Cultural heritage as digital noise: nineteenth century newspapers in the digital archive

Johan Jarlbrink and Pelle Snickars
Department of Culture and Media Studies, Umeå University, Umeå, Sweden

Received 5 September 2016
Revised 2 March 2017
Accepted 19 March 2017

Abstract

Purpose – The purpose of this paper is to explore and analyze the digitized newspaper collection at the National Library of Sweden, focusing on cultural heritage as digital noise. In what specific ways are newspapers transformed in the digitization process? If the digitized document is not the same as the source document – is it still a historical record, or is it transformed into something else?

Design/methodology/approach – The authors have analyzed the XML files from *Aftonbladet* 1830 to 1862. The most frequent newspaper words not matching a high-quality references corpus were selected to zoom in on the noisiest part of the paper. The variety of the interpretations generated by optical character recognition (OCR) was examined, as well as texts generated by auto-segmentation. The authors have made a limited ethnographic study of the digitization process.

Findings – The research shows that the digital collection of *Aftonbladet* contains extreme amounts of noise: millions of misinterpreted words generated by OCR, and millions of texts re-edited by the auto-segmentation tool. How the tools work is mostly unknown to the staff involved in the digitization process? Sticking to any idea of a provenance chain is hence impossible, since many steps have been outsourced to unknown factors affecting the source document.

Originality/value – The detail examination of digitally transformed newspapers is valuable to scholars depending on newspaper databases in their research. The paper also highlights the fact that libraries outsourcing digitization processes run the risk of losing control over the quality of their collections.

Keywords Sweden, Archives, Documents, Accuracy, Print media, Auto-segmentation, Character recognition equipment, Large-scale digitization

Paper type Research paper

Introduction

In October 1847, the telegraphic wire in St Germain outside of Paris was struck by lightning. The Swedish newspaper *Aftonbladet* reported that a telegraph assistant at a station nearby had discovered the demolished telegraph printing several letters on its own. Yet, according to the paper, "since they were not coherent, he decided to signal the phrase used for 'I do not understand.'" In doing so, however, he received "a heavy electric shock, which was followed by a loud bang, sounding like a gunshot" (*Aftonbladet*, 1847).

Within a digitization project initiated by the National Library of Sweden, the 1847 October copy of *Aftonbladet* was digitized in 2013 at the Swedish Media Conversion Centre. The newspaper *Aftonbladet*, founded in 1830, was one of the key titles in nineteenth century Sweden. It is often described as the first modern newspaper – consequently, it was also the first newspaper to be completely digitized by the National Library. Then again, if a telegraph struck by lightning in the late 1840s produced some real uncanny results, the same can be said of present day digitization processes. The digital version of the paper with the lightning-telegraph incident, in fact, literally reported that the struck assistant "saw a dazzling light along the wires on the walls conducting electricity de visI devärdigavid värdigavid dejemte fullkomen ihåförvintparkerslagna förvintparkerslagna parkerslagna ken – tas till70 70 misvårt fruktarsnart tAf eoch sisrans njes ej […] which fell down in pieces, burning the table and the floor."

This research has been funded by The Torsten Söderberg Foundation.

Journal of Documentation
Vol. 73 No. 6, 2017
pp. 1228-1243

0022-0418
DOI 10.1108/JD-09-2016-0106

The mysterious words in the middle of the quote are not Swedish, and no reader of *Aftonbladet* in 1847 would have found them in any newspaper copy (hence no reference to the quote). As the non-coherent letters printed by the telegraph they seem to have been generated by an external disturbance – which, however, occurred 166 years later through the very act of digitization at the Swedish Media Conversion Center. Today, these "sentences" can be found in the newspaper database, "Svenska dagstidningar" at the National Library. Many texts and words now part of similar digitized newspaper databases share the same fate; and some are of a similar weird kind. What was never printed in old newspapers has today become part of the historical record.

In this paper we argue that the digitization of historical newspapers is not a neutral process where data are transferred from one medium to another. On the contrary, when newspapers are digitized they are transformed. Like telegraphic signals they usually resemble what was transmitted, but sometimes not. In this paper, we are consequently interested in noisy media, and the ways that digitized nineteenth century Swedish newspapers can today be perceived as a sort of waste(d) heritage. Likewise, there exists a fascinating media historical analogy since basically all of the printed newspaper issues reporting on the failing St Germain telegraph in 1847 were also turned into waste (or waste paper) after a few days. And the very few copies that survived are now slowly disintegrating in library repositories.

As is well-known, libraries all over the world are today digitizing their historical newspapers for preservation, as well as making them digitally available for research (and pleasure). The results of these efforts are useful databases that make it possible to search millions of newspaper pages online. Yet, as we argue in this paper, contemporary digitization processes can also be seen as a continuation of the process turning newspapers into waste. As the digitized newspaper report from 1847 displays, digitization can generate its own sort of waste – usually in the form of digital noise. In addition, digitization today results in physical paper copies being wasted, almost as soon as the pages have been scanned. In most cases preservation of cultural heritage is the opposite of destruction (Assmann, 2008). But for newspaper digitization, preservation and destruction goes hand in hand.

In this paper – prompted by a media archeological interest to dig deeper – we ask ourselves in what ways are newspaper documents being transformed in the digitization process? What kinds of errors do actually occur? If the digitized document is not the same as the source document – is it still a historical record, or is it transformed into something else? How is it possible to practice source criticism when the mechanisms and algorithms for selecting, capturing, processing and storing the historical data are hidden behind graphical user interfaces? In short, the process of digitization, optical character recognition (OCR), article segmentation, modes of presentation etc are all infrastructural settings that transforms old newspapers into new objects with a media specificity different from the original paper prints. As a consequence, the growing reliance on digital reproductions of old newspapers raises questions – from both a heritage and research perspective – regarding the function of such scanned documents, especially the relation between newspaper source documents and digital reproductions (Mussell, 2012).

Media technologies are seen as user friendly as long as one does not have to bother about the way the underlying technologies work. The archive, as Jussi Parikka (2012) makes clear, could be seen as such a media technology, since it "is the implicit starting point for so much historical research that it itself, as a place and a media form, has been neglected, become almost invisible" (p. 113). It can thus be argued that digital archives are more invisible than traditional archives, since the mechanisms regulating them are virtually hidden behind a graphical user interface. This is obviously in conflict with historical methodologies emphasizing on source criticism, and questions regarding the selection and processing of sources. The way digitized data (as newspaper files) are created, stored, processed and

formatted naturally have implications on how historical records as digital newspapers can be accessed and used, the histories that can be explored and the stories that can be (re)told. "If history is a matter of what is stored, and if what is stored is a matter of the media available (stone, papyrus, DNA, bone, film, floppy disks)," John Durham Peters (2015a) has stated, "then changes in the infrastructure will mean changes in the historical record" (p. 83).

