites are updated, and new pages and links created, moved or broken several times in between two snapshots). 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 early 21st 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.42

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.

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. The web however is still both an important source in itself and also serves as a very significant index to valuable content, social relations, genres and activities on the web and in society at large, including also non-web digital platforms.

In spite of the fast-growing range of archiving projects experimenting with a variety of archiving strategies build 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.43 Thus, we cannot tell whether the
materials preserved are those worth to be preserved. There is also lack of criteria for deciding, which materials should be considered worth to preserve.

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.

6. Web archives are always flawed
The growing array of archiving strategies cannot hide the fact that web archives are always flawed. First of all they are flawed as a result of the criteria used for selection of materials. Something will always be missing. Second, 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. Third, archived materials are flawed due to the variety of editable timescales, which can be ascribed to any part of any message. This is not least the case for materials, which include real time data. Fourth, they may be flawed due to the modifications of web materials either before the archiving process as links may be broken, materials deleted or moved to another address taking place in the timespan between the collections of different parts of the materials. Fourth, 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. Finally, they are of course also flawed due to physical problems resulting in informational noise.
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 materials by cutting off link connections and thus 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 dramatically reduced in so far the materials are archived and fenced off from the ever oscillating live web, if not in real time then with a minimal delay.

7. Alternatives or supplements
The establishing of general, often national Web archives is not the only method for preservation and organisation of the knowledge resources on the web. The development of the web has been accompanied by the development of a variety of strategies aiming to optimize and preserve the use of digital materials as a knowledge resource. The overarching issues relate to the constitutional role of hypertext, which in many ways turns upside down the original ideas of computational processes.

The notion hypertext was originally coined by the philosopher Ted Holm Nelson and conceptualised 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, Nelson 1993). With the idea of a global, interlinked ‘docuverse’ he seemingly took the classical ideal of knowledge organisation 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. But 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 et al. 2001, 3; Berners-Lee et al. 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 created and 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 build 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.47

A fourth strategy is to rely only on 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 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 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. General web archives 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 organisation 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 organisation sources as they develop in the future. Since most general archives are national archives there is a need for facilitate interoperability in between these and the internet archive and other on-going archiving initiatives.
Web archives are multiple source knowledge organization systems. The principles of general web archives are seldom discussed from the perspective of Knowledge Organisation. There are reasons for this. First, if Knowledge Organisation limits itself to deal with validated scientific and scholarly resources, web archives stretch far beyond this domain. Such archives include 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 organized for this purpose. Each archive is built according to a specific set of principles for selection, preservation, presentation (knowledge visualization) and search facilities provided by the particular archive. These principles represent a particular type of knowledge organisation, which organize source materials for a huge variety of possible research projects. 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 KOS’s. 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 cannot be appropriately described within the vocabularies of previously developed KO’s. 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, 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 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. This is the case even if such processes are performed via long chains of automated and responsive sequences as in “self-driving” cars.
If, as argued for instance by Hjørland, (2017, 12) KOS’s should be understood as independent of the specific media, the question is whether and how the characteristics of different media enter into the conceptualisation of KO and in particular how to include the features made possible only by networked digital media.

Web archives differ in many ways from former types of archives. First, because the materials have to be archived on the fly because of their ephemeral character. Second, web archive materials are collected by help of mechanical methods, which (among other things) rely on knowledge of the precise location of the materials to be archived while the archiving process at the same time cut off links to the surrounding web. Third, the overwhelming amounts imply that only a tiny part of the materials can be considered as validated resources. Fourth, because the materials are networked and complex. Fifth, because the editor position remains open in the lifecycles of digital materials as there is no physical closure, but only coded closure of any single element and sequence in every single document.

On a par with all types of digital materials they also differ from printed materials because digital materials are invisible and have to be made visible - or otherwise perceptible - via an interface, which always represent a particular, selective perspective or interpretation of the sources. Digital media finally also differ from print media because the material manifestations remain editable and can be utilised for semiotic variation.

General web archives will always include materials from multiple sources and the materials may have passed through a series of transformations and remix processes. Like other multiple source knowledge systems web archives may be characterized due to their particular hypertext configurations: of the intrinsic timescales; of the ‘spatial’ scale local—national—global in respect to domain addresses; in respect to content; ns in respect to the multiplicity of sources each possibly with its own updating frequency and so forth.

Web archives includes broken links to the surround web, while digitized archives and other types of digital born archives even if the materials may be circulated on the web are already fenced off in so far hypertext is only used for navigation in the materials or internal feature only in a delimited set of materials.
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 organisation Ibekwe-SanJuan and Bowker (2017) argue that Big Data create a need to rethink the standpoint from which the KOS’s are designed. As indicated in the title, the source of the requirements to rethink the principles of KOS’s 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.51

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 conceptualisation 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 2017a) – 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 in so far 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 Birger Hjørland’s 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 conceptualisation of the data materials reaching 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 globalised system of all sorts of KOS’s. 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 KOS’s – 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, general perspectives. There should be no single paradigm for KO. Rather they should stretch from clearly specified and closed KO’s 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 organisation 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 21th century exponentially growing amounts of digital materials are immersed in a globalized hypertext landscape – and there is a need for further analysing the implications for the development of KOS’s, not simply the multiple source and partly real time-based systems but the whole array of new formats for the range of possible KOS’s.
If Knowledge organisations 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 KOS’s.

