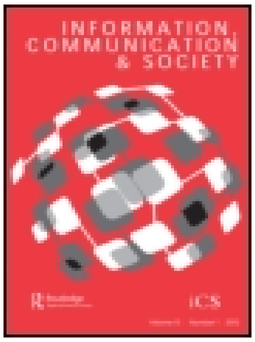

Routledge
Taylor & Francis Group

Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: http://www.tandfonline.com/loi/rics20

Critically assessing digital documents: materiality and the interpretative role of software

James Allen-Robertson

To cite this article: James Allen-Robertson (2017): Critically assessing digital documents: materiality and the interpretative role of software, Information, Communication & Society, DOI: 10.1080/1369118X.2017.1351575

To link to this article: http://dx.doi.org/10.1080/1369118X.2017.1351575

Published online: 17 Jul 2017.

INFORMATION, COMMUNICATION & SOCIETY, 2017
https://doi.org/10.1080/1369118X.2017.1351575

Routledge
Taylor & Francis Group

Critically assessing digital documents: materiality and the interpretative role of software

James Allen-Robertson

Department of Sociology, University of Essex, Colchester, UK

ABSTRACT

As a contribution to the ongoing tradition of critically assessing documents for research, this paper aims to highlight materiality as a key factor in the co-shaping of knowledge derived from digital documents. The paper first builds upon prior debates in document studies with work from the fields of Science and Technology Studies, and Communication Studies, to establish the role of document materiality in the interpretative process. By first establishing digital documents' material reality as electrical signal, the paper then discusses the interpretative role of software, in both the representation of that signal for human interpretation and the production of the document through software tools. Finally, the paper considers the implications for persistence and access to digital documents posed by their material reality and the private archival contexts in which they often reside.

ARTICLE HISTORY
Received 20 December 2016
Accepted 3 July 2017

KEYWORDS
Digital document; materiality; software; critical assessment

Introduction

An encroaching reality for documentary researchers in sociology is that the practice of digital documentation is rapidly increasing. Individual's lives are documented in unprecedented ways through social media, blogging, video and audio. Institutions are also generating huge amounts of data, some of which is presented in the style of a 'document' through reports, white papers and so forth, but much of it is 'data' fed into other algorithmic processes for logging, monitoring and to spur further automated action. The internet is our predominant form of archiving, storing and indexing the everyday. As more human and non-human agents participate in this archival process, its scope grows in both breadth and depth. Whilst the possibilities for qualitative research expand with this ongoing documentation, through, for example, visualisation and computational linguistics from scholars in the Digital Humanities and Digital Sociology (Berry, 2012; Marres, 2017), it is imperative that we recognise and critically assess these sources of knowledge as new forms.

The critical assessment of knowledge sources has a long tradition in the social sciences, predominantly in their application to qualitative research and the traditional methods of interviews, observation, and the analysis of non-virtual written or visual documents. This body of work recognises the unavoidable distortions introduced by the human actors that

CONTACT James Allen-Robertson jallenh@essex.ac.uk Department of Sociology, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK

both produce knowledge and the researchers that interpret it. Accepting these distortions, researchers have deployed a reflexive approach to their knowledge sources that recognises that factors such as the identity of the author and the interpretative position of the reader necessitate an awareness of the social construction of meaning and the need for linguistic reflexivity (Alvesson & Sköldberg, 2010; Platt, 1981). These distortions apply equally to digital documents.

Qualitative researchers who have engaged with the internet as a source or space of research have also furthered our critical assessment of virtual sources, recognising how both well-established and new concerns apply to our virtual sources. These concerns extend the assessment of 'authenticity' and 'authority' to digital spaces (Hine, 2000), the ways in which internet-mediated communication might shape individual experience, communities and the structure of texts (Jones, 1999), and the opportunities for conducting traditional research methods in, for example, accessing previously overlooked groups and the facilitation of new 'viral' sampling techniques (Kazmer & Xie, 2008; Palys & Atchison, 2012; Seymour, 2001). Other scholars have looked at the changing nature of research methods, particularly in relation to social media as a source of rich meta-data heavy digital objects, such as tweets (Bruns, 2012). Reflexive researchers engaged in using digital sources and tools are retaining a critical stance in assessing these digital sources. However, this paper suggests that we should also consider the material as a hitherto overlooked factor that co-shapes our interpretations of digital documents.

Today many of the documents that are produced by us, and about us, are products of digital electronic computing. No longer familiar in their materiality the documents of our lives arise and persist as signals confined within software and hardware assemblages. Yet in a vast number of social spheres, we have adopted the idea of the 'digital document' quite readily, acclimatised through increasingly user-friendly software that express and mimic the typographic conventions of print culture. This familiarity disguises the significant material difference from non-virtual document objects. Discussion of documents as qualitative sources of knowledge has predominantly focussed on their content, rather than them as material objects in use (Coffey, 2014; Platt, 1981; Plummer, 2001; Scott, 1990). Prior (2008) has argued for an extension of this understanding, 'repositioning' documents in a way that also recognises them as both containers of knowledge and objects of action within the social world. By extending this recognition of documents as objects in their own right, this paper argues that by recognising documents as material objects of interpretation, we can then ask what might be the 'consequences of form' (Dourish & Mazmanian, 2013, p. 96) when our documentation predominantly comes in the form of electrical signal rather than as ink on paper.

