Hypertext Configurations: Genres in Networked Digital Media

Niels Ole Finnemann

The Royal School of Library and Information Science, University of Copenhagen, DK 2300 Copenhagen, Denmark. E-mail: finnemann@hum.ku.dk

The article presents a conceptual framework for distinguishing different sorts of heterogeneous digital materials. The hypothesis is that a wide range of heterogeneous data resources can be characterized and classified due to their particular configurations of hypertext features such as scripts, links, interactive processes, and time scalings, and that the hypertext configuration is a major but not sole source of the messiness of big data. The notion of hypertext will be revalidated, placed at the center of the interpretation of networked digital media, and used in the analysis of the fast-growing amounts of heterogeneous digital collections, assemblages, and corpora. The introduction summarizes the wider background of a fast-changing data landscape.

In the early 21st century the processes of the digitization of culture and society have entered a new phase. Today, digital media are used everywhere. We scan the entire world from outer space to the interior of the body. A growing number of people participate in the daily production of digital materials, intentionally as well as unintentionally. Recent studies estimate that the production of digital materials in the last few years surpasses the total amount of information produced in the previous history of mankind (Dienes, 2012; Hilbert, 2012; Hilbert & López, 2012; Kitchin, 2014b). The exponential growth, however, is not simply "more of the same." The data of today are more diversified due to ever more different purposes articulated in distinct software paradigms, resulting in a growing diversity of knowledge formats. With this comes the so-called "messiness" of the fast-exploding production of "big data" that has now become an issue of a more fundamental character, giving rise to methodological debates on how to deal with increasingly "unstructured" and "data intensive" materials (Aiden & Michel, 2013; Boyd & Crawford, 2012; Frické, 2015; Gitelman, 2013; Hey, Tansley, & Trolle, 2009).

A further implication is that research itself increasingly produces messy data sets by compiling research corpora derived from multiple sources. This may be in the analysis of real-time processes of long-term developments, of many sorts of human activities that are performed using different digital media platforms, or as part of efforts to deal with "grand challenge" issues that demand inter- and cross-disciplinary analyses, as in climate research (Borgmann, 2015; Burrows & Savage (2014); Kitchin, 2014a; Steffen et al., 2015).

Today, we have neither developed an overview of the existing variety of data materials, nor do we have general criteria for what to preserve. Insofar as society at large wants to keep some of these materials, whether due to cultural heritage perspectives, future research, or future commercial value, a number of fundamental questions need to be raised regarding the criteria for selection and how to describe, analyze, and visualize the variety of materials considered worth preserving. Such questions are pertinent to the work of libraries, archives, museums, and existing and new research infrastructures (RI), as well as public institutions and corporations, including small and medium enterprises (SME), which otherwise may not be able to deal with their own data. Furthermore, these questions touch on core issues in all academic disciplines, and not least in the library and information sciences. Fundamental questions are thus raised about how to describe these materials and how to organize the knowledge resources of the 21st century.

The aim of this article is to contribute to the development of a conceptual framework for characterizing digital materials.

The hypothesis is that a wide range of heterogeneous data resources can be characterized and classified due to their particular configurations of hypertext features, such as scripts, links, interactive processes, and time scaling, and that the hypertext configuration is a major but not sole source of messiness. The notion of hypertext will thus be placed at the center of the interpretation of networked digital media, and used in the analysis of heterogeneous digital collections, assemblages, and corpora.

The point of departure is taken from the broader notion of networked digital media (Finnemann, 2014a) that are presented as based on notions of hypertext, interactivity, and

search and pattern recognition, which, even if they can be traced far back in the history of digital media, are given new dimensions—not least because of the editable hypertext time scales.

In the second section the timescales implied in hypertext relations are analyzed on the basis of three examples: the global currency trading system based on Knorr Cetina's work (2009, 2014), Facebook, and web materials. The three cases are chosen as examples of different configurations of multiple source knowledge systems (in the following referred to as MSKS) that are formed around ongoing hypertext interaction between multiple anchor points and destinations.

The third section discusses whether hypertext configurations can be considered as genres and suggests six parameters for characterizing different genres of MSKS. It is proposed that there is a need for developing the study of digital materials into a field in its own right on a par with the study of other types of mediated materials (Finnemann, 2011). The paper concludes with some remarks on the wider perspectives and further questions to be addressed.

The Hypertext Nature of Networked Digital Media

Although a dominant buzzword in the late 1980s, the term "hypertext" was used less often during the 1990s. In those latter years, one of the most significant instantiations—web protocols—paved the way for a multidimensional expansion of hypertext relations. This is the case with respect to reach (from local to global, private to public, and whom to whom) and of modes (intrinsic within conceptual wholes—works—and extrinsic in between conceptual wholes). As will be discussed, the World Wide Web (WWW) protocols also stretch the multiple timescale(s) of hypertext relations.

Barnet (2013) describes the conceptual history of hypertext from Bush's (1945) idea of Memex and Nelson's (1965) coining and further elaborations of the term "Hypertext" to Story Space created by Joyce (1987), Bolter (1991), and others in the late 1980s. While this is the history of the emergence of a conceptual framework, centered on stand-alone machines or a single user and a "one-system" perspective, there is a growing need to conceptualize the much more widespread and mundane uses of hypertext in networked digital media.

A New Paradigm

Hypertext and interactivity is presented in Finnemann (2014a, 2014b) as a basis for a third paradigm in the history of digital media formed around networked digital media, searches, and pattern recognition. The two former paradigms are the "computer classic" idea of a rule-based data processing machine—derived from Alan Turing's concept of a universal computing device—and the human computer interaction paradigm focusing on the interface and the notion of the computer as a tool or tool box (Bannon & Pylyshyn, 1989).