The problem is "that while we think we are searching newspapers," it has been argued, "we are actually searching markedly inaccurate representations of text" (Hitchcock, 2013, p. 14). Such inaccurate representations, however, are also interesting as a point of departure, and as an analytical and empirical object of study. Hence, in the present paper, noisy media is of interest to us. Our paper departs from – and describes – the ways in which nineteenth century newspapers have been digitized at the National Library of Sweden, with a particular focus on both the resulting (and erroneous) XML files (from *Aftonbladet* 1830 to 1862), as well as the institutional setting at the Media Conversion Center, where the actual digitization activities takes place. We are, in short, interested in the informative capacity of digital reproductions, as well as the practices around the digitization process, and the ways that frequent distortions and errors presently shape what is regarded as heritage.

Media transformations – in theory and practice

Among scholars researching digitized cultural heritage it is well known that so called "digital surrogates" do not, and cannot, duplicate or replace original paper documents (Conway, 2015; Dahlström, 2009; Manoff, 2006). Digitization is a way to provide access – and not a preservation strategy in the traditional sense (preserving the integrity of the physical object). However, within a heritage culture permeated with digitization activities that are nowadays customary, one might argue that the notion of "surrogates" is increasingly contested. Online digitized newspapers can, indeed, skew historical research (Milligan, 2013). Yet, they are steadily becoming the norm for research (rather than studying original prints) – which makes it particularly important for heritage institutions to keep an eye on actual output.

Even if digitization activities have been geared toward facilitating access, especially when it comes to fragile newspapers, internal motivation for digitization (at heritage institutions) has usually been to preserve what will otherwise turn into paper flakes and dust. In a policy document defining the digitization strategy of the National Library of Sweden it is, for example, stated that "a fundamental part of the library's mission is to preserve the cultural heritage for the future. Preservation is therefore one of the most important selection criteria. We prioritize the digitization of material on low-quality paper, mainly newspapers, magazines, printed adverts, and audiovisual material where the recording media are deteriorating" (Kungliga biblioteket, 2016). When digitization is put forward as a preservation strategy it is often interpreted as a specific form of "information capture;" the preservation of material "on" paper – not the paper itself. From a heritage institution perspective, the transfer of information – texts, images, sound or video – from one medium to another, is hence usually what digital preservation has come to signify. At least in practice.

One of us authors (Pelle Snickars) has worked a decade within the Swedish heritage sector (and a number of years at the National Library), and can indeed assert that the notion of access is akin to institutional lip service; a kind of insincere support but not really put into practice (nor backed by funding). In essence, the main reason why issues around access have been ignored has to do with copyright issues – in the case of digitized old Swedish newspapers, the limit for freely available issues are now 1901 (but prior the restriction was set at the year 1864). In short, since legislation has made it difficult for (especially European) heritage institutions to handle copyright, preservation has usually become the guiding principle of digitization activities. As the policy document from the National Library of Sweden affirms, preservation is almost embedded in the DNA of heritage institutions (as well as among professional archivists

and librarians) – and arguably one of the reasons why the National Library has paid so little attention to actual output of its noisy newspaper heritage.

Nevertheless, there are also good reasons to question the distinction between digitizing for access and digitizing for preservation. Paul Conway has, for example, argued that the split "is artificial and misleading." Following his line of reasoning, in the digital world access is always default and the "obvious outcome of digital transformation, even if access is fully realized only through functioning electronic networks and the legal frameworks that manage permissions." In essence, access is a "given in the digitization process" (Conway, 2015, pp. 54-55). We argue that somewhat paradoxically, this is exactly why preservation has become the measure of the value of digitization efforts (in the case of newspaper digitization activities at the National Library of Sweden). Through institutional heritage processes of what has been described as "archivalization" (Ketelaar, 2001) access has constantly been overlooked and even neglected – since it is a given – with collateral results.

Most heritage institutions are aware of the transformations of content that the digitization process generates. Yet, as we will show in this paper, practical matters tend to forsake theoretical insights. As previous research on digitized newspapers has shown, whenever information is digitized and transferred, new noisy "information" will be added with inaccurate OCR as the most observed and common noise category. Evaluations of OCR accuracy have been carried out in several countries with a focus on national digitization projects, among them Australia, the UK, Finland as well as the digital heritage portal, Europeana (Holley, 2009; Tanner *et al.*, 2009; Kettunen *et al.*, 2014; Pletschacher *et al.*, 2015). In most evaluations a small text sample has been selected, manually keyed or corrected, and compared with the result produced by the OCR engine. In general, inaccurate OCR research and reported result have varied between 50 and 70 percent accuracy on a word level, and between 71 and 98 percent accuracy on a character level (Kettunen *et al.*, 2014; Holley, 2009). In yet another study, an average word accuracy of 78 percent was reported, including a significant word accuracy (150 common stop words excluded) of 68 percent, and a 63 percent accuracy for words starting with a capital letter (proper nouns) (Tanner *et al.*, 2009). Distinctions between words, significant words and proper nouns are highly relevant when the usefulness of a database for key word search is evaluated. Pletschacher *et al.* (2015) calculated the accuracy with several user types in mind. For a phrase search to be successful individual words had to be accurate, but also the word order. For the word order to be correct text regions must be sufficiently segmented (consecutive regions not split, nonconsecutive regions not merged). The success rate for a phrase search, taking text segmentation into account, was calculated to be 62 percent.

It is important to stress, that the model user directing these evaluations, searched for individual words or phrases using a database interface. Digitized newspapers can be explored in many other ways, however. The model user in our analysis, for example, represents a researcher not using the search interface at all. Researchers within the growing field of digital humanities often use digital methods for massive text analysis and they need access to the complete text files. The user who is close reading individual texts can overlook the noise as long as they can find what they are searching for. Researchers doing digital text analysis are not reading individual texts – they want to find linguistic patterns in a large text corpus. Accurate OCR is essential, and so is correct text segmentation. Newspapers are seldom digitized with this user group in mind, but the databases are potentially rich resources for digital text analysis – if the noise levels can be managed.