The aim of this paper is to contribute to the critical assessment of documentary sources by highlighting the less considered role of materiality and its continuing role in the interpretation of digital documents. By first establishing materiality as a key factor in document interpretation and then extending it to digital materiality, the paper then addresses the challenges posed by digital materiality. These include the role of software interpretation as co-reader of digital documents; the influence of software as co-author of digital documents; and the entanglement of software and social processes in the persistence and discovery of digital documents. As such, the aim of this paper is to contribute to the continuing debates regarding the limitations of knowledge obtained from the sources we examine, to further build upon our understanding of the mediating role of computer-mediated communication in the practice of research, from a foundation that identifies materiality as a key factor that co-shapes the knowledge derived from digital documents.

Documents as material technology

The emergence and growing prominence of digital documents has highlighted concerns whether the digital document is in some way different from printed documents and so necessitates a redefinition of the term. The fundamental question of what is and is not a document has been a matter of discussion from European Documentation scholars such as Otlet (1934) and Briet (Briet et al., 1951/2006, p. 9), Sociologists such as Sidney and Beatrice Webb (Webb & Webb, 1932) and more contemporary scholars working under a broad range of disciplines that could be classified as Information Studies (Buckland, 1998; Day, 2014; Frohmann, 2009; Lund, 2009).

European documentation was predominantly concerned with both defining what we meant by the term 'document', and with developing efficient practices to utilise documents as sources of knowledge. For Otlet (1990, p. 83), the technological design of an organisational system played a crucial role in the utilisation and production of knowledge through efficiently bringing together users with a particular 'informational need', and the appropriate document that could satisfy it (Day, 2014). However, Otlet's thinking was also guided by an essentialist understanding of documents that framed them as containers of extractable objective knowledge. His aim was to produce an informational system that linked together these extracted pieces of knowledge in such a way that it would objectively reflect reality itself (Ducheyne, 2009). Under Otlet's framing of documents, though his system's material design was imperative in the joining of users to knowledge, the materiality of the object itself was considered superfluous to the knowledge it 'contained'.

Briet (Briet et al., 1951/2006) shared Otlet's interest in information systems but broke from the metaphorical model of containers and conduits and essential knowledge content for a more interpretative position. For Briet what counted as knowledge was changeable, subject to shifting information needs brought about by wider technological, political and cultural change. Since we could have no certainty as to what did, or what would, count as knowledge, or documentation of that knowledge in future social contexts, the very definition of a document must be flexible. For Briet, any material object could qualify if they were interpretable as meaningful;'[a] document is a proof in support of a fact' (Briet et al., 1951/2006, p. 9). As such under Briet's position, any material object can be interpreted as documentary. For Buckland (1998), considering whether digital documents challenge our understanding of documents, Briet's position indicates that the transition to digital documents is unproblematic. If a digital object can be interpreted by the user as a meaningful object, as evidence of a fact, then it is a document, rendering considerations of 'what' that document is superfluous, what matters is the interpretation of it. Similarly, Frohmann (2009) argues that to assert a definition of documents based upon rigid categorisations is futile. We create criteria to ensure clarity of meaning, yet the criteria are not always already there but are instead a matter of context, produced *in situ*. We have flexibly adapted and integrated the concept of the 'digital document' because it makes sense in our current social context. This paper concurrently supports and challenges this interpretative position. If like Otlet we consider that a document is a container for some essential extractable piece of information, then a shift to digital documentation, as a shift from one container to another, is broadly unproblematic. However, if we recognise that information arises from a confluence of material form and its interpretation, then that radical change in the systems and forms of documentation necessitates consideration. Whilst the knowledge derived from a document is interpreted, not essential, this does not necessarily discount the need to recognise that the predominant form of *what* we interpret has changed, and that our interpretation may in some way be influenced by the material form.

This is not an issue specific to digital documents but is part of a wider 'dematerialisation' of digital media where our focus emphasised the informational, the epiphenomena of computing machinery and de-emphasised the machinery itself. This dematerialisation has been furthered by a rhetoric of immateriality through popular dichotomies such as 'physical/digital' and phrases such as 'in the cloud' that position the digital as 'pure' information, discounting its ontology as a material, mediating thing. Strands of work in the field of media, communication and Science and Technology Studies (STS) are challenging this frame of immateriality, appealing to both historical and mechanistic approaches to resituate the digital into the corpus of clunky imperfect meaning mediating technologies (Allen-Robertson, 2015).

Furthermore, there has already been significant work headed by Lievrouw (2014) to connect these material-focussed approaches in Communication Studies with STS, where interest in materiality has seen resurgence with the Actor-Network Theory of Latour (2005) and the socially and materially aware concept of 'affordances' from Hutchby (2001). This confluence of the material-focussed strands of communications studies, with the broader study of technology as material and social constructs, reframe media as material technological systems that have a significant role in generating information through the confluence of their materiality and social use. Under this frame media objects are no longer identified simply as the information they signify, but as material signifying *objects*. Ontological privilege is reclaimed by the materials with the recognition that they are arranged and collide in specific ways to generate phenomena that might then be interpreted as content. For many sociological researchers signifying objects such as books, diaries, tapes, film, papers, photographs, newspapers and audio remain the most informationally valuable documents because behind their production is an intent to store and communicate meaning. However, as asserted by this emerging body of work, the intention of a digital document's design does not exclude them from consideration as a material object like any other. The underlying understanding of what constitutes 'materiality' has been a matter of substantial debate within a wide range of academic fields, due particularly to the broader constructivist turn (Sterne, 2014). Kirschenbaum's (2008) framework of materiality recognises that signifying objects have 'forensic' and 'formal' materialities, the former being the atomic physical substrate and its operation, whilst the latter is the shape and organisation of the inscribed message. These materialities are interdependent, the forensic impinges upon the affordances of what message may be inscribed, whilst the formal message may impinge upon the operation of the mechanism itself (Allen-Robertson, 2015). Furthermore, these more object-oriented aspects of materiality must be recognised as interdependent upon social practice and social arrangements (Lievrouw, 2014). It is a complex issue, a full examination of which is beyond the scope of this article which seeks to specifically highlight the continuing role of materiality's forensic and formal influence in digital documents (for more detailed discussion, see Allen-Robertson, 2015).