The notion of networked digital media can be traced back to Turing's (1936) idea of a choice machine, which differed from the idea of the computer as a rule-based, automatized machine (Finnemann, 1999a) even if the ideas were interpretations of the very same machine. That the choice of a first step is always a human choice about how to start a digital device—even a robot—may sound trivial, within the context of networked digital media the choice of the next step allows the reading position to be turned upside down because readers may continue reading, leaving the material unaltered, or they may modify the material either at the present location or at some other destination by means of an interactive intervention. Thus, the working mode of the author, the editor, the curator, the navigator, and the censor are available as options in the reading position.

This is so because the Internet allows ongoing hypertext interactions between multiple anchor points and destinations, which brings the "choice machine" perspective to the fore and reduces the automated programming perspective to that of a special case—albeit still of significant innovative relevance. With the focus on the choice of the next step, the fundamental interconnectedness of hypertext, interactivity, and search also becomes immediately clear: Interactive relationships are hypertextual in character, as are search processes.

Hypertext Time Relationships as a Basis for Networked Digital Media

Hypertext can be seen from the author's perspective, from the perspective of the "work" (e.g., the storage on the server), and from the perspective of the "reader." These perspectives can be seen at the level of processing and at the levels of functionality, of features, and of meaning. Cronin (2001) addresses the issue of multiple authorship of research results and foresees that "In the future, it is quite likely that the concept of the 'author-function,' ..., will vary from one 'epistemic community' ... to the next" (p. 567). The focus is on the acknowledgment of contributions to large-scale collaborative research projects that produce a concerted work, such as a film. Clearly, hypertext facilitates the development of multiple authorship forms, but worse, it also makes it more difficult to keep track of the production process. This is because networked digital media allow each participant to switch between the author/editor mode, the navigator mode, and the reader mode, and to incorporate multiple sources both in real time and in edited timescales.

In the following, the main focus will be on the reader position, and the point of departure will be the distinctions of *author/editing* mode, *navigation* mode, and *reading mode*, which together constitute *modal shifts* as a constituent of hypertext reading (Askehave & Nielsen, 2005; Finnemann, 1999b; Kong, 2015; Mitsikopoulou, 2015; Zhang & O'Halloran, 2012). Each of the modes includes a spectrum: The *editing* mode includes the spectrum author-editor-curator, the *navigation* mode includes the spectrum navigation-browsing, and the *reading mode* includes the

spectrum close-reading, skimming (as overviewing) and scanning (of particular indexical markers). Today, these modes are also supplemented by automatized routines such as "distant reading" (Moretti, 2013). Thus, the hypertext reading position includes all text reading modes and adds navigation modes and author/editing modes. Finally, the author-editing mode also includes the spectrum of changing the functional architecture, a particular feature, and/or the content.

The reason for privileging the reader position is that this is the initial position of any use of a corpus of digital materials and when it comes to the analyses of hypertext time dimensions it is often not possible to either establish previous relationships (of production and remix) or a future state (of the "work"), as this state (unlike the case for print) depends on the action of the reader. The reading time establishes a distinction between the past and the future of the corpus because it opens up possible changes to the material and, thus, the continual ongoing development of material over time.

The very same relationships constitute a significant source of noise. A long time span between the definition of the anchor and reading at the destination will increase the noise. This is often also the case for real-time relationships between multiple sources.

Example 1: The Global Currency Trading System

To describe multiple source real-time hypertextual exchanges, Knorr Cetina (2009, 2014) has suggested the term "synthetic situation" based on "scopic coordination." They are defined as the more or less instantaneous coordinating activities made possible by distributed, but coordinated, screens. She does not use the notion hypertext, and she wanted the argument to be neutral regarding particular media, but her most interesting examples are instantiations of hypertext relationships in digital media. A prime example is the global currency trading system, established with the help of the Bloomberg and Reuters's services. Dealers all over the world use these services, with a few market-maker centers in, for example, London, New York, and Tokyo. The organization is a multilayered system of networked digital media, which in this case are synchronized, and integrate all sorts of financial news for each of the hundreds of dealers. The information is presented on six to eight screens filled with hundreds of individually updated cells and charts. The real-time processing of constantly updated information and trade actions facilitates the intensified, seemingly spontaneous, communication of the dealers "as if a traders' *brain* was attached to the market … unthinkingly" (Knorr Cetina, 2014, p. 52). The reading mode is described as enforced by flow-like subconscious activity. It still includes a modal shift related to the positions of the dealer switching between observing/scanning (reading modes) and enacting (author/editing modes), as the operations of any online dealer are fed into the system at the same time as inputs are generated by preprogrammed algorithms. It also includes facilities that

allow private dialog between the dealers. The ongoing updating of the various sources of information reflects global market fluctuations and constitutes a system that requires that the dealers respond to the changing stream within less than a second. Thus, the system provides a strongly delimited "window of interaction" demanding what is denoted "response presence" (as distinct from "embodied presence").

We may call this "big data" but, more precisely, we could speak of a new type of global knowledge system based on the use of a particular hypertext configuration to integrate information from many different sources as a kind of organized and selective crowdsourcing. Whether, and how, this kind of "typified rhetorical action" (Miller, 1984, p. 159) qualifies as a new genre will be discussed in the forthcoming section.

