Using digital sources

Nix, Adam; Decker, Stephanie

DOI:
10.1080/00076791.2021.1909572

Document Version
Peer reviewed version

Citation for published version (Harvard):
Nix, A & Decker, S 2021, 'Using digital sources: the future of business history?', *Business History*. https://doi.org/10.1080/00076791.2021.1909572

Using digital sources: The future of business history?

As historians start researching the late twentieth century, they are increasingly finding traces of the past created digitally. At the same time, use of computers to digitise analogue material means that many pre-digital sources have been reproduced digitally. As such, future historical research will increasingly include digital forms of evidence and computer-based research tools. This paper explores how such resources might be used within business history, bridging the gap to digital history, and reflecting upon their methodological implications. We present a framework for distinguishing between sources, elaborating their differing digital characteristics and historical authenticity. We then draw on our own use of digital company records and media archives to outline two different ways digital sources can be interrogated by business historians. We argue that digital sources afford unique insights and new opportunities for historical knowledge production, but to access them, business historians will likely adapt aspects of their future practice.

Keywords: Digital history; digital humanities; digitisation; born-digital; reborn-digital

Introduction

Business historians have a long tradition of reflecting on the implications of modern organisational record-keeping and communication practice, doing so even before the advent of ubiquitous personal computing and the internet (Harvey & Jones, 1990; Turner, 1978). However, now more than ever, the digital nature of more recent organisational existence is intersecting the historiographic research of organisations (Kirsch, 2009; Moss, 2009). Furthermore, archives, libraries and publishers are gradually digitising their collections, adding as they do, to an increasingly *digital* history. While there is an established and lively discourse

on this 'digital shift' within these professions (Corrado & Sandy, 2017; DCDC, 2019; Prom, 2016), the far-reaching implications (and increasing engagement) with digital historical research also raises increasingly important questions for business historians. To date, however, the use of digital sources is an overlooked methodological issue within business history, despite the encouraging move towards greater reflexivity more generally (Decker et al., n.d., 2015; Maclean et al., 2017; Rowlinson et al., 2014).

By digital sources, we refer to any historical materials that contain digital elements, whether as a circumstance of their original creation or retrospective alteration. Under this broad classification, digital sources share the same underlying historiographic basis as their analogue equivalents, constituting partial traces of an ontologically inaccessible past (Lipartito, 2014; Mills et al., 2014; Trouillot, 1995). Nonetheless, intangibility, editability and easy duplication are just a few of the characteristics that distinguish digital from non-digital sources, and make working with them a unique form of historical research (Brügger, 2012). That these sources offer new and valuable affordances is something that digital historians have long promoted, alongside the various web and computer-based research tools that facilitate engagement with them (Ayers, 2001; Cohen et al., 2008; Duranti, 2001; Rosenzweig, 2003). However, these developments are yet to be widely recognised within business history, even if the potential for new insight has already been demonstrated, for instance, by work leveraging digital archives (Vallejo Pousada & Larrinaga, 2020), government databases (Benke, 2018), and computer-assisted analysis (Tumbe, 2019).

Because the tools and approaches designed to leverage digital sources often address theoretical and methodological concerns of other disciplines, the analytical opportunities they offer to historical researchers are not always obvious. Topic modelling and corpus linguistics, for example, can provide an overview of the most commonly used topics or words in a collection of digital texts (Flaounas et al., 2013; Graham et al., 2012), and specific tools like

Google's *Ngram Viewer* provides an accessible way to statistically explore historical language trends (Lansdall-Welfare & Cristianini, 2020). However, such tools provide limited insight into the key questions that underpin critical source analysis: Who wrote it? What is the text about? Why was it written down (and preserved)? When and where was it created? Despite this, we maintain that digital sources and some of the tools developed to interrogate large amounts of digital text can be immensely useful to business historians, but that this requires new and adapted methodological practices from historians. These are only rarely discussed, not widely known, not present in research training and their integration into historical research is still developing as more and more digital archives become available.

In this paper, we explore the methodological aspects of using digital data as historical sources, starting with an elaboration of key ideas from digital history (Brügger, 2012, 2018; Cohen et al., 2008; Rosenzweig, 2003; Sternfeld, 2011). We believe more conceptual and linguistic clarity is needed if we are to use digital sources effectively, and this framework provides business historians a basis for such methodological reflexivity (Schwarzkopf, 2012; Sternfeld, 2011). Next, we elaborate two examples of research using digital sources, focusing first on digital records from Enron (emails and telephone transcripts), which we analysed with critical source analysis (Kipping et al., 2014). Because of the nature of their production, preservation and access, these sources required careful methodological consideration in ways that went beyond classical source analysis. We then elaborate upon our experience searching digital media archives by drawing on a computer-aided, linguistic analysis as an alternative approach to classic historical research. In doing so, we demonstrate that, by complementing our methods with those familiar to disciplines such as linguistics or computer science, digital source can afford new and different insights into historical texts (Tumbe, 2019). We close by reflecting on the many new opportunities and choices that exist for digital historical research with the caution that these new research practices will be selective in their consumption and

representation of the underlying historical records. Thus they require a greater methodological awareness than more familiar paper-based archives (Decker, 2013; Kipping et al., 2014; Lipartito, 2014).