As not all interdependent factors can be addressed at once, this article uses Hutchby's (2001) concept of 'affordances', identified by Lievrouw (2014) as an effective balance in the classic STS technological versus social determinism debate, in order to speak of the material whilst retaining recognition of the social. Under Hutchby's framework, technology's role in society is recognised to be a product of both its material construction and its social use. The object does not have inherent 'capacities', specific uses or impacts designed in, but its material construction and design will limit the possible range of interpretations and influence. This echoes much of the literature that considers technology to be a social construct, a view supported by historical case studies demonstrating the differing ways in which material technologies were used and understood by different social groups (Bijker, Hughes, & Pinch, 1987). However, for Hutchby (2001), as vast as the possible range of uses could possibly be, it is not an infinite range. Thus, rather than determine use, technologies 'afford' or have affordances that frame the range of possible use.

If we extend 'affordances' to documents, as objects from which we can derive meaning through their use, we can see how the evidentiary nature of a document arises from the confluence of material form and social interpretation. A photograph at the material level is a flat surface of varying inks, but what it depicts is determined by the observer, and may differ depending upon who is observing it. However, the possible interpretations of what the photograph depicts is not an infinite range, and will in some way be guided by the material constitution of the object itself. It is medium by its design, document by its interpretation, but fundamentally it is a material object in use, no different from any other material form.

Levy (2001) illustrates the implications of material and contextual change in paper documents by tracing the changing form of Walt Whitman's *Leaves of grass* across various print editions. The many print editions do ostensibly the same job of presenting a reader with typographic forms on a flat surface to communicate the poems. However, though all editions are identified to be the same (i.e., they are all Whitman's *Leaves of grass*), they differ in the kinds of inferences that might be drawn by a reader. Different editions changed the aesthetic presentation, the order of poems, omitted or added poems, and provided explanatory footnotes, titles and photographs added where previously there were none. All these features played a role in what inferences may be drawn from the book as an interpreted object. Rather than rely on just the textual aspects, literary interpretation became intertwined with the practices of bibliography, printing and publishing (McKenzie, 1999).

Going beyond typographic features, the material composition of the document itself may also support productive interpretation. Material elements either constitutive of the original artefact or present due to its persistence through time and space can be crucial in the generation of information. Elements such as the object's chemical composition, damage such as watermarks, paste marks or rust, or even the arrangement of an artefact set, and the containers they reside in, can all be informational to the right user. Brown and Duguid (2000, p. 173) detail how for one researcher trying to map outbreaks of eighteenth-century cholera, even scent, was a significant attribute. A hint of vinegar on the paper would indicate the letter had been subjected to the postal system's rudimentary attempts at disinfection, and thus that the local area was attempting to reduce further transmission of the disease.

> The sizes, shapes and weights of records structure physical interactions between records and their users, and changes in their presentation and physical condition may provide evidence of their histories of use and stewardship … Materiality, the material expression of human ideas, is therefore perhaps the most primary of sources regarding the circumstances of the records' creation … (Rekrut, 2014, p. 238)

For archivists, these material affordances of documentary objects have come into relief precisely because of their attempts to digitise their collections. Digitisation, by producing a second object intended to communicate the same information using a different material form, has demonstrated the way in which form and content are intertwined. Often the initial motivation to digitise is predicated on the recognition that changing from one form of use object to another offers a range of affordances different from the original object. Museum archives may digitise to improve access by making the documents network transmutable, and therefore viewable at home via the web. However, the production of these materially different embodiments of documents has also led archivists to question what constitutes the 'significant properties' (Yeo, 2010) of the original archive materials. Confronted with the technical reality that a digital copy cannot embody every element of the material artefact, such as scent, weight, three-dimensional shape and physical condition, they must make crucial decisions regarding what properties should be represented. It is a social editorial process that must work within the affordances of the software and hardware available to produce the object (Yeo, 2010).

McGregor (2014), drawing on her experiences whilst in the midst of a large digitisation project, observed that the process of translating an archive of magazines, a highly visual medium, into a database, disrupted the spatial and temporal arrangements of how one reads a magazine. The digital versions prioritised the textual over the visual, and the unique over the repetitious so important to print serialisation. Typographic elements of the magazines that repeated every week were discarded as their repetition implied a lack of importance, whilst the text which did differ with each edition was considered of greater importance. Qualities such as the arrangement of the text on the page, use of colour and photography, all indicative of the magazine's moment of publication, became lost in a relational database of 'content'. The database structure afforded new uses, and therefore new information generated through use, but there was also a loss of affordances as some properties of print were not translated into the database structure. The result is an obstruction of a reading that includes a historical 'sociomaterial sensitivity' (Carlile Nicolini, Langley, & Tsoukas, 2013) to them as discrete actors in the world, leaving them instead severed from origin and context. The object is reframed under the context of the information system it has entered, and its aesthetic content amalgamates with the politics and marketplace of the custodian (Sassoon, 2007; Stewart, 1984). Just as the introduction of hanging filing systems in the early twentieth century afforded drastically different ways of organising, manipulating and using the contained documents (Yates, 1989), the archival systems within which these documents reside shape them. As such, digitisation projects change both the material form of the document and the range of possible interactions the user may engage in.