The sources incorporated in the system have different timestamps but the crucial time dimension is the "reading time," or the live copresentation of any relevant kind of information in different locations. The screen setup and the interface structures are possibly the only constant elements throughout the day. The information is constantly updated from a variety of sources each with their own rhythms. This can be considered paradigmatic of a class of systems based on a response presence that allows the agents less than a second—far below ordinary reading speed.

Formerly, response presence was limited to face-to face contact and to telephone communication; Today we have a growing range of—increasingly global—systems, which are also based on huge and deliberately composed combinations of background information that is constantly fed into the present situation. We have MSKS monitoring weather, climate, and traffic, be it flights, trains, ships, or cars. These may vary with respect to domain and with respect to interactive intervention forms, with or without an enforced type of response presence. Pilots flying aircraft may often use automated procedures but they also have a system for complex monitoring, integrating external information with the information from their own operations.

Response presence may be more or less strongly enforced by the system. Many computer games are based on similar response-presence demands. If the gamers do not respond within 1, 3, 5, or 10 seconds they may lose the game, miss a new level, or lose their most valuable earnings. The reading mode can be deliberately modified by programmers but is more or less intensified by the interplay between gamers online, as global market activity defines time in the case dealers, discussed earlier.

Synchronized, global MSKS that depend on the configurations of hypertext represent a significant aspect of globalization today. However, they are only one among a wider range of MSKS that are based on distinct hypertext configurations.

Among the global MSKS, a growing number are based on response presence within hours, a day, or a few days. This is the case for various forms of social media sites (Instagram, Facebook, and Snapchat), which must be read

within a few hours if you are not going to miss the window of interaction. Crowdsourcing platforms normally leave the response time to be defined in the particular project. Insofar as a news site provides a window of interaction, it is usually to be visited on a daily, "real-time" basis. Multiple source news sites (such as gdelt.com, Google news, Newsvine) mostly provide a window of interaction for modifying the settings of selected news types. Response presence is not always enforced. It may be an issue of negotiation, as in the case of texting and e-mail. Real-time systems include a variety of response presence forms, some defined by the stream (risk of missing information and response option), some by activities (risk of loss of money), and some by negotiation.

Example 2: Facebook

Facebook provides a mixture of enforcement and negotiation. The response presence is more loosely governed and may be stretched for some hours without losing too much of the info-stream. It also depends on different Facebook practices within the overlapping individual networks of "friends" and friends' friends. Facebook acts as a coeditor, and also collates series of messages from a variety of sources with some sort of simultaneity ranging from immediate interactions to delayed interferences, although most often within a day.

In the first example the relation between dealers and inputs were symmetric. Facebook is asymmetric, based on globally aggregated inputs that are "locally" distributed to many, partly overlapping, small networks. At the same time, Facebook crowdsources and privatizes all the traces left. These traces consist not simply of messages and their content, but also of fluctuating network relationships between the messages generated among the subscribers by their friending and grouping activities, such as likes and sharing, privacy settings, and so on.

Facebook also delivers organizational memory structures such as a cumulative timelines, an activity list, a list of friends, a personal archiving facility, and related features. So far, there is no augmentation of the data delivered by the subscribers beyond its manipulated distribution among individually created networks, and Facebook's soft editing practices that may include nudging, sequencing, and possibly prioritizing messages according to more or less unknown and changing criteria—and, of course, relevance-based advertisements. Strict censorship is also practiced. Augmentation takes place primarily on the hidden side of the information delivered by subscribers.

Twitter delivers a simpler and open system of networking via hypertext tags, leaving the connection between followers and followed to individual choices and restricting each message to a limited number of characters. The subscriber is able to select a particular configuration of in-stream messages as a kind of selected crowdsourcing. Even if the Twitter format is simple, the usages are extremely complex, as shown by Tufecki (2014), who presents a series of obstacles for big data analysis of Twitter. These complexities relate, not least, to subtle utilizations of hypertext syntax, for instance, as in subtweets, shorthand, and the hiding of denominators. The dominant forms of social media corpora are semi-open corpora that allow a relatively high degree of subscriber influence on the functional network architecture, and not simply on delivering content.

These essential elements of intensive hypertext systems are difficult to capture, preserve, and document. Are they of relevance? It seems they are. During its first 10 years Facebook has developed into a significant platform for negotiations and discussions on all sorts of political, social, and cultural affairs. This is not primarily due to the merit of Facebook, but rather a result of decisions taken each and every day by a large proportion of the world's population. The question, therefore, is how these human networks, mediated by social media platforms, can be sufficiently documented for study, and how this can be combined with study of the actual content that people place on the agenda.

Facebook is a website, or rather a collection of distributed sites, including apps that stretch outside the web. It is a legal entity, which does not produce content but maintains certain editorial principles, primarily in the form of moral censorship. Over the years it has become increasingly integrated into the wider public and interferes with legacy media, primarily by means of links to external sources shared by the subscribers. This is possible only because Facebook is embedded in the wider web universe that can also be characterized via the hypertext configuration. This, however, will require a distinction between open and closed hypertext corpora.

Example 3: Web Materials, Websites, and Archived Web Materials

Brügger and Finnemann (2013) argue that the distinction between digitized and digital-born materials is based on two different criteria for the use of hypertext. Hypertext relationships cannot be an intrinsic part of digitized materials (even if used when dealing with such materials), whereas there will always be such relationships in digital-born materials.

This distinction is also valid for the interface; interfaces of digitized materials are extrinsic, they are not part of the nondigital original and they are often defined a posteriori. For digital-born materials, an interface is always included as an intrinsic part of the materials. In both cases the interface specifies a repertoire of search facilities, navigation, editing, and reading modes, and in both cases this interface can be supplemented or modified.