Whilst an exhaustive summary of the many analytical possibilities is beyond the scope of this article, we seek to open up a wider debate as to which tools and approaches offer relevant methodological pathways as business historical research enters the digital era. With this, we contribute a better methodological understanding, not just of key concepts in digital humanities that are relevant to business historians, but also business history-specific insights into the different ways in which we can research digital sources, and how existing tools and approaches can help us answer questions important to the field.

An increasingly digital history

Among historians, the field of *digital history* has been the primary nexus for the use of digital sources and technologies to engage with the past. Much of the early dialogue here concerned an evolution beyond the field's pre-digital norms, promoting the digital preservation, creation and dissemination of historical knowledge (Ayers, 2001; Cohen et al., 2008; Dougherty & Nawrotzki, 2013). Along these lines, digital history has been defined as "an approach to examining and representing the past that works with the new communication technologies of the computer, the internet network, and software systems." (Thomas in Cohen et al., 2008:454). As this suggests, this initial wave of interest was at least partly a reaction to the mainstream emergence of web-enabled personal computing and an exploration of its potential usefulness. Beyond professional historians, digital history has also been promoted for its potential democratising effect, facilitating a break from the norm of limited access archives and peer-review, and encouraging greater public inclusion and engagement (Bolick, 2006; Seefeldt & Thomas, 2009). Thus, digital history has emerged as a distinct community of practice, teaching

and researching through an alternative model for historical scholarship (Cohen & Rosenzweig, 2006).

Beyond 'digital history', historical research in general has experienced a digital transformation, which is perhaps most noticeable in the collections of digitalised material that are now an established means of engaging with the past (Putnam, 2016; Schwarzkopf, 2012; Sternfeld, 2011). Such trends – themselves a product of society's wider digital transition – go beyond sources and affect the entire practice of history (Fellman & Popp, 2013; Norton & Donnelly, 2018). Indeed, at a basic level, the incorporation of computers (and the internet) into the historian's 'methodological toolbox' is now widespread (Brügger, 2012), variously aiding research, writing and reviewing activities. For instance, they have been seen as a solution to greater transparency, with Smith & Umemura (2019) arguing that business historians should routinely make digital images of their sources available, in the interest of effective criticism. Beyond this, communication techonolgies also affect how research is resourced, as seen during the recent Coronavirus lockdown, when a crowdsourcing request prompted thousands to help remotely digitise handwritten rainfall records dating back to the 1820s (Amos, 2020; Dunn & Hedges, 2016). The use of advanced computing is also no longer a fictional notion, with technologies like machine learning and blockchain increasingly employed to organise and protect the integrity of digital collections (Bui et al., 2019; Spencer, 2017; The National Archives (UK), 2017). While many such initiatives remain the exception within historical research, they are suggestive of the systemic influence digital technologies are having within our field.

Though these broader considerations of a digital history represent important issues, we necessarily limit our attention to digital sources as a discrete aspect of digital historical research. Digital sources include any trace of the past that exists in a digital form, whether because of its original creation or the retrospective efforts of archivists or other interested

parties. Implicit within digital history is an engagement with such primary and secondary historical materials (Cohen et al., 2008). While this might include digital historical *representations* of analogue sources (Sternfeld, 2011), histories of the late twentieth century and beyond will progressively use traces of the past that have only ever existed in a digital format (Rosenzweig, 2003). As with the outputs of pre-digital innovations like the printing press, typewriter and fax machine, these sources are products of underlying technologies, the conditions and affordances of which shape their nature (Coopersmith, 2015; Fayard & Weeks, 2007; Thompson, 2017). In this way communication genres like email have completely changed the culture of personal and organisational correspondence (Byun & Kirsch, 2020; Moss, 2009; Yates, 2005). Similarly, the advent of social media has changed how social actors engage with organisations, generating potentially valuable but unique digital traces in the process (Bressers & Hume, 2012; Laurell et al., 2019; Onaga & Shell, 2016).

Books, newspapers and other typically secondary sources are also an important aspect of the digital humanities, and mass-digitisation projects of the Hathi Trust, Google (Books), and other institutions are regularly cited for their transformative influence (Blevins, 2019; Mullen, 2014). Indeed, libraries and publishers have been among the most engaged communities in relation to the digital transition, and the process of searching and reading published material is now an inherently digital one (Abbott, 2014; Nicholas & Clark, 2015). As with all historical evidence, the primacy of a source depends upon the nature of the questions asked of it, and not any intrinsic variance (Lipartito, 2014). Nonetheless, contemporaneous and incidental fragments of normal life – the sources more commonly primary to historical research questions – are subject to far greater variability in terms of production, preservation and accessibility. However, intrinsically digital sources like emails and websites are less prominent within the digital history literature, and relevant

historiographic consideration is largely the product of the sub-field, web history (Brügger & Milligan, 2018; Milligan, 2019).