For Levy (2001), though the differing editions of *Leaves of grass* each altered the range of interpretive possibilities for the reader, it was the digitised online copy that was the most drastically different edition. The copy was produced by running one of these print editions through optical character recognition (OCR), and scanning the graphical elements. The text generated by the OCR was wrapped in HTML code to make it suitable for display on the web. Each poem occupied its own web page, connected to the others through an index page, but also accessible via a search engine to locate specific strings of text within any of the poems. Each poem retained its footnotes, but utilised hyperlinks to allow swift access from within the poem text. The images, once *in situ* within the folds of the pages, were provided on a separate single page, decontextualised from their relationship to the text. Levy notes that these images were eventually removed from the digital version altogether when the market pressure of hosting costs became prohibitive. Though the typographic content was broadly the same as the printed version, the decoupling of it from one form and instantiating it in another drastically altered the nature of the work itself. In its new form, users were offered significantly different affordances of using the document (automated searching, indexing and hyperlinking across poems) yet the typographic arrangement of individual poems, the structural arrangement or even availability of the non-textual elements were also subject to the limitations of HTML and the wider archival context that required the continued operation of servers and internet connectivity for it to persist. *Leaves of grass* was remade, discarding all features except text and image and persisted only in part, due to the priorities of its new archival context.

Digitisation projects provide us with a comparative illustration of the ways in which the material form might impact upon our range of interpretative affordances, not necessarily as a ‘loss’ but a shift in the interpretative possibility afforded. For digitally native documents we must recognise that even without a point of comparison to illustrate it, our range of agency and interpretation are also shaped by their material features. For digital documents, as we will see, a key aspect of digital materiality is its intrinsic reliance upon mediation, and non-human interpretation.

The interpretative role of software

For any digital document, whether natively produced, or produced in reference to a prior documentary form, its material reality as electrical signal means that it comes under the classification of ‘data’. Rather than being distinct from the normal operation of computing, digital documents are data that have been interpreted as having substantial document-like qualities. This interpretation may occur because the observer notes the mimicry of prior document forms, or simply, as under Briet’s terms, consider it to be of evidentiary value. Yet before the user can interpret the object to come to such conclusions, the data must first be interpreted from signal, to data, to document by software. A digital document may mimic prior forms, but only because software has interpreted the data into such a form. A digital document may be interpreted as being of evidentiary value, but only once software and interface have rendered electrical signal into human sensible form. Following on from the recognition of materiality as contributory to our interpretation and use of documents, digital and non-digital alike, this paper’s second contribution is the recognition of software as an intermediary interpreter necessitated by the digital document’s material form. This interpretative process happens both at the point of a document’s consumption, as it is interpreted by a user, but also in the process of production, shaping the range of possible expressions available to the author. In occupying both these roles, software is the unnoticed co-reader and co-author of digital documentary forms, introducing an often-overlooked layer of complexity into our assessment and interpretation of digital documents.

As co-reader

‘Raw data’, as Gitelman (2016) points out, do not exist. Data are always the product of a directed attempt (whether human overseen or human designed) to measure and classify some phenomena as meaningful. Equally, a digital document does not exist without the directed attempt to take electrical signals and represent it as something meaningful at human scale. The term ‘digital document’ is something of an anachronism. It is an imposition of print culture conventions upon calculating machines that makes little sense at the technical level where very little is meaningful at human scale.

Digital documents must first be mediated by many layers of software interpretation before we can derive any meaning from them. Consider, for example, a digital document that at the interface level appears very much like its paper equivalent, a PDF of a print journal article. The PDF file draws much of its authority through its mimicry of paper, invoking print’s ‘long complex association with what is’ (Gitelman, 2014, p. 113), but it is a mimicry that is only possible through software’s mediation. The PDF persists (i.e., It remains accessible to you over time) inscribed on your computer hard drive, not as text but as a signal. The signal is possible because the hard drive surface is made up of billions of magnetically charged granules. Some are charged positively, some negatively. A collection of consecutive granules similarly charged represents one ‘bit’, the smallest measurement of data. An average journal article PDF will require somewhere between approximately 800,000 and 8,000,000 bits. One square inch of hard-disk drive surface could easily accommodate 12,000 copies of this PDF, if you were so inclined. Not only is this electrical signal the only aspect of the document that, broadly speaking, persists over time as the inscription from which the document will be generated, its unmediated form would undoubtedly elicit a very different interpretation through human use than the fully fledged, mediated, machine-interpreted PDF as experienced via the interface (Allen-Robertson, 2015).