Arguably, the problems with digitization projects are by now well known, at least within the scholarly archival science community. But as this paper shows, theory and practice are indeed different things. Why otherwise would the resulting OCRs of newspaper pages from the National Library of Sweden be packed with errors? On the one hand, we suspect that the library has been unaware of the tremendous amounts of noise they have added in the

digitization process. On the other hand, we also suspect that since the National Library primarily perceives digitization as a preservationist activity, access has simply turned secondary. In other words, the institution is aware of problems caused by digitization, but still delivers noisy files – predominantly for the reason that the deployed software simply cannot (at present) do a better job. Either way, the erroneous results are surprising, not the least since large-scale digitization projects have been on the political heritage agenda in Sweden for a number of years. The establishment of Digisam – a co-ordination secretariat for digitization, digital preservation and digital access to cultural heritage – was, for example, part of the Digital Agenda for Sweden (initiated by the Government in 2011). Work performed at Digisam have followed a vision where "cultural heritage is digitized, accessible and usable for everyone [and where] there is a coordinated, cost-effective infrastructure to support digitization, use and preservation of high quality" (Digsam, 2014).

Exploring noise

Our paper explores the digitized newspaper collection produced by the National Library of Sweden from a perspective where we media archeologically concentrate on and accentuate heritage as noise. To uncover what is hidden and taken for granted during the process of digitizing newspapers, we are using media theoretical perspectives focusing on media as technical infrastructures, rather than content and representations (Krämer, 2015; Peters, 2015b). Following this tradition the functions of technical infrastructures are revealed in primarily two ways: media technologies are examined when they are still under construction and before they are presented to end users, or approached when they either break down or generate noise.

Mapping noise, however, is a tricky endeavor. Misinterpretations generated by OCR follow a pattern to some degree, but they are also random enough to make systematic noise mapping difficult. Digitized newspapers can look very different from day to day. One issue seems perfectly alright, but the next day's issue is corrupted on every single line. Some OCR misinterpretations are more frequent than others, but the variations are innumerous. Our aim was hence to explore noise – rather than strictly measure it. Digital methods have been used to identify frequent errors, as well as corrupted versions of single words.

The digitized paper analyzed here is *Aftonbladet*. Since the digital result depends, to some extent, on the character of the printed newspaper, a short description of the layout could be useful: In the early 1830s, the paper was printed in three columns in a quarto format. Over the years the format changed to (an ever growing) folio with new columns added – in the early 1860s they were seven. The sparse layout with much white space during the first years developed into compact pages with five times as many words on an average page (from an average of 1,332 words per page in 1831 to 6,982 in 1860). Antiqua was used for almost all texts in *Aftonbladet*, but some of the headlines were printed with fraktur.

In order to understand more about the digitized newspaper beyond the graphical user interface presented at "Svenska dagstidningar," we analyzed the XML files from *Aftonbladet* 1830 to 1862, approximately 10,000 issues of four pages each. A reference corpus of digitized nineteenth century texts was put together in order to filter out as many of the correct words from the newspaper corpus as possible. No digital high-quality corpus of Swedish newspapers exist, but the Swedish Literature Bank – a website for reliable digital versions of Swedish classics – provides access to a variety of transcribed and proofread texts, such as novels, plays, historical writing, travel writing, and some journalism (litteraturbanken.se). In all, 200 texts from the Literature Bank (from the 1820s to the 1880s), were hence used to create a corpus of 370,000 unique words. Naturally, this is only a fragment of all the possibly correct words in the newspaper corpus, and many of the names of people and places are not included. The reference corpus was not used as a key,

defining all the non-matching words in the newspaper corpus as corrupt, but rather as a list of stop words, making it possible to exclude some of the most frequent correct words from further analysis. The 500,000 most frequent newspaper words not matching the references corpus was selected to zoom in on the noisiest part of the digitized newspapers.

The variety of the interpretations generated by OCR was examined further, focusing on a single word and all the corrupted versions of it. We used the word "telegraf" (as the electric telegraph was our initial research topic) in *Aftonbladet* 1830 to 1862 as a starting point. In order to find as many versions of this word as possible we searched with a Levenshtein distance of three, an edit distance which makes it possible to find every version where three or less letters are missing, replaced or added. Irrelevant matches were removed, as well as compounds and suffixes (e.g. "telcgrjifmadcrrättelscr," i.e "telegraph messages").

In addition to these digital methods we visited the digitization facility to get a better understanding of the process, how the work is organized, and how the staff think about their tasks. We did not conduct a thorough examination of every step of the process, but we were able to grasp some of the institutional and practical difficulties characterizing the mass digitization of historical newspapers.

In total, 15 million pages have so far been digitized by the National Library of Sweden, and many millions more will be digitized in the years to come (if funding is secured). The newspaper database we dissect is a longed for resource among Swedish (and international) scholars. Still, historical newspapers have proven very difficult to digitize with good quality. The present result, as we will show, in fact contains extreme amounts of noise: millions of misinterpreted words generated by OCR, and millions of texts chopped off randomly by the auto-segmentation tool. These multiple errors and the noisy digital heritage they create, can be seen as traces of the digital processes, and be examined to reveal how tools are programmed. Hence, in this paper, we try to answer questions like: What happens with the newspapers during the digitization process? What kinds of software are used during the process and how do they reformat the newspapers? What is it actually that generates all the noise? And how can cultural heritage institutions guarantee quality and provenience when digitization depends on commercial software packages they cannot access and control? Following classical information theory, a newspaper (signal) is the information to be transmitted, while noise becomes the product of the channel (the digitization process). Analyzing errors means "listening to the noise of the transmitting system itself," following media archeologist Wolfgang Ernst (2013) – with the ultimate aim to uncover the operations of the system that generates all noise (p. 68).

Noisy media theory

In classic information theory, the signal is usually defined as useful information and noise as a disturbance. Noise has generally been understood as distorting the signal, making it unintelligible and/or impossible to understand. Eliminating noise was, in short, paramount within information theory – but, it also lead to the fact that noise *per se* became an analytical category. Information theory was, in short, always interested in noise. As is well known, Claude Shannon's article, "A Mathematical Theory of Communication" (1948) – which envisioned a new way to enhance the general theory of communication – basically concentrated on noise. As has often been stressed, Shannon was not interested in messages with "meaning." All semantic aspects of communication were deemed "irrelevant to the engineering problem" – but noise was not. Shannon (1964) stated that he wanted "to include a number of new factors, in particular the effect of noise in the channel," where the fundamental problem of communication, to his mind, was that of "reproducing at one point either exactly or approximately a message selected at another point" (p. 31). As is evident, contemporary digitization activities display a number of resemblances and affinities to these remarks and arguments.