Since we often want to delimit a digital corpus (even if digital-born) to a fixed form, we need a concept for deliberately established hypertext-free chunks of digital materials, whether or not they are copies of nondigital originals. In the following, such chunks are denoted as closed corpora. In some cases there are hypertext relationships within such a closed corpus insofar as they are "frozen" as internal and noneditable after the day of publication. Hypertext relationships will still be needed to deal with such materials, but

they are kept separate from the frozen hypertext relationships within the corpus and excluded as intrinsic features.

Any digital corpus, whether digitized or digital-born, has a history prior to the actual reading of the corpus, and this history will include a distinction between links considered to be external—useful for dealing with the materials—and links that are an intrinsic part of the materials.

Website and web universe. A website may be defined by a domain address or by an editorial regime that stretches across several URL domains, but in both cases a website constitutes a delimited space composed differently from the web as a whole. The composition always includes specifications concerning interactive facilities including response-presence demands, often related to updating policies. A website is not a closed corpus in itself, but it may be defined as closed if there is no further updating. If it is not closed it is still delimited by the distinction between internal and external links. The time dimension may, in some cases, also imply subtle issues as to whether a later updating is intrinsic or extrinsic. If this is undertaken by the editorial regime that created the corpus it may be seen as intrinsic (and the content of the corpus is time-sensitive), whereas an update is extrinsic if a curator undertakes it. In this case an unaltered version can still be identified.

The web universe as a whole is a different kind of corpus, which exists as such only because of hypertext interconnections. For this corpus the distinction between internal and external links is irrelevant, but it also differs because it is fluctuating. Even if we may want to consider the overall global archive, we could never capture the component parts consisting of individual sites as a whole at any given "same" moment, nor could we relate to the particular content of all websites in such a moment. Thus, web materials in general constitute one kind of corpus, whereas a website constitutes a different kind.

Even if we cannot embrace the web universe as a whole, it is still useful to consider its characteristics as a corpus because the online web link relationships allow for very complex aggregations of data from many different sources that can be brought together by an individual request. Such aggregations can only be archived "manually" and "on-the-fly" within the short lifetime of the materials, if at all. We cannot restore a Google search of yesterday.

Online web materials and archived web materials. As a further consequence of this, the hypertext configuration of any sort of web archive is also constituted differently from the universe of the web itself. The web universe is centered on the presence of materials at the time of reading. This presence is established by means of hypertext connections. To archive any fraction of the web, be it a single page, a site, or a number of sites, you need to carve it out by cutting the links to the surrounding web. These broken links delimit the archived corpus as distinct from the online "equivalent." This is also the case for national or general web archives. Thus, web archive studies will always deal with the sorts of

messiness related to broken links, scripts that do not work in the archive, and messy interactive exchanges. The anchor-destination relationship, inherited in web archives, differs from the anchor-destination relationships in live web materials. It is also well known (Brügger, 2011) that general archives will contain multiple instances of the same website, while the web will often only contain the most recent instance. The archive thus allows for long data studies that cannot be performed on the live web. This is also an option for many other sorts of long-term collections of materials.

While the web protocols widened the reach of hyperlinks to a global scale, they did not provide a similar extension on the time axis. The materials of yesterday (or the addresses of these materials) are very often either modified or deleted (Ashenfelder, 2011). This is precisely why efforts have been made to preserve web materials in general, either through the Internet Archive, archive.org, or outside the US in general web archives established around national domains, cooperating in the International Internet Preservation Consortium (IIPC). The web itself provides hypertext relationships on a short-term timescale with global reach, whereas the archives provide a more scattered universe as they are subsets carved from the live web.

To sum up: The web universe as a whole is constituted of links. Websites are constituted by the distinction between internal and external links. Web archives are established, delimited, and distorted by broken links. Web archives also add a secondary dimension, as the archived materials are not archived simultaneously but over time. Thus, some links establish connections that have not existed on the live web insofar as the content has been modified or removed in between the archiving of the anchor and the destination (Brügger & Finnemann, 2013).

Hypertext Configurations as Genres

The examples discussed so far show that crucial differences between digital corpora are constituted as different configurations of hypertext, and that the complex multiple time dimensions inherent in all hypertext systems are particularly relevant for identifying such differences.

The question is whether these configurations should be considered as new genres or as additional features added to the repertoire of functionalities in existing genres. In contemporary functionalist genre theory (Askehave & Nielsen, 2005; Askehave & Swales, 2001; Giltrow & Stein, 2009; Swales, 2009), genres are described as constituted by purpose, functionalities, and narrative. Yates, Orlikowski, and Rennecker (1997), who are concerned with organizational practices, place particular focus on "the socially recognized purpose" and "common characteristics of form" (p. 1). These are functional theories but they are in accordance with cultural historical activity theory (Andersen, 2015; Bawarshi & Reiff, 2010; Bazerman, Bonini, & Figueiredo, 2009) insofar as both traditions have shifted away from identifying the genre by looking at the similarities "between documents" to examine social action seen as "typified rhetorical

actions based in recurrent situations" (Russel & Fischer, 2009, p. 163 [see also Miller, 1984]).

In the literature, hypertext systems have been described as a more or less new genre emerging out of existing genres, breaking various limitations that were formerly considered to be constitutive of textual genres (Barnet, 2013; Bolter, 1991; Bolter & Grusin, 1999; Giltrow & Stein, 2009). Thus, linearity is often replaced with nonlinearity or multilinearity (Landow, 1992). In these theories hypertext—in a variety of instantiations—is seen as an additional functionality that can be implemented in a closed corpus or a work.