The PDF, as experienced by the user at screen level, is the result of layers and layers of interpretative software and hardware operations. At the point of ‘opening’ the PDF, the most immediate level of software to the user is the application layer (such as *Adobe Reader*), itself dependent on the operating system (such as *Microsoft Windows*), which is also dependent on layers of deeper processes that transform hardware signals into something comprehensible to the operating system. Each stage is one of structured, procedural, algorithmic invariable interpretation (Goldschlager & Lister, 1988; MacCormick, 2012). The result of this assemblage’s concurrent processing is a PDF, not as a fixed permanent object, but as a ‘performance’ (Chun, 2011b) that ceases the moment the processing ends. So long as all those processes operate in the way expected by all dependent processes, the PDF will display in the same way every time, and as intended by the producer. However, the movement of signal from one stage to the next is not just communicative, but transformative, the nature of that transformation reliant upon the design of that software process. Many academics who use PowerPoint may have experienced an instance of the file not appearing as expected, usually when the presentation file was made in a version of PowerPoint that differed from the version trying to interpret it. Even minor changes to the code that comprises the application ‘PowerPoint’ can lead to significant differences of software interpretation, and thus appearance when the file is opened.

At the level of persisting signal there is no differentiation between text, or image, or PowerPoint. It is only when that signal is interpreted by software that expects a specific arrangement of signal that it becomes differentiated as text or image. This is the nature of file 'formats', the predefined standards that establish what each part of the signal 'means to the interpreting application software. These constraints are often not noticed as file formats 'often take on a sheen of ontology when they are more precisely the product of contingency' (Sterne, 2012, p. 8). Yet these constraints are productive in determining not only what is a valid signal, but in establishing the rules for how it should be interpreted.

These rules are not just inward facing, but can be contextually aware, able to draw inferences about the user, as well as the hardware and software environment it is operating in. For example, websites can alter themselves depending on who they believe is viewing them based on locational data, cookies that have tracked their browsing habits and even the model of machine used to browse with (Angwin, Parris, & Mattu, 2016). Even the dimensions of the screen can play a pivotal role in the representation of the object as many websites now implement 'responsive design' in their HTML code, altering the size and position of page elements, or even adding or subtracting them depending on the amount of screen space available.

Digital documents then are the confluence of correctly ordered signal, layers of transformative software operations, a rigid structure of expectation codified as a file format and contextualising data. As both simultaneously machine data and expressed data (Levy, 2001), the digital form does not mould to the requirements of the document to be represented, the document must adhere to the limitations of the digital. Rather than offering us a somehow 'purer' form of information, the digital document is probably the most mediated form of communication we have ever produced.

As co-author

Software not only mediates as we consume, interpret and generate information with digital documents, but it is also important to recognise that software plays a key role in structuring the authorship of these objects as well. The materiality of a medium and the techniques of expression it affords will in turn place limitations on the kinds of expression, and therefore kinds of information generation it facilitates. Gitelman (2014) illustrates this process well in her discussion of 'blank' forms within paper-based bureaucracy. Pre-printed blank forms, present in any bureaucratic structure, establish a normative expectation of the kinds of information that are entered upon them, through limiting the range of responses or operationalising phenomena in a way that imposes a worldview upon it. Author agency, as Ben Kafka puts it, gets 'refracted through the medium' of paperwork (Kafka, 2012, p. 111). The interpretation of that document must account for both the intent of the author completing the blank, but also the intent of the author of the blank.

Today every digital document begins as a pre-structured blank, offering 'interaction' with pre-existing mechanisms, but denying 'authorship' of the mechanisms themselves (Murray, 1998). Some blanks seem familiar such as the 'template', the input box for a social media post and the pre-existing structure for a blog. However, the blank is also the ruleset that structures your response, the drop-down list of pre-written responses, the limitation of expressing a response through a thumbs up or thumbs down, and the hard-coded expectations of what is and is not valid input for your personal profile. This blank may produce a document that includes not just your response, but data about your response, embedding contextual information such as date, time, location and linkages to other documents or users deemed relevant at the time, producing a digital object that carries a rich cache of machine authored data (Bruns, 2012).

The blank is not just the pre-existing spaces in which we provide 'content', it is also the range of authoring techniques our tools make available to us. The use of word processing software is the use of a framework of predetermined options that seeks to totalise an understanding of language that will pre-empt the archetypal user (Fuller, 2003, p. 149). For Manovich (2013), software's design and affordances shape the kinds of interactions we can have in terms of expressing a message. What we once identified as the properties of the medium (e.g., television) are now the properties of a category of software (e.g., video streaming) situated within other categories of software (e.g., operating systems, device drivers, and codecs). By changing the properties of either the medium-specific software, or the general software architecture that supports it, we change the medium itself, the kinds of techniques we can deploy and the kind of documents that can be authored (Manovich, 2013). This range of techniques will both afford an expression, but also indicate conventions of expression, either through soft power facilitation (such as the provision of a letter template) or unavoidable restrictions (140-character limit). These selective facilitations of use afford us both new possibilities and new limitations. Whether we are shifting professional practices in response to hard character limits on Twitter (Mirer & Bode, 2015) or being threatened with cultural exclusion due to a lack of language encoding (Junker & Luchian, 2007), the author of any digital document is an entanglement of a human author and their software tools.

Mediating persistence

Persistence over time is intrinsic to a document's use as evidence, and in ensuring access for its future use. Equally, persistence is also dependent on access, as Otlet and Briet understood in their drive to develop effective information retrieval systems, a document that persists but that has no access may as well not persist at all. The persistence of digital documents is dependent upon an entanglement of their ontological reality as software-interpreted electrical signal, and the socially shaped priorities of the archival systems in which they reside.