Classical information theory became popular when Shannon and his co-author, Warren Weaver, published their book, *The Mathematical Theory of Communication* in 1949. The same year, Weaver had published an article analyzing the ways in which a more "human communication theory might be developed out of Shannon's mathematical theorems" (Rogers and Valente, 1993, p. 39). In it Weaver stated that "information must not be confused with meaning," but more importantly (for our paper), he wrote a longer passage on "the general characteristics of noise." How does noise "affect the accuracy of the message finally received at the destination? How can one minimize the undesirable effects of noise, and to what extent can they be eliminated?," Weaver asked. If noise was introduced into a system – like a digitization process – then the "received message contains certain distortions, certain errors, certain extraneous material, that would certainly lead one to say that the received message exhibits, because of the effects of the noise, an increased uncertainty." Yet, as Weaver (1964) paradoxically stated, if uncertainty is increased, then information is also increased – "as though the noise were beneficial!" This type of uncertainty which arose because of errors or "because of the influence of noise," Weaver however described as an "undesirable uncertainty" (pp. 8, 19).

Within classical information theory, noise could in other words also be described as beneficial. In general, however, noise was a dysfunctional factor; the task was combating noise. Consequently, Shannon and Weaver's mechanistic model of communication mostly dealt with the signal-to-noise ratio within various technical systems. Obviously, their model was indifferent to the nature of the medium. It has, however, since been argued that the arrival of a new medium always changes the relation (or ratio) between noise and information. Digitization processes are no exception. Within German media theory and media archeology, noise has, for example, often been used as a productive analytical category.

Media archeology is part of a gradually shifting emphasis toward media specific readings of the computational base and the mathematical structures underlying actual hardware and software, a transition with analogies to Shannon that also resonates with an increased interest in technically rigorous ways of understanding both software and the operations of material technologies. Analyzing accidents, errors and deviations has, for example, been one strategy to approach systems and technologies that are hard to grasp as long as they function properly. As Jussi Parikka has written (in his English introduction to Wolfgang Ernst's writings), "more than once, Ernst asks the question 'Message or noise?' " – a question that, according to Parikka (2013), is "about finding what in the semantically noisy is actually still analytically useful when investigated with the cold gaze of media archaeology" (p. 36). Another German media theorist, Sybille Krämer (2015), has even stated that various forms of analyses under the hood is the *only* way to make the functions of media technologies visible: "only noise, dysfunction and disturbance make the medium itself noticeable" (p. 31).

One does not, however, has to accept these media theoreticians' definitive claims, to make noise beneficial in an analysis of the digitization technologies that transform printed texts to digital files. Misinterpretations produced by the OCR make explicit what graphical elements the software interprets as important and "meaningful," and errors in the auto-segmentation show what the tool is programed to recognize as a "text." No more, no less. Our perspectives and analyses in the following are thus far more profane and empirical – yet, still informed by the noisy media theories described above.

Digitizing newspapers, generating noise

Newspapers are traditionally described in quantitative terms, according to page numbers and weight, in meters of columns and meters of shelf space. For digital newspaper databases, large numbers are often a goal in themselves. The results of projects are given in the millions of pages digitized so far – as well as the many millions waiting to be digitized. The Library of

Congress, for example, announced in 2015 that its database, Chronicling America (with historic newspaper pages from 1836 to 1922), "now features more than 10 million pages – 74 terabytes of total data – from more than 1,900 newspapers in 38 states and territories" (Library of Congress, 2015). In addition, the British Newspaper Archive presents what almost look like a live-updated figure of pages now available online – 18,522,948 pages in March 2017[1]. The National Library of Sweden has in a similar way been announcing an ever increasing number of searchable newspaper pages. At "Svenska dagstidningar" users are told to search with "*" to receive current and up-to-date numbers of available newspapers pages, 15,109,119 in March 2017[2]. The enormous amounts of data, and the expectations of what these numbers might make possible, in many ways corresponds with beliefs and ideas constituting "Big Data" – big size is what matters (Boyd and Crawford, 2012).

Naturally, researching digitized cultural heritage *en masse* can lead to fruitful results; various forms of literary analyses has, for example, taken advantage of the fact that a large body of novels are now available in digital form. Matthew Jockers has, for instance, argued convincingly for a history of literature turned into data processing in his book, *Macroanalysis* (2013), with its analyses of linguistic patterns in 3,346 nineteenth century novels. Using digital methods it is possible to map topics in the novels, the popularity of individual topics, the difference between American, English and Irish novels, and so on (Jockers, 2013). As is well known, Franco Moretti's (2013) label for this type of study – distant reading – has become a digital humanities buzz word. As a "condition of knowledge," distant reading allows researchers to "focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems" (pp. 48-49). The skepticism from traditional scholars of literary history can perhaps be explained by the tendency in distant reading to reduce novels to what newspapers have always been: millions of words.

Both Jockers and Moretti have admitted that some things are lost when literature is analyzed as data from a distant. Still, they claim that benefits of their methods outweigh eventual limitations. There are, however, important differences between digitized novels and digitized newspapers. First of all, the possibilities of analyzing newspapers using digital methods are limited. A novel is usually a distinct textual unit, with a narrative starting on one of the first pages, and closing on one of the last. Text in more than one column is rare, and variations in font sizes and styles are limited. Illustrations are not very common. The number of possible novels to analyze may seem huge, but they are nevertheless limited compared to other forms of print media.

A nineteenth century newspaper is the opposite of the novel in these respects. A single newspaper issue may contain several hundred different texts – articles, telegrams, news items, advertisements, etc. Texts are printed in seven or eight compact columns, with individual texts running from one column to the next. Font sizes and styles show a great variety, even within one single text. Adverts, in particular, experiment with graphical elements, illustrations and font styles. To convert nineteenth century newspaper pages into machine-readable text is therefore a great challenge. Since funding is limited and the papers to be digitized are (in)numerous, most libraries choose to proceed as cost effectively as possible: "We cannot slow down to make things perfect" (Verheusen, 2008, p. 33).