Askehave and Nielsen (2005) propose that the web constitutes a basis for a variety of hypertext features that should be taken into account in the genre analysis of web materials. Following Finnemann (1999b), they state that hypertext from a reader's perspective is constituted by modal shifts between reading modes and navigation modes. Although from their perspective the reading mode may be analyzed as usual, they claim there is a need to perform a distinct analysis of the navigation mode. The argument is demonstrated in an analysis of front pages, which they describe as a new web-based genre adding the navigation mode both to the level of purpose, the repertoire of functional features (link/moves), and to the rhetorical strategy. Thus, they conclude that hypertext genres need to include the medium as a distinct level due to the navigational repertoire:

> An account of the generic properties of genres on the web involves a three-level analysis of both modes: In the reading mode, the text must be characterised in terms of its communicative purpose, moves, and rhetorical strategies; and in the navigating mode, the medium must be characterised in terms of its communicative purpose, links, and rhetorical strategies (Askehave & Nielsen, 2005, pp. 127–128).

Since their case is built on front pages, they are mainly concerned with the menu-link structures used to navigate the site, which they consider to be a medium-specific feature external to the text, and they explicitly leave out the third mode, that of editing. If the editing mode is included it becomes clear that the text itself is also editable and, therefore, should be included as part of the hypertext form of digital materials. Thus, in digital media hypertext is not an addition to text, but the landscape in which the text is immersed.

It may be argued that text-based genres and practices are inscribed into hypertext-based digital media where they can be treated in much the same way as they were treated before. If so, they will have to be treated with the help of hypertext features and interfaces. However, they can also be dealt with in new ways, as shown, for instance, in the examples mentioned earlier.

The MSKS described in this article are all strongly dependent on the social action of the users and they pose a series of interactive structures for these actions. One may even argue that it blurs the distinction between a document and a "typified rhetorical action," as these actions are constantly fed into a document that is, thereby, modified.

It is too early to tell how the genres created around analog legacy media will evolve either outside digital media or inside, for example, e-books. It is not too early, though, to conclude, as documented in the preceding sections, that there is a fast-growing array of "new types of typified social actions" that are genuinely built on the hypertextual basis of networked digital media. Many of these genres emerge out of previous known genres, as also formerly shown in the history of film (drawing on literature and theatre) from theories of path dependency (David, 1985) and from the hypertext literature (Bolter & Grusin, 1999). In new genres there will always be some traces of previous forms of human communication, but the range of functional features available in digital media reaches well beyond the conditions given by print, radio, film, television, and organizational practices formed around nondigital types of documents.

As a consequence, a theory of digital media genres needs to include both a theory of genres in which hypertextual properties are dealt with as external to the purpose and, thus, are not essential to the functions, and a theory in which these properties are exploited in one way or another as integral to the purpose, and also as essential for purpose, functionality, and narrative. PDF-files, e-books, you tube videos (not Youtube as a whole) can be considered as part of the first category, whereas the systems dealt with in the previous sections can be considered as part of the second.

Six Dimensions of Internet-Based Multiple Source Knowledge Systems

In the examples discussed in this article one can identify six major dimensions of MSKS that together serve as a delimitation of MSKS from other genres and as variables that are incorporated in different ways in different kind of typified social practice that utilize MSKS.

The six characteristics are:

- Time sensitivity
- Local-global sources
- Public–private
- Whom-to-whom
- Editability/interactivity
- Messiness.

The list is not exclusive. Further studies may show that other characteristics are equally important. The claim is that these six characteristics can be used to identify different types of typified social action involving MSKS.

Time sensitivity. This dimension includes MSKS that are based on response presence, ranging from milliseconds (the currency system) to hours (Facebook), a day or a few days (news sites), and MSKS without any built-in time sensitivity (be they cumulative or rarely updated). The backward time sensitivity—can you go back to former states—varies between more or less full-scale backward backup and no backup at all. Only some sources provide backup facilities

(Wikipedia, for instance, has a log of changes; most blog servers also provide permanent backup) while others do not. Thus, MSKS in most cases are based on messy data with respect to former states and the provenance of the present state. Insofar as they deserve to be preserved, they pose new questions as to how to preserve and how to document the messiness. The web as a whole cannot be backed up and former states can only be partly restored with the help of web-archiving strategies, which, however, are also flawed due to the fluctuating character and variety of updating practices as well as the widespread use of interactive and dynamic features on the live web. Thus, hypertext is both the constituent of the WWW, and of a growing variety of web-based genres, and a main source of the messiness of web data as well as many other sorts of data. For delimited fractions of web materials it is possible, to some extent, to archive at least fractions of the materials that may serve as "ruins" and remain to be studied. The future time sensitivity of MSKS depends on both the present and future practices, which in most cases, but not always, will be formed around the in-built editorial facilities.

Local-global sources. The currency trading system and Facebook represent two different types of social practice with respect to utilization of the seamless variable web-scale reaching from local to global. In the case of the currency system, both the multiple sources and the agencies operate within a global reach dimension. In the case of Facebook, the platform is global, whereas the subscriber networks, that provide the content and the communicative network relations, are "local." They are not local in the traditional sense of being bound to a local place, the neighborhood. They are local in the sense that the overlapping networks of "friends" and "groups" are established on the basis of mutually agreed individual choices. They are built on some sort of affiliation that is often based on belonging to some sort of "local" community. Thus, there is a gulf between the platform and the variety of communicative practices performed on this platform, as well as between these practices and the communicative practices of the same people on other platforms, whether on legacy media, on social media, or on other parts of the web. According to some studies, Facebook today is a main news source for younger generations in the US (Pew Research Center, June 2015). For this reason, the issue of whether Facebook should be considered an emerging MSKS in itself or, alternatively, as a kind of intermediating system, should be left to further analysis. There are increasing gulfs between particular software platforms and the social functions based on the changing usages of multiple platforms.