Digital documents present a fundamental challenge to their persistence through the realities of their material construction. Whilst legacy archives must work to ensure that their documents do not degrade, the very existence of a digital document must be constantly re-asserted. As Chun (2011a) argues, we conflate the 'memory' of computational systems with the idea of 'storage', implying that once created, a digital document will exist within hardware until it is ready to be retrieved. However, like our own memories, digital documents must be constantly refreshed to persist. Short-term memory modules in computers quickly lose their data the moment electrical current is lost. Long-term memory such as hard-disk drives can retain signals without constant current, but not indefinitely. The signals only persist because they are audited, reorganised and rewritten by the host computer (Allen-Robertson, 2015). These signals are 'enduringly ephemeral' (Chun, 2011a), constantly refreshed yet still degenerating. Furthermore, their reliance upon assemblages of software for interpretation means that software must also persist to guarantee it can be accurately performed. Digital Humanities scholars, in particular, have grappled with the issue of stabilising the ephemeral born-digital resource, a process which itself has illuminated the ephemerality of non-digital texts in need of preservation (Nowviskie, 2015). Various strategies to manage this issue have been proposed, such as providing 'emulation' of the original software environment, or through migrating the documents to newer currently supported document types (Anderson, 2015). Yet both strategies require significant labour to initialise and to maintain as the webs of interdependent software packages and formats shift and change over time.

The information systems that support our digital documents are myriad and dispersed, with a vast range in terms of their visibility and purpose. The publicly accessible internet is our most prominent 'archive of archives' (Gane & Beer, 2008). Traditional 'legacy' (Prelinger, 2016) archival institutions have long concerned themselves with this responsibility of ensuring documents persist through materially aware practices and technologies of preservation. Ensuring a document persists is key for their larger roles of providing access, and maintaining the 'evidence' that is foundational to the secondary knowledge generated from them. However, the internet is not, according to Ernst and Parikka (2013), an archive in the traditional sense of legacy archives. Rather than a 'static accumulation of dossiers' contemporary digital archives are a 'dynamic connection of documents and links' (Ernst & Parikka, 2013, p. 84). Digital archives are broadly indifferent to the content held. Rather than aspiring to ensure the persistence of cultural memory with access to documents already within the archive, emphasis is placed on the newest and latest documents, dynamically restructuring the archive to prioritise the now, with little regard for what has already passed (Prelinger, 2016).

The slow, thoughtful legacy archival concerns of persistence are fast being outpaced by a focus on 'relevance', a priority that emphasises the present rather than the past (Prelinger, 2016). The persistence of digital objects in general is subject to a range of socially situated constraints and contingencies, such as external privacy and security concerns, financial sustainability, continued provision of a knowledgeable and skilled workforce, and political change and instability (Adu & Ngulube, 2016). For private archival spaces, such as YouTube, Twitter and Facebook, there is an even greater range of influences. These privately owned spaces are archival in much of their function, storing, ordering and making accessible vast amounts of digital documentation, operating as keepers of contemporary everyday culture. Yet their identification as 'platforms' indicate the multitude of roles and priorities they maintain. YouTube must balance users, policymakers, advertisers and clients as they both exhibit a public facing front-stage, and a profit-seeking 'backstage' (Gillespie, 2010). The term 'platform' does crucial discursive work in framing the site as a politically neutral surface, rather than recognise the many competing interests in the kinds of content that persists and is accessible on the site. Even if the documents persist within internet archives, our ability to locate and use them can be highly mediated by the systems themselves. Google's search algorithms return results to us based on an ever-fluctuating algorithmically determined measurement of relevance (Introna & Nissenbaum, 2000), playing a role in precisely which documents are considered relevant. Furthermore, contemporary search engines also know the users issuing the queries, making relevance a highly fragmented and recursive measure as past searches are folded into future responses (Day, 2014). The result is that the search platform becomes a collaborator in document discovery, delivering documents based on a mix of the host's political and economic interests, and a shallow understanding of what it believes the researcher wants to receive.

Chun's argument of the enduringly ephemeral remains salient as the digital document is maintained only through constant attention, both technically and socially. Furthermore, these strategies must also account for the increasing likelihood that a digital document may be dependent upon, or solely be accessible via private servers outside the jurisdiction of archivists or researchers. We often presume, with data surveillance so dominant a theme in our contemporary society, that all these data being generated are being archived by somebody as a matter of course. Yet archival spaces that emphasise relevance, the contemporary, and access over persistence and provenance have no onus to retain any documents that outlive their host's own definition of value (Prelinger, 2016).

Conclusion

Digital documents, by their prominence in the everyday lives of people, present themselves as the future of documentary research. They present significant opportunities for a 'data dense' understanding of contemporary and past social issues as it seems each movement in the social world has the potential to generate all kinds of data from a multitude of perspectives and positions. Yet documentary researchers in sociology and elsewhere need to be aware that these documents are not simply a re-inscription of the paper documents that documentary research was built upon. The critical assessment of documentary sources must also recognise the material as a contributory factor in the interpretation of meaning particularly in the case of digital documents. Digital documents, far from the rhetoric of pure information, are highly mediated objects with a materiality that plays a significant, if often unseen contributory role in the interpretative process. Their materiality as inaccessible electrical signal necessitates layers of software interpretation that is varied and changeable based on a multitude of internal and contextual factors. The authorship of these documents is also highly mediated by the conventions and affordances designed into the tools used to produce them. In these mediating processes, software becomes the unseen co-author and co-interpreter of digital documentary sources. The material realities of our digital documents have also introduced a significant range of challenges to their persistence and with it a challenge to their suitability as evidence. Their persistence requires significant labour and resource, both in the reaffirmation of their existence through the rewriting of their electrical signal, but also in their persistence within the private archives that maintain our access, and direct our attention, to them. Though digital, there are layers of necessary, very solid, infrastructure that must be attended to if the digital will persist (Hu, 2015).