The total number of words in the XML files of *Aftonbladet* is 194 million, with nine million words being unique. An often used Swedish dictionary from 1851 (as a comparison) lists 62,000 words Dalin (1850-1853). Then again, many of the words printed in the newspaper will not be found in a dictionary: names of people and places, foreign words and terms, and so on. The number of unique words is nevertheless preposterous. Henceforth, we asked ourselves what kind of words actually constitutes the historical record. Frequent OCR misinterpretations are well known errors within all textual digitization projects – such as "c" interpreted as an "e," "h" as "li," etc. Such graphic similarities are extremely difficult for

machines to detect (Tanner *et al.*, 2009). The XML files from *Aftonbladet* confirmed these types of errors. The most frequent word in the Swedish language is "och" – that is "and". Not surprisingly, this is also the most frequent word in the newspaper files. Corrupt versions of this word are also common, revealing errors such as "oeh," "ocli," "ech," "eeh," "osh" and "»ch." About 20 other versions can be easily identified – among them somewhat paradoxically "ocr."

Yet, computational misinterpretation of the word "och" is not a profound problem. It is most often a so called "stop word," and not important at all as a search term. To identify more disturbing errors we proceeded to further examine our top 500,000 list of words not matching the reference corpus. As expected we found a lot of cities, names of streets and people in the sample, but some of the most frequent names were also the ones most frequently misinterpreted: "Slockholm" instead of "Stockholm," "Liibeck" instead of "Lübeck," etc. Many of the most frequent words in the sample were not necessarily misinterpretations, but rather correctly interpreted fragments of words: "Stockho," "ningar," "ngen," "onera." These text fragments – and there are several thousands of them, occurring thousands of times – are most likely to be missed when texts are searched and processed with digital tools. As a consequence, when searching the newspaper database a number of articles containing the word "Stockholm" printed in the paper, will not be found (and displayed as a result) in the digitized edition. Most of the items on our list consisted of only a few letters, however, and it was not possible to tell if they were really fragments of words, or short words misinterpreted by OCR – or both: "elt," "ined," "frin," etc.

Another frequent problem that we discovered was words made up of more than letters. Question marks, exclamation marks, semicolons, hashtags, figures, and other signs and symbols replaced letters in countless ways. Every single combination was not very frequent, but the variations were indeed many. "$atan" ("gatan" = street) occurs 39 times, and "©lägenheten" – that is "olägenheten," an inconvenience – only eight. Then again, unique words starting with a question mark in our corpus amount to almost 1,100, and these occurred over 100,000 times. Given these errors, it might not come as a surprise that the Euro symbol appeared already during the nineteenth century: "€ologne."

Other words in the corpus we studied were born digital in more than one sense: HTML entities. The characters " > " and " < " have for some reason been transformed into the entities ">" and "<" – not always, but still thousands of times. It is not, however, always the printed characters that are interpreted in this way. Sometimes a printed letter, or a combination of several letters, is read as " > " or " < ," and then converted into the HTML entities, generating "e >tra" out of "extra," or "fl <ska" out of "flaska." A simple search for these entities in the complete database gives almost 4,000,000 hits. In all likelihood, a bug somewhere in a software has sprinkled the historical newspaper pages with the characteristics of modern day online media.

Our search for "telegraf" with a Levenshtein distance of three revealed 590 different versions of the word. The initial "t" was interpreted as 25 different characters, the first "e" as 22, and the "l" as 19 different characters. Some interpretations were more common than others ("l" for "t," "t" for "l," and so on), but our results nevertheless indicates that a letter can be transformed into almost any other character. In fact, three errors in a single word are not uncommon, as in "ttlsqraf," "ttlagrsf" and "tsiegray." These are all OCR variations of "telegraf" – still, it might be worth stressing that the spell-checker of Google Docs actually identifies these "words" as "telegraph!"

Many of the 590 different versions of the word "telegraf" were generated from chopped-off words, merged with fragments of other words, for example, "TelegDalarSandhaier." This compound was created from words printed in a report on ships entering the Stockholm archipelago, sent by a telegraph and published on July 30, 1853. What was printed in the paper was a "Telegraf-rapport," the telegraph stations "Dalarö," "Sandhamn," and a German

ship, Spieler (corrupted by OCR). All of these words appeared as the first word on four of the lines in the column. Yet, the auto-segmentation tool marked the first letters on every line, and constructed a new word out of the fragments, and made them a part of the column appearing to the left – where an article about wood exports was printed. This is exactly the same type of error that brought the noise into the report about the 1847 telegraph hit by lightning in our introduction. The fragments of words interrupting the story were printed in the column to the right of the telegraph article. There one could (in the real newspaper) find a news item about the early winter in St Petersburg. Part of the news from the Russian capital merged with the story from St Germain, just as the telegraph report on arriving ships became part of the text analyzing wood exports. New texts are hence generated automatically – and often even made up of completely new words.

Obviously, the 590 different interpretations of "telegraf" makes it totally impossible for a user to find all relevant articles and texts when he or she is searching the database "Svenska dagstidningar." In addition, texts re-edited by the segmentation tool also makes it difficult to analyze the corpus using digital methods. Auto-segmentation is part of the process of turning printed characters into machine-readable text. For someone searching the database, text segmentation provides a help to locate search words on the page, and most of the times it does not matter if the segmented text does not match perfectly with the text as it was once printed in the paper.

But for scholars who want to process massive amounts of texts digitally, texts blocks generated by auto-segmentation will cause a lot of problems. Hence, we asked ourselves what kind of texts is actually analyzed if the segmented texts do not match with the texts as they were printed in the paper? In order to select and locate a sample of auto-segmented texts blocks in *Aftonbladet* we used the 590 versions of "telegraf" (and 400 versions of "elektrisk," "electrical") as search words. 1,250 texts blocks were found and read manually. Not a single one of them matched with a text as it was once printed in the newspaper. Some of the text blocks contained several separate news items grouped together, while others divide long articles in smaller parts. The problem identified above, with one column being merged with part of a neighboring column, was also common. Thin black lines separate columns in print, but the segmentation tool often splits them vertically, three or four characters into a column. The chopped-off words generated this way are usually merged, creating very long strings of fragments. These strings are either inserted in the middle of a text from a neighboring column, or attached at the end of text blocks. Statistics on word length reveal that 0.2 percent of all the words in the XML files are longer than 30 characters – the longest one 270 characters. This figure of 0.2 percent may seem small, but it represents 388,000 words, made up of millions of fragments of other words. For a method such as topic modeling to be useful, scholars would have to edit the text blocks manually before doing a digital analysis. Otherwise the results will be fictitious, not the least since they will discover patterns in "texts" never written.