Public–private. Since the global currency system is constituted as a market, it is, by default, open to anyone, but because it is also transformed into one overarching MSKS based on costly resources, the market functions more like a professionalized, semiprivatized gray zone.

In the case of Facebook, there are also gray zones that are partly left to the interpretation and delimitation of the individual subscribers. The web as a whole is a free public service presented by the inventor as a gift to mankind and, in principle, open to any citizen. At the same time, the service allows for any kind of delimitation and establishing of private spaces, whether as closed spaces or as commercial spaces. A service such as Google stands out as an example of a system available to everybody but for individually defined needs, limited only by Google's personalization strategies that tends to fence people in according to their own search history. Thus, Google is primarily a publicly available private-service MSKS but, as is well known, the privacy is heavily limited by the collection and commercialization of the traces left by usages. Facebook—and the Internet in general—is built around a both–and mixture on the three scales of local–global, public–private, and whom-to whom.

Whom-to-whom. The currency system today is a hierarchical system with the MSKS described by Knorr Cetina (2009, 2014) at the center, aggregating all sorts of information and bids from dealers wherever they are located. On one dimension it is a system of everybody communicating with everybody else. On the more limited operating dimension, it is an asymmetric system of many information sources communicated to a few operating dealers who are also communicating with each other in a symmetrical few-to-few system. Facebook facilitates a much more complex configuration of overlapping networks that evolve over time both with respect to subscribers and subscribers' friendships relations, and with respect to themes and subjects of communication, group formation, and relations to legacy media, marketing, and business activities. The web as a whole is an everybody-to-everybody system, limited by language, culture, variety of needs, shared ideas or interests, and so on. A particular website may establish its own delimitations.

Editability. The editable character of digital materials, which is an inherent part of these materials due to the fundamental hypertextual nature of digital media, is probably the most distinctive of the features discussed because it incorporates the author and editor modes as options in the reader position.

The editing mode includes the adding of new content, the modification or deletion of existing content, and the modification of the functional architecture of the system. The role of interfaces to digital media is not simply to make the materials visible and accessible, it is also to specify a limited set of options for editorial modifications, be it of content, of particular features, or of that part of the functional architecture that is provided as software. Thus, any system specifies a—limited—range of editorial facilities.

In the currency system we may suppose that the fields of information collected on the screens have to be edited from time to time, but when the system is running it is open only for change to the values within each field, whether the source is stock market information, news, or input from the dealers in action. In the case of Facebook we have a

different system in which the subscribers provide their messages (texts, images, videos, links) in a range of typified text fields, but also add to the development of networked communities, formation of groups, and ranking patterns via "like" functions. Again, one might illustrate the distinctiveness by including Google Search, which is enacted by the instruction of a user but can only be modified by changing the search parameters. Many Google services actually qualify as MSKS, using the characteristics of MSKS in different ways.

Messiness. The messiness of data has always been an issue, once referred to in the saying "garbage in garbage out." More significant, perhaps, is that modern information theory, as established by Claude Shannon, had as its main issue the question of how—automatically—to correct for the noise produced by the channel (Shannon & Weawer, 1969 [1949]). Until recently, messiness has mainly been seen as an issue to be dealt with by adding coded redundancy or by cleaning up data before they are treated for analytical purposes. With networked digital media and, not least, "big data," the messiness of data has now become an issue of a more fundamental character.

Schlobach & Knoblock (2012) identify a range of sources for this messiness, explaining that they pose a number of new questions, as compared to the previous semantic web, which have succeeded in producing data in "structured, linked, and standardized ways," but are now confronted with data that:

> ... comes with a plethora of contradicting world-views and contains incomplete, inconsistent, incorrect, fast-changing, and opinionated information. This information not only comes from academic sources and trustworthy institutions, but also is often community built, scraped or translated. In short, the data is messy and difficult to use (p. 1).

As indicated in their brief summary, there are many sources of the messiness and it may take many different forms, an issue that is far beyond the scope of this article. However, one of the main and recurrent sources is to be found in hypertext because this is a constituent feature of digital media. The authors' mention of new, untrustworthy sources, including all sorts of civic groups, may also be an indication that digitization has entered a new era, one in which messiness is an inescapable condition, and that the analysis of the messiness should be included in the future analysis of heterogeneous digital materials.

In the case of Facebook, we are confronted with a third source of messiness, since it is not possible to establish a complete corpus either for any single user or for any network of subscribers. One is left to extract prespecified selective data (if you are allowed) or to deal with abstract network and distribution patterns (Christakis & Fowler, 2009) that have to ignore the meaning of the messages, such as the meaning of a "like." In the case of web materials, a particular type of messiness appears in the dilemma between the impossibility of studying the whole web, except meagre measurements, and the fact that any delimitation is provided by breaking links that disturb the contextual dimension and may imply that the excavated corpus is incomplete insofar as the content also depends on external sources.