These influences are best illuminated in the efforts to digitise that are able to demonstrate a shift in interpretative potential (Levy, 2001; McGregor, 2014), and when expectations built into software design may exclude forms of expression (Junker & Luchian, 2007) or encourage particular types of authorship (Mirer & Bode, 2015). When it comes to the digitally native document, our awareness of co-authorship and interpretation lessens as we lack a point of comparison. The precise details of how contingent our software entangled practices are becoming, are difficult to identify. Yet we must maintain a critical awareness of how our practices, or even the way we think (Hayles, 2012) may be being influenced by the shift in predominant document form. This awareness should be fostered in a period where techniques of text decomposition and natural language processing subject our texts to intense methods of software reinterpretation, and inscrutable processes of machine learning guide our interpretation (Rieder, 2017). Whilst these are more drastic developments of software interpretation, we should not forget the subtler, mundane software interpretivity of the everyday that perform and produce everyday documentation. Just as Otlet and Briet recognised, the relationship between knowledge, documents and the systems that manage them, we must not fall into the trope of believing the digital to be purer, clearer information, but recognise its intrinsic nature as mediated and contingent, ephemeral and opaque in its authorship. Though not possible here, further development of practical strategies for dealing with these issues, or even experimental research that might evaluate the impact of differing media form in user interpretation, requires much greater attention, and I intend to contribute to this discussion in a future paper.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Economic and Social Research Council [grant number ES/M010236/1].

Notes on contributor

James Allen-Robertson is Lecturer in Media and Communication and Digital Sociologist in the Department of Sociology at the University of Essex. His early book *Digital culture industry* (2013, Palgrave Macmillan) charted the role of piracy in the development of digital media distribution. His more recent work explores issues of power and agency in the relationship between humans and their media machines, and seeks to develop new algorithmic methods for social science research [email: jallenh@essex.ac.uk].

ORCID

James Allen-Robertson http://orcid.org/0000-0002-1668-8140

References

Adu, K. K., & Ngulube, P. (2016). Key threats and challenges to the preservation of digital records of public institutions in Ghana. *Information, Communication & Society*, *20*(8), 1127–1145. doi:10.1080/1369118X.2016.1218527

Allen-Robertson, J. (2015). The materiality of digital media: The hard disk drive, phonograph, magnetic tape and optical media in technical close-up. *New Media & Society*, *19*(3), 455–470.

Alvesson, M., & Sköldberg, K. (2010). *Reflexive methodology*. London: SAGE.

Anderson, D. (2015). The digital dark age. *Communications of the ACM*, *58*(12), 20–23.

Angwin, J., Parris, T., & Mattu, S. (2016). When algorithms decide what you pay. ProPublica.org. Retrieved from https://www.propublica.org/article/breaking-the-black-box-when-algorithms-decide-what-you-pay.

Berry, D. (2012). *Understanding the digital humanities*. Basingstoke, UK: Palgrave Macmillan.

Bijker, W., Hughes, T., & Pinch, T. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. Cambridge, MA: MIT Press.

Briet, S., Day, R. E., Martinet, L., & Anghelescu, H. G. B. (1951/2006). *What is documentation? English translation of the classic French text*. Lanham, MD: Scarecrow Press.

Brown, J. S., & Duguid, P. (2000). *The social life of information*. Boston, MA: Harvard Business School Press.

Bruns, A. (2012). How long is a Tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information, Communication & Society*, *15*(9), 1323–1351.

Buckland, M. (1998). *What is a 'Digital Document'?* Retrieved from http://people.ischool.berkeley.edu/~buckland/digdoc.html

Carlile, P. R., Nicolini, D., Langley, A., & Tsoukas, H. (Eds.). (2013). *How matter matters: Objects, artifacts, and materiality in organizational studies*. Oxford: Oxford University Press.

Chun, W. H. K. (2011a). The enduring ephemeral, or the future is a memory. In E. Huhtamo & J. Parikka (Eds.), *Media archaeology* (pp. 184–203). Berkeley: University of California Press.

Chun, W. H. K. (2011b). *Programmed visions: Software and memory*. Cambridge, MA: MIT Press.

Coffey, A. (2014). Analysing documents. In U. Flick (Ed.), *Qualitative data analysis* (pp. 367–379). London: SAGE.

Day, R. E. (2014). *Indexing it all: The subject in the age of documentation, information, and data*. Cambridge, MA: MIT Press.

Dourish, P., & Mazmanian, M. (2013). Media as material: Information representations as material foundations for organizational practice. In P. R. Carlile, D. Nicolini, & A. Langley (Eds.), *How matter matters: Objects, artifacts, and materiality in organizational studies* (pp. 92–116). Oxford: Oxford University Press.

Ducheyne, S. (2009). 'To treat of the world': Paul Otlet's ontology and epistemology and the circle of knowledge. *Journal of Documentation*, *65*(2), 223–244.

Ernst, W., & Parikka, J. (2013). *Digital memory and the archive*. Minneapolis: University of Minnesota Press.

Frohmann, B. (2009). Revisiting 'what is a document?' *Journal of Documentation*, *65*(2), 291–303.

Fuller, M. (2003). *Behind the blip: Essays on the culture of software*. New York, NY: Autonomedia.

Gane, N., & Beer, D. (2008). *New media*. Oxford: Berg.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364.

Gitelman, L. (2014). *Paper knowledge: Toward a media history of documents*. Durham, NC: Duke University Press.