Institutional noise

To get a better understanding of the digitization process we visited the facility where Swedish printed newspapers are presently turned into data. The digitization facility is located in Fränsta, a town on the northern shore of the river Ljungan, about 370 kilometers north of Stockholm. The location is due to regional political rationales – still, the Swedish Media Conversion Center (Mediakonverteringscentrum (MKC)), is also a division of the Swedish National Archives. It is in fact the largest center in Europe specialized on digitizing (paper based) cultural heritage. Besides digitizing the records of the National Archives itself, MKC provides its services to commercial companies as well as different Swedish public agencies, such as the Swedish Tax Agency (old church records) or Lantmäteriet, the Swedish Mapping, Cadastral and Land Registration Authority (regarding maps). Since 2004, MKC has

regularly digitized historical newspapers on behalf of the National Library of Sweden. Initially, work was done on an experimental basis, and later the National Library became something of a regular partner. About 70 employees work at MKC in two shifts with various digitization projects.

Björk (2015, pp. 168-169) points out that a linear perspective on digitization processes might be misleading since several steps are actually parallel. New information is produced at every stage, generating different layers of representations: metadata, transcription, and image. Nonetheless, the staff at MKC often refers to their facility as a "factory," complete with an assembly line and different workstations. The first step within the digitization process takes place at the library repositories where librarians pick out the volumes to be digitized. Since the National Library was provided with the newspaper collection from Uppsala university library a few years ago, they have had two copies at their disposal of almost all newspapers. Basically, the copy most suitable for scanning has been chosen and transformed into a "disposable copy." These newspaper copies are cut into loose sheets at MKC in order to speed up the scanning process. When the "content" is captured – the disposable copy is discarded. To digitize newspapers hence literally means to save and protect content, i.e. texts and images printed on paper, from the fragile paper medium *itself*. In short, what is digitized are not newspapers, but the part of (and in them) labeled as content (Manoff, 2006).

As the local management showed us around at the MKC facility in April 2016 we followed the paper trail from the room where the "raw material" arrived from the library, to the rooms where the digital files were processed, packaged and made ready for delivery back again to the library. The first steps involved preparation – inspection, metadata corrections, and, if possible, the cutting of bound volumes into loose sheets. Then came the actual image capture – with scanner or camera, depending on the source document – as well as image quality control (of every image or a sample). To create high-quality images in a digitally sustainable format (at present, JPEG2000) has been paramount for the National Library (Kungliga biblioteket, 2011). International comparisons have made clear that results are impressive; the process generates high-quality images and the process has been described as a "ground breaking project" (Tanner, 2013).

Another important goal has been to make newspaper pages searchable. The results, however, have not been as impressive (at all) as the actual image capture. Compared to similar international digitization projects, newspaper pages have usually been segmented (or zoned) manually in order to improve quality. This time consuming part of the process has often been carried out at data centers in low-income countries such as India or Cambodia[3]. The National Library of Sweden and MKC, however, decided on an early stage to set up a workflow where as many steps as possible were automated, including text segmentation. The segmentation tool used is called Zissor, developed by a Norwegian company with the same name. Zissor is delivered as an integrated package including software for OCR developed by ABBYY. This software package generates auto-segmented pages and OCR-read texts – and as we soon became aware of, a lot of frustration for staff.

How the Zissor tool works, what it accomplishes and brings about is surprisingly mostly unknown to the staff at MKC. Even if all resulting XML files are full of references to the "Segmentation Component for Zissor Content Conversion System" nobody at MKC seems to understand what the software package actually does. Staff members insert some basic instructions, but the options provided by the interface are very limited. The settings for the auto-segmentation tool, for example, includes the most common font styles and sizes for a specific newspaper and time period (five years or more), the number of columns, and so on. Ironically, the staff giving and setting these instructions are not sure what difference they make, however. The limited options provided by the interface does not seem to match the graphical variety of the newspaper pages.

ABBYY represents similar problems. The staff at MKC decides which dictionary to use for a specific time period, but ABBYY will only consult it if the OCR interpretation of a word is estimated to be "uncertain" by the tool itself. Since many "certain" interpretations are in fact misinterpretations, the dictionary is usually of no help at all. At our visit, we suggested that some simple and common sense rules – no numericals or punctuations in the middle of a word, and no acceptance of symbols unknown when the paper was printed during the nineteenth century – would improve results significantly. The ABBYY interface, however, does not allow for such changes. Nor any other changes, in fact. These can only be altered by the Norwegian software developers.

Somewhat paradoxically, the integrated package of tools at MKC thus seems to be too fixed and too flexible at the same time. The interface only allows the staff to give a limited number of instructions – but the tools themselves are very flexible (as we described above) when they generate text blocks and a constant flow of new words. Since the software is more or less working independently from the staff at MKC, it comes as no surprise that small talk about these tools almost resembled speaking about living creatures with agency: "Zissor is the black box in our process;" "ABBYY gives its suggestions – and then we have to stick to that."

Provenance, fundamental to the way any National Library (or traditional library, for that matter) works, describes the chronology of ownership, and the way it (as well as custody or location) affects any historical object: where does a book come from? who has printed it? owned it? and so on. Such a provenance chain – however rudimentary – remains difficult to follow and describe when it comes to digitized newspapers at MKC in Fränsta. In fact, being on location and studying the digitization process, it swiftly become apparent to us that sticking to any idea of a provenance chain is impossible, since so many steps have been outsourced to unknown factors affecting the source document. On the one hand, the sacrosanct software conditions output in concealed ways and on the other hand institutional factors do the same: the library hiring MKC, MKC hiring Zissor, and Zissor hiring ABBYY. They all affect the digitization process – in different and increasingly vague ways. In short, the further up these actors are in the digitization chain, the less they seem to know of the processes that turn papers into digital data.