Further examples include corpora created via crowdsourcing procedures that are characterized as procedures for aggregating information or knowledge from multiple sources—a crowd that is usually established via a public call (Estellés-Arolas & González Ladrón-de-Guevara, 2012). Crowdsourcing is not necessarily related to networked digital media, but a range of new forms have developed, exploiting the hypertext characteristics of these media. For general search engines, "the crowd" is constituted by the array of interlinked websites, their results fluctuate throughout the day, and each specific result is only available on a particular, "personalized" machine. In other cases, such as the online encyclopedia, Wikipedia, the crowd is everybody, and the results are public and cumulative over time. The history of the corpus is recorded and can be traced. In other cases, again, the result is private and, ultimately, not constituted as a corpus at all.

Analysis of the examples also shows how the ongoing processes of digitalization challenge existing archiving strategies and raise new issues around establishing documentation, preservation, and institutionalization, not least for synchronized global knowledge systems.

These examples are not considered to be exhaustive; on the contrary, one may argue that the questions raised call for the establishment of the analysis of digital materials as a new field in its own right. The analysis of hypertext configurations seems to be fruitful for characterizing a large set of digital materials but should be supplemented with analyses of the agencies, purposes, data formats, and tools involved.

Conclusion

It has been demonstrated that it is possible to distinguish between different types of heterogeneous digital materials due to the configurations of multiple, hypertext timescales, not least when observed from the reading position, which introduces a separation of the past and future of any corpus due to the window of interaction that both allows and limits the possible switches between reading, navigating, and editing modes.

Most, if not all, digital corpora and collections can be grouped on a scale ranging from corpora to be read in real time to corpora to be read asynchronously, at any later time. This includes single source systems. The notion of "real time" turns out to be a variable notion, however, involving rather different ways of reading. For instance, the reading mode can be strongly controlled by the data stream. The currency trading system is the most extreme case, both because response presence is limited to less than a second and because it is enforced both by streaming and by the interactions of the globally distributed dealers.

The notion of asynchronous reading also has to be modified, as the reading of a corpus at different times is not necessarily a reading of the same, unaltered corpus. Even if the corpus is unaltered between two readings, new metadata and annotations may have been added or removed. Google Search and Google Translate are examples of global multiple source sites to be read or used at any time, but that are sensitive to the particular reading time. Cumulative systems (Google Books, some cultural heritage systems) are examples of a less complicated dependency on reading time.

Hypertext systems are inherently unstable, and the instability increases due to an increase of sources incorporated in the configuration. Thus, the question is raised as to how they should be documented, what can be preserved, and who should be responsible for the preservation. As Facebook is increasingly used as a forum for public debates, intermingling with Twitter, Instagram, mass media, and so forth, it also raises issues of how to archive such materials in ways that actually allow us to study these distributed public fora of relevance for writing or for otherwise visualizing the history of the 21st century. Instability, however, is not always and only related to the hypertext nature of digital media. It may also be the result of the character of the data and the methods used to generate the data. Thus, there is also a need to reconsider the epistemologies of coordinated multiple source hypertext systems and establish the study of networked digital media in a field of its own.

References

Aiden, E., & Michel, J.B. (2013). Uncharted. Big data as a lens on human culture. New York: Riverhead Books.

Andersen, J. (Ed.). (2015). Genre theory in information studies. Bingley, UK: Emerald Group Publishing.

Ashenfelder, M. (2011). The average lifespan of a webpage [Web log post]. Retrieved from: http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/

Askehave, I., & Nielsen, A.E. (2005). Digital genres: A challenge to traditional genre theory. Information Technology & People, 18(2), 120–141.

Askehave, I., & Swales, J.M. (2001). Genre identification and communicative purpose: A problem and a possible solution. Applied Linguistics, 22(2), 195–212.

Bannon, L., & Pylyshyn, Z. (1989). Perspectives on the computer revolution. Wilmington, NC: Intellect Books.

Barnet, B. (2013). Memory machines. London: Anthem Press.

Bawarshi, A.S., & Reiff, J. (2010). Genre: An introduction to history, theory, research, and pedagogy. Anderson, SC: Parlor Press.

Bazerman, C., Bonini, A., & Figueiredo, D. (Eds.) (2009). Genre in a changing world. Perspectives on writing. Fort Collins, CO: WAC Clearinghouse and Parlor Press. Retrieved from: http://wac.colostate.edu/books/genre/

Bolter, J.D. (1991). Writing space: The computer, hypertext, and the history of writing. Hillsdale, NJ: Erlbaum.

Bolter, J.D., & Grusin, R. (1999). Remediation: Understanding new media. Cambridge, MA: MIT Press.

Borgmann, C. (2015). Big data, little data, no data. Scholarship in the networked world. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, Communication and Society, 15(5), 662–679.

Brügger, N. (2011). Web archiving—Between past, present, and future. In M. Consalvo & C. Ess. (Eds.), The handbook of Internet studies (pp. 24–42). Oxford, UK: Wiley-Blackwell.

Brügger, N., & Finnemann, N.O. (2013). The web and digital humanities: Theoretical and methodological concerns. Journal of Broadcasting and Electronic Media, 57(1), 6680. doi:10.1080/08838151.2012.761699

Burrows, R., & Savage, M. (2014). After the crisis. Big data and the methodological challenges of empirical sociology. Big Data & Society, April-June, 2014: 1–6. doi:10.1177/2053951714540280

Bush, V. (1945). As we may think. The Atlantic, July 1945. Retrieved from: www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Christakis, N.A., & Fowler, J.H. (2009). The surprising power of our social networks and how they shape our lives. New York: Little, Brown & Company.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication. Journal of the American Society for Information Science and Technology, 52(7), 558–569.