List of references:


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 (2), Article 2. [http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1085&context=yjolt](http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1085&context=yjolt)


Websites and portals:

- Archive-IT: https://archive-it.org/
- 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/
- Networked Knowledge Organisation (NKO): http://nkos.slis.kent.edu/
- UN Sustainable development goals: https://sustainabledevelopment.un.org/
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 2014).

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 (math, images, diagrams, many different spoken and typed languages etc. For Limitations of big Data analyses see e.g. boyd and Crawford 2011, Moretti 2013, Kitchin 2014, Gatto 2014, Ibekwe-SanJuSan and Bowker 2017. 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 Lyman et al. 2003; Mannerheim 2000; Masanes 2006; 2; Hilbert & Lopez. 2012; Brügger 2018, 55.

Costa, Gomes and Silva 2017: 191. 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.

Masanes (2005, 72-74) identifies changes in ‘authorship form’, ‘content shaping’, ‘convergence’ and ‘technique’ as four major factors making web archiving more complex than archiving manuscripts and printed documents.

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.

Cornet, Gomes and Silva 2017:198-199.

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.’

See Kirschbaum, 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 (Kirschbaum et al p. 33).


http://netpreserve.org/about-us/members/

The Common Crawl Foundation is a non-profit organization founded in 2007. Commoncrawl.org holds xxx amounts (source) The 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/


Robot.txt is a de facto standard based on consensus within the www developer community in the early 1990’es 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 foundations’ quoted from https://www.iso.org/standard/55211.html


A list of countries with/without legal deposit laws for web archives can be found at The International Internet Preservation Consortium (IIPC) web pages: 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 Denmark see Schostag and Fønss-Jørgensen, 2012. The Internet Memory Foundation performed a survey on web archiving in 2013, 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

Quote from Internet archive, front page, https://web.archive.org/
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.

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 organisations and facilitated their opening towards broader audiences.

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).

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 6.

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.

Among a growing range of strategies for statistical analyses of large cultural datasets see for instance Halevy, Nordvig, and Fernando Pereira, 2009 Christakis and Fowler, 2009; Moretti, 2013; Aiden & 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 sampling 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.

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


For an analysis of such intricacies in Google’s Usenet-web archive see Paloque-Berges 2017, 229-251.

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 both domain centric harvesting of national domains and other domains.

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.

In Finnemann 2017a 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 are intrinsic in the web archive materials (cf. Gail Hodge 2000). See also the NKO homepage at http://nkos.slis.kent.edu/.

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 to topic centric in depth harvesting of full sites. In other cases more levels are included to ensure that more sites a harvested in 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.

For an overview of combinations of archiving strategies in general web archives see Gomes et al. 2012; and the website of IIPC (The International Internet Preservation Consortium).

See for instance Halevy et al. 2013.


Hypertext is often described as a mechanical connection between an input location (the anchor) and a destination. This is basically a ‘go to’ command included in most programming languages leaving the next step to the user. The process however may include a coded instruction of what to do at the destination or even at several destinations. The link instruction can include instructions for any sort of action to be performed at the specified destinations. This is what makes hypertext something else than a ‘go to’ command, or a foot note reference technique. All elements anchor, location, coded instruction, specified destination, and action to be performed at the destination are editable. Cf. Finnemann, 2017b.

This and other examples of multiple source systems based on networked digital media are discussed in N.O. Finnemann, 2017b.

The relation between the ephemeral and persistent character of web materials and the implications for web archiving is discussed in Steven Schneider and Kirsten Foot, 2005 and Masanés, 2006. Masanés, 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.’

Taylor 2012
37 As reported at the IIPC member site: http://netpreserve.org/about-us/members/internet-archive/
38 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 never exist at the same time on the web.
40 The Danish case is documented in Birte Christensen-Dalsgaard et al 2003. The report includes (p. 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 in which it was also stated that the archived materials should be considered cultural heritage. See also Schostag and Fønss-Jørgensen 2012; Finnemann, 2001; Brügger, 2001. For a detailed discussion of how to delimit a ‘national’ web domain see Niels Brügger 2017c.
44 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.
45 For analyses of the lack of archivability see e.g. Zierau 2011, Kelly et al. 2013.
46 For a more elaborate discussion see Brügger 2017a. For archiving of websites using source code, see Anne Helmond 2017.
47 The knowledge Organisation of web archives also include the collections principles and strategies as well as the visualization facilities and the organisation 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.
48 According to Costa, Gomes and Silva 2017:193 the size alone makes traditional methods for cataloguing ‘too time consuming and expensive’ and they argue for automatic indexing. They do not address the ephemerality of the materials, nor the deformations inherited in the process of harvesting.
49 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 (boyd and Crawford 2011, Kitchin 2014).