In the age of paper-based collections libraries formed a "control zone" where quality, stability and authenticity of information was guaranteed: "an item maintained in the control zone must always be, by definition, the unaltered original" (Atkinson, 1996, p. 259). As pointed out by Carl Lagoze (2014), this control zone is increasingly collapsing in a digital environment (Lagoze, 2014). Digitization, understood as "translation" rather than "transfer," is never a neutral process where the original stays unaltered (Dahlström, 2009, p. 31). Translation, however, can be more or less controlled, producing more or less accurate results. It has been stated by Geoffrey Bowker (2005) that "raw data is an oxymoron." Data are never raw, on the contrary, "data should be cooked with care" (Bowker, 2005, p. 184). Tom Boellstorff (2013) has developed this idea further and pointed out that a lot of data today is in fact "rotted." This category represents "the unplanned, unexpected, and accidental" transformations of data, moving between the "intentional and unintentional" (Boellstorff, 2013). This is a useful way to describe the creation of noise in digitized newspaper databases. Noise is produced when software is used to make pages machine readable and searchable. The processes are automated and the noise creation is not intentional – yet, since the tools are programmed to perform in a specific way the result is not unintentional either.

Rotted paper was a well-known problem already at the turn of the century 1900. Newspapers printed on paper made from wood pulp turned yellow and fragile after only a few years (Hill, 1910). The sticky tape used to patch such deteriorating pages together is, in fact, still visible on many digital reproductions. Nevertheless, this kind of preservation could "only temporarily delay the effects of the forces of evil living inside the paper"

(*Arkiv och biblioteksfilmning*, 1951, p. 51). The ongoing digitization of historical newspapers might thus be seen as a final solution to the problem of rotted newspapers. However, as we have argued in this paper, digital preservation generates problems of its own – alas, rotted paper is replaced by rotted data.

Conclusion

During the nineteenth century journalists and newspaper editors frequently copied and pasted news items. These were inserted into and tailored to fit the specificity of the medium of the daily press – form, in short, affected content. In a similar manner as during the nineteenth century – moving from scissors to scanners – digitization today transforms content. In many ways, it seems apt that the segmentation tool used at the digitization factory, MKC in Fränsta is called Zissor. On a daily basis up to 50,000 newspaper pages are digitized, complete with OCR interpretation and automatic article recognition by the Zissor software. As we have shown, however, in this paper, the resulting XML files are extremely full of noise. Form indeed affects content. The scanned, digital collection of *Aftonbladet* contains millions of misinterpreted words, as well as millions of random texts created by the auto-segmentation tool. In fact, the resulting XML files are best described as newspaper content never written.

Even if this paper takes a critical stance and acknowledges that the digitization of nineteenth century newspapers at the Swedish Media Conversion Center has resulted in tremendous amounts of errors and a truly noisy heritage, we are not critical of these digitization activities *per se*. In fact, our scholarly work has been done in constant dialogue with the National Library of Sweden. This paper can hence be perceived as a way to admit, concede and hopefully help the National Library (and fellow academics) to facilitate both a better understanding of digitization as a media transformation process, as well as improving it. Basically, the newspaper scanning procedures at MKC have been geared toward producing high-quality images – not proper textual content computationally gleaned from the sheets. Our research, however, has made it all too apparent for the National Library that the textual output is excessively noisy. As a result of a workshop at the library in March 2016, where we presented similar findings as the ones in this paper, an external scholarly group of experts (including one of us) has therefore been set up at the National Library to scrutinize OCR interpretation, with an eventual assignment to produce guidelines and quality measures regarding textual output of newspaper digitization activities.

In addition, since the National Library has been keen on producing high-quality images of nineteenth century newspapers, there has always been an implicit idea that they might process the resulting image files again (if new software promises to produce superior results). In theory, as media archeologist Wolfgang Ernst (2013) keeps reminding us, digital archives are always in motion: "The esthetics of fixed order is being replaced by permanent reconfigurability" (p. 99). Yet, grey theory in our described case could easily become concrete practice. New software that can re-process image files point toward the permanent reconfigurability Ernst is writing about – not the least since software always gets updated. At the level of digital documents change has already in many ways become the new norm. If a heritage institution as the National Library of Sweden for more than a decade made real efforts to try to digitize newspapers collections, they did not really envision the ways in which the resulting files (might) alter, permute and even re-arrange their entire archival infrastructure. With hindsight Paul Conway's (2015) description of digital collections as organic entities hence seems increasingly apt: "behind the scenes, in server rooms and on the desktops of systems administrators, 'artificial' digital collections are organic entities that grow and change their shape as new materials are added, new contextual relationships are established among objects, and new procedures are brought to bear on the organization and management of these large collections" (Conway, 2015, p. 65).

Mutable digital collections thus blur the boundary between a fixed and traditional archive (ordinary newspapers) and a digital one (scanned newspapers) – yet at the same time also point toward the need for heritage institutions to reconceptualize what a digitized archive actually is, since large collection of "digital surrogates" naturally also hold archival qualities. Then again, in terms of noisy heritage this causes and creates entirely new perplexities for researchers too. On the one hand, as we have shown in this paper the digital copy is far from identical in relation to the original paper version; letters never printed in old newspapers are now indeed part of the historical record. But if the same digital material is processed again, on the other hand, the new digital copy will result in a copy that also differs from the original digital copy. Hence, Warren Weaver's undesirable uncertainty seems to be returning. Once upon a time, the hallmark of the press as a medium was that most printed copies were identical. Digitizing old newspapers, however, always means that we will be distancing ourselves from both originals as well as more or less auratic digital copies. Within the digital newspaper archives to come – produced by the National Library of Sweden and based on the nineteenth century newspapers digitized at MKC in Fränsta – there will hence all likely not only be a difference between original and copy. There will also be a deviation between the different interpreted copies of the original scanned newspaper page.

Notes

1. www.britishnewspaperarchive.co.uk (accessed March 2, 2017).
2. tidningar.kb.se (accessed March 2, 2017).
3. "Turen går til Indien" (2014), https://blog.avisdigitalisering.dk/2014/02/01/turen-gar-til-indien/; "In the digitization factory" (2010), www.betatales.com/2010/10/07/in-the-digitalization-factory/ (both accessed June 28, 2016).

References

Aftonbladet (1847), "Åskan och elektriska telegrafen", *Aftonbladet*, Stockholm, October 16, p. 3.

Arkiv och biblioteksfilmning (1951), Statens offentliga utredningar, Arkiv och biblioteksfilmning, Stockholm.

Assmann, A. (2008), "Canon and archive", in Erll, A. and Nünning, A. (Eds), *Cultural Memory Studies: An International and Interdisciplinary Handbook*, Walter de Gruyter, Berlin and New York, NY, pp. 97-107.

Atkinson, R. (1996), "Library functions, scholarly communication, and the foundation of the digital library: laying claim to the control zone", *The Library Quarterly: Information, Community, Policy*, Vol. 66 No. 3, pp. 239-265.