David, P. (1985). Clio and the Economics of QWERTY. The American Economic Review, Vol. 75, No. 2, Papers and Proceedings of the Ninety-Seventh Annual Meeting of the American Economic Association. (May, 1985), pp. 332–337.

Dienes, I. (2012). A meta study of 26 "how much information" studies: Sine qua nons and solutions. International Journal of Communication, 6, 874–906.

Estellés-Arolas, E., & González Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. Journal of Information Science, 38(2), 189–200.

Finnemann, N.O. (1999a). Modernity modernised. In P.A. Mayer (Ed.), Computer media and communication—A reader (pp. 141–160). Oxford, UK: Oxford University Press.

Finnemann, N.O. (1999b). Hypertext and the representational capacities of the binary alphabet. Århus, Denmark: Center for Kulturforskning, Aarhus Universitet. Retrieved from www.hum.au.dk/ckulturf/pages/publications/nof/hypertext.htm

Finnemann, N.O. (2011). Mediatization theory and digital media. Communications, 36(1), 67–89.

Finnemann, N.O. (2014a). Digital humanities and networked digital media. MedieKultur, 30(57), 94–114

Finnemann, N.O. (2014b). Research libraries and the internet: On the transformative dynamic between institutions and digital media. Journal of Documentation, 70(2), 202–220.

Frické, M. (2015). Big data and its epistemology. Journal of the Association for Information Science and Technology, 66(4), 651–661.

Giltrow, J., & Stein, D. (2009). Genres in the Internet. Amsterdam: John Benjamins Publishing.

Gitelman, L. (Ed.). (2013). Raw data is an oxymoron. Cambridge, MA: MIT Press.

Hey, T., Tansley, S., & Trolle, K. (Eds.). (2009). The fourth paradigm. Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Hilbert, M. (2012). Big data for development: From Information- to Knowledge Societies. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145&rec=1&srcabs=2229952&alg=1&pos=3

Hilbert, M., & López, P. (2012). How to measure the world's technological capacity to communicate, store and compute information? Part I–II: Results and scope. International Journal of Communication, 6, 936–955, 956–979.

Joyce, M. (1987). Afternoon: A story. Storyspace software. Watertown, MA: Eastgate

Kitchin, R. (2014a). Big data, new epistemologies and paradigm shifts. Big Data & Society, 1, 1–12.

Kitchin, R. (2014b). The data revolution. London: Sage.

Knorr Cetina, K.K. (2009). The synthetic situation: Interactionism for a global world. Symbolic Interaction, 32(1), 61–87.

Knorr Cetina, K.K. (2014). Scopic media and global coordination: The mediatization of face-to-face encounters. In K. Lundby (Ed.), Mediatization of communication (pp. 39–62). Berlin: Mouton de Gruyter

Kong, L. (2015). A multimodal study of bank websites (unpublished doctoral dissertation). The Hong Kong Polytechnic University, Hong Kong.

Landow, G.P. (1992). Hypertext: The convergence of contemporary critical theory and technology. Baltimore, MD: Johns Hopkins University Press.

Miller, C. (1984). Genre as social action. Quarterly Journal of Speech, 70, 151–167.

Mitsikopoulou, B. (2015). Rethinking online education: Media, ideologies, and identities. London: Routledge, Taylor and Francis.

Moretti, F. (2013). Distant reading. London: Verso.

Nelson, T.H. (1965). Complex information processing: a file structure for the complex, the changing and the indeterminate. ACM '65 Proceedings of the 1965 20th national conference. pp. 84-100. doi 10.1145/800197.806036

Pew Research Center. (June 2015). *Millennials & Political News*. Retrieved from: http://www.journalism.org/files/2015/06/Millennials-and-News-FINAL-7-27-15.pdf

Russel, D.R., & Fisher, D. (2009). Online multimedia case studies for professional education. Revisioning concepts of genre recognition. In J. Giltrow & D. Stein (Eds.), Genres in the Internet (pp. 163–191). Amsterdam: John Benjamins Publishing.

Schlobach, S., & Knoblock, C.A. (Eds.). (2012). Editorial: Dealing with the messiness of the web of data. Journal of Web Semantics, 14, 1–2.

Shannon, C.E., & Weawer, W. (1969 [1949]). The mathematical theory of communication. Urbana, IL: University of Illinois Press.

Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M., & Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. Science, 13 Feb 2015: Vol. 347, Issue 6223. Retrieved from http://science.sciencemag.org/content/347/6223/1259855. doi 10.1126/science.1259855

Swales J.M. (2009). Worlds of genre – Metaphors of genre. In C. Bazerman, A. Bonini, & D. Figueiredo (Eds.), Genre in a changing world (pp. 147–157). Fort Collins, CO: WAC Clearinghouse and Parlor Press.

Tufecki, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. ICWSM 2014 (pp. 505–514). Ann Arbor, MI, June 1–4, 2014.

Turing, A.M. (1936). On computable numbers — with an application to the Entscheidungs problem. Proceedings of the London Mathematical Society, (1936-37), 2(42), 230–265 (with corrections in Vol. 43, 1937, 544–546).

Yates, J., Orlikowski, W.J., & Rennecker, J. (1997). Collaborative genres for collaboration. Genre systems in digital media. Proceedings of the 30th Hawai International Conference on System Sciences. 1060-3425/97 IEEE computer society (ISSN 0-8186-7862-3/97).

Zhang, Y., & O'halloran, K.L. (2012). The gate of the gateway: A hypermodal approach to university homepages. Semiotica, 190, 87–109.