Gitelman, L. (2016). Raw Data' is an oxymoron. In W. H. K. Chun, A. W. Fisher, & T. W. Keenan (Eds.), *New media old media* (pp. 167–176). London: Routledge.

Goldschlager, L., & Lister, A. (1988). *Computer science: A modern introduction*. London: Prentice Hall.

Hayles, N. K. (2012). *How we think: Digital media and contemporary technogenesis*. Chicago: University of Chicago Press.

Hine, C. (2000). *Virtual ethnography*. London: SAGE.

Hu, T.-H. (2015). *A prehistory of the cloud*. Cambridge, MA: The MIT Press.

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology*, *35*(2), 441–456.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The Information Society*, *16*(3), 169–185.

Jones, S. (Ed.). (1999). *Doing internet research*. London: SAGE.

Junker, M.-O., & Luchian, R. (2007). Developing web databases for aboriginal language preservation. *Literary and Linguistic Computing*, *22*(2), 187–206.

Kafka, B. (2012). *The demon of writing: Powers and failures of paperwork*. New York, NY: Zone.

Kazmer, M. M., & Xie, B. (2008). Qualitative interviewing in internet studies: Playing with the media, playing with the method. *Information, Communication & Society*, *11*(2), 257–278.

Kirschenbaum, M. G. (2008). *Mechanisms: New media and the forensic imagination*. Cambridge, MA: MIT Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford: Oxford University Press.

Levy, D. M. (2001). *Scrolling forward: Making sense of documents in the digital age*. New York, NY: Arcade.

Lievrouw, L. A. (2014). Materiality and media in communication and technology studies: An unfinished project. In T. Gillespie, P. J. Boczkowski, & K. Foot (Eds.), *Media technologies* (pp. 21–51). Cambridge, MA: MIT Press.

Lund, N. W. (2009). Document theory. *Annual Review of Information Science and Technology*, *43* (1), 1–55.

MacCormick, J. (2012). *9 Algorithms that changed the future*. Princeton, NJ: Princeton University Press.

Manovich, L. (2013). *Software takes command: Extending the language of new media*. London: Bloomsbury.

Marres, N. (2017). *Digital sociology*. Cambridge: Polity.

McGregor, H. (2014). Remediation as reading: Digitising the western home monthly. *Archives and Manuscripts*, *42*(3), 248–257.

McKenzie, D. F. (1999). *Bibliography and the sociology of texts*. Cambridge: Cambridge University Press.

Mirer, M. L., & Bode, L. (2015). Tweeting in defeat: How candidates concede and claim victory in 140 characters. *New Media & Society*, *17*(3), 453–469.

Murray, J. (1998). *Hamlet on the Holodeck: The future of narrative in cyberspace*. Cambridge, MA: MIT Press.

Nowviskie, B. (2015). Digital humanities in the anthropocene. *Digital Scholarship in the Humanities*, *30*(Suppl_1), i4–i15.

Otlet, P. (1934). *Traité de documentation*. Brussels: Editiones Mundaneum.

Otlet, P. (1990). *International organisation and dissemination of knowledge: Selected essays of Paul Otlet* (W. B. Rayward, Trans.). Amsterdam: Elsevier.

Palys, T., & Atchison, C. (2012). Qualitative research in the digital era: Obstacles and opportunities. *International Journal of Qualitative Methods*, *11*(4), 352–367.

Platt, J. (1981). Evidence and proof in documentary research 1: Some specific problems of documentary research. *The Sociological Review*, *29*(1), 31–52.

Plummer, K. (2001). *Documents of life 2: An invitation to a critical humanism*. London: SAGE.

Prelinger, R. (2016). The disappearance of archives. In W. H. K. Chun, A. W. Fisher, & T. W. Keenan (Eds.), *New media old media* (pp. 199–204). London: Routledge.

Prior, L. (2008). Repositioning documents in social research. *Sociology*, *42*(5), 821–836.

Rekrut, A. (2014). Matters of substance: Materiality and meaning in historical records and their digital images. *Archives and Manuscripts*, *42*(3), 238–247.

Rieder, B. (2017). Scrutinizing an algorithmic technique: The Bayes classifier as interested reading of reality. *Information, Communication & Society*, *20*(1), 100–117.

Sassoon, J. (2007). Photographic meaning in the age of digital reproduction. *Archives & Social Studies*, *1*, 299–319.

Scott, J. (1990). *A matter of record: Documentary sources in social research*. Cambridge: Polity Press.

Seymour, W. S. (2001). In the flesh or online? Exploring qualitative research methodologies. *Qualitative Research*, *1*(2), 147–168.

Sterne, J. (2012). *MP3: The meaning of a format*. Durham, NC: Duke University Press.

Sterne, J. (2014). What do we want?' 'Materiality!'. In T. Gillespie, P. J. Boczkowski, & K. Foot (Eds.), *Media technologies* (pp. 119–128). Cambridge, MA: MIT Press.

Stewart, S. (1984). *On longing: Narratives of the miniature, the gigantic, the souvenir, the collection*. Baltimore, MD: Johns Hopkins University Press.

Webb, S., & Webb, B. (1932). *Methods of social study*. London: Longmans.

Yates, J. (1989). *Control through communication: The rise of system in American management*. Baltimore, MD: Johns Hopkins University Press.

Yeo, G. (2010). 'Nothing is the same as something else': Significant properties and notions of identity and originality. *Archival Science*, *10*(2), 85–116.