Björk, L. (2015), *How Reproductive is a Reproduction? Digital Transmission of Textbased Documents*, Swedish School of Library and Information Science, Borås.

Boellstorff, T. (2013), "Making big data, in theory", *First Monday*, Vol. 18 No. 10.

Bowker, G.C. (2005), *Memory Practices in the Sciences*, MIT Press, Cambridge, MA.

Boyd, D. and Crawford, K. (2012), "Critical questions for big data", *Information, Communication & Society*, Vol. 15 No. 5, pp. 662-679.

Conway, P. (2015), "Digital transformations and the archival nature of surrogates", *Archival Science*, Vol. 15 No. 1, pp. 51-69.

Dahlström, M. (2009), "The complete edition", in Deegan, M. and Sutherland, K. (Eds), *Text Editing, Print and the Digital World*, Ashgate, Farnham, pp. 27-44.

Dalin, A.F. (1850-1853), *Ordbok Öfver svenska språket*, Vol I-II, Stockholm, available at: https://spraakbanken.gu.se/swe/resurs/dalin (accessed June 28, 2016).

Digsam (2014), "Digisam's guiding principles", available at: www.digisam.se/digisam-s-guiding-principles-now-translated-into-english/?lang=en (accessed March 2, 2017).

Ernst, W. (2013), *Digital Memory and the Archive*, University of Minnesota Press, Minneapolis, MN.

Hill, F.P. (1910), "Deterioration of newspaper paper", *Bulletin of the American Library Association*, Vol. 4 No. 5, pp. 675-678.

Hitchcock, T. (2013), "Confronting the digital", *Cultural and Social History*, Vol. 10 No. 1, pp. 9-23.

Holley, R. (2009), "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs", *D-Lib Magazine*, Vol. 15 Nos 3/4.

Jockers, M. (2013), *Macroanalysis: Digital Methods & Literary History*, University of Illinois Press, Urbana, IL.

Ketelaar, E. (2001), "Tacit narratives: the meaning of archives", *Archival Science*, Vol. 1 No. 2, pp. 134-155.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. and Kervinen, J. (2014), "Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods", paper presented at IFLA, Lyon, available at: https://helda.helsinki.fi/handle/10138/136269 (accessed February 21, 2017).

Krämer, S. (2015), *Medium, Messenger, Transmission: An Approach to Media Philosophy*, Amsterdam University Press, Amsterdam.

Kungliga biblioteket (2011), "Kravspecifikation DD v.1.2", available at: http://digidaily.blogg.kb.se/dokument/ (accessed June 28, 2016).

Kungliga biblioteket (2016), "Digitalisering på Kungliga biblioteket", available at: www.kb.se/om/verksamhet/digitalisering/ (accessed March 2, 2017).

Lagoze, C. (2014), "Big data, data integrity, and the fracturing of the control zone", *Big Data & Society*, Vol. 1 No. 2, pp. 1-11.

Library of Congress (2015), "Online resource of historic newspapers posts 10 millionth page", available at: www.loc.gov/today/pr/2015/15-171.html (accessed June 28, 2016).

Manoff, M. (2006), "The materiality of digital collections: theoretical and historical perspectives", *Portal: Libraries and the Academy*, Vol. 6 No. 3, pp. 311-325.

Milligan, I. (2013), "Illusionary order: online databases, optical character recognition, and Canadian history, 1997-2012", *The Canadian Historical Review*, Vol. 94 No. 4, pp. 540-569.

Moretti, F. (2013), *Distant Reading*, Verso, London.

Mussell, J. (2012), *The Nineteenth-Century Press in the Digital Age*, Palgrave, London.

Parikka, J. (2012), *What is Media Archaeology?*, Polity, Cambridge.

Parikka, J. (2013), "The media-archaeological method", in Ernst, W. (Ed.), *Digital Memory and the Archive*, University of Minnesota Press, Minneapolis, MN, pp. 1-22.

Peters, J.D. (2015b), *The Marvelous Clouds: Toward a Philosophy of Elemental Media*, The University of Chicago Press, Chicago, IL.

Peters, J.D. (2015a), "Proliferation and obsolescence of the historical record in the digital era", in Tischleder, B. and Wasserman, S. (Eds), *Cultures of Obsolescence: History, Materiality, and the Digital Age*, Palgrave, New York, NY, pp. 79-96.

Pletschacher, S., Clausner, C. and Antonacopoulos, A. (2015), "Europeana newspaper OCR workflow evaluation", *HIP '15 Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*, pp. 39-46.

Rogers, E.M. and Valente, T.W. (1993), "A history of information theory in communications research", in Schement, J.R. and Ruben, B.D. (Eds), *Between Communication and Information*, Transaction Publishers, New Brunswick, pp. 35-56.

Shannon, C.E. (1964), "A mathematical theory of communication", in Shannon, C.E. and Weaver, W. (Eds), *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, IL, pp. 29-125.

Tanner, S. (2013), "World class digitisation in Sweden", available at: http://simon-tanner.blogspot.se/2013/03/world-class-digitisation-in-sweden.html (accessed June 28, 2016).

Tanner, S., Muñoz, T. and Ros, P.H. (2009), "Measuring mass text digitization quality and usefulness: lessons learned from assessing the OCR accuracy of the British Library's 19th century online newspaper archive", *D-Lib Magazine*, Vol. 15 Nos 7/8.

Verheusen, A. (2008), "Mass digitization by libraries: issues concerning organisation, quality and efficiency", *Liber Quarterly*, Vol. 18 No. 1, pp. 28-38.

Weaver, W. (1964), "Recent contributions to the mathematical theory of communication", in Shannon, C.E. and Weaver, W. (Eds), *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, IL, pp. 1-28.

About the authors

Johan Jarlbrink is an Associate Professor in Media History and a Senior Lecturer in Media Studies at the Umeå University, Sweden. The main focus of his research is the use and transformation of nineteenth century newspapers: reading and collecting, copy-and-paste and archiving, newspapers as waste and empirical material in historical research, etc. Johan Jarlbrink is the corresponding author and can be contacted at: johan.jarlbrink@umu.se

Pelle Snickars is a Professor of Media and Communication Studies at the Umeå University, Sweden. His research focuses on the relationship between older and new media, media economy, and the digitization of cultural heritage. International publications include *The YouTube Reader* (2009) and *Moving Data: The iPhone and the Future of Media* (2012).