Distinguishing digital sources

The plural and diverse nature of digital sources significantly complicates their effective articulation as a methodological issue, particularly among historians yet to engage significantly with digital historical research. In this regard, there is currently a gap between business history and fields like digital history, where the difference between digital materials is widely understood. This notwithstanding, business historians have moved noticeably towards greater methodological reflexivity and explication of their practice in recent years, and as digital sources become more prominent, they become increasing relevant to this trend. To facilitate an integration of the digital past into this discourse, we set out a framework for historical materials' potential or actual digital existence, which introduces a conceptual and linguistic basis for understanding different types of digital source. For business historians, this will aid more critical and effective usage of digital sources, facilitating greater reflection on the nature of a source's digital existence, and highlight its particular affordances and limitations.

In elaborating on the different types of digital sources (see Figure 1), web history provides the useful notion of *digitality*, which captures a distinction between sources that are *digitised, born-digital* or *reborn-digital* in nature (Brügger, 2012, 2018). While these categories have been described in relation to websites and web archives, their interrelationship and significance to sources relevant to business history requires further consideration. In expanding on these concepts, we consider digitality in combination with *format authenticity,* which considers whether a source has been edited to take a different (digital) form and is key to appreciating the selection decisions underlying access to digital material (Schwarzkopf, 2013). Thus, digitality explains whether the digital elements of an original source are intrinsic (e.g., a webpage) or optional (e.g., a digitised handwritten letter), and format authenticity

presents a basis for understanding the implications that changes in format have on the use and accessibility of such historical material.

[Figure 1]

Analogue sources are entirely non-digital in nature; however, they can become digital, albeit in a reproduced form. This occurs through a process of digitisation, whereby a digital copy of an analogue document is created (often photographically) and stored electronically. Such sources are already a common component of established archives such as the Hagley Museum & Library in Delaware, US, or the National Archive (UK), and in their most simple form may just constitute digital scans, which allow for remote and multi-user access. More sophisticated digitisation offers still greater digital functionality. For example, with optical character recognition (OCR) software, a handwritten letter can be reproduced so that a computer-readable, digital version is available within an online repository (Cassell & Symon, 2004; Gollins & Bayne, 2015). Such processes make pre-digital sources accessible remotely by multiple simultaneous users and allow contents to be searched computationally (Putnam, 2016). Historians can use this functionality to search and select from a large collection of sources in a manner equivalent to searching a library database for literature (Moss, 2015; Putnam, 2016). This moves away from traditional archival finding aids, containing information *about* sources, towards a process of searching directly *within* them (Nicholson, 2013).

Because the digitisation process is fairly uniform, virtually any analogue source can become digital, whether audio-visual, handwritten, or typed. While the resulting digitised sources can only ever be edited representations of historical material, not authentic sources in-and-of-themselves, their easy access and usability has made them a popular option for research (Laurell et al., 2019). However, digitisation is relatively resource intensive, making archival discretion a necessary factor in deciding which sources should be reproduced. In this way, digitisation represents a further level of selection, building on the initial survival of an analogue

original and the subsequent decision to formally preserve it (Schwarzkopf, 2012). While such judgements undoubtedly consider historical accuracy, the easy access and flexibility they provide to 'important' sources is by no means an unambiguous consideration. Archival decisions are subject to political and cultural conditions (Donnelly & Norton, 2012), which are relevant in appraising the representativeness of digitisation as a discretionary source format. Moreover, given well-resourced (often Western) institutions have a far greater capacity for digitisation, there is a risk of compounding archival and historical bias embedded in our current practice (Breckenridge, 2014; Cummings et al., 2017) .

The term *born-digital* is widely used throughout the digital humanities to describe those sources that have only ever existed digitally, having been created in the post-analogue past and preserved within their original digital format (e.g., Boss & Broussard, 2017; Brügger & Milligan, 2018; Thomas, 2016). Amongst the many possible examples, emails, web-based media, and word-processed text all represent born-digital material. In addition to their intrinsic searchability, born-digital sources contain characteristics such as editability, interactivity and hypertextuality, that are material to their original character (Brügger, 2012). For instance, websites are not static artefacts, but rather continually changing bodies of information, being adapted in line with the motivations of those who curate them (Milligan, 2019). Born-digital sources can also exist as part of a complex interconnected corpus, as seen with email or social media, where a single message sits within a highly structured network of wider intercommunication (Jaillant, 2019; Prom et al., 2018). While preservation of born-digital sources is vital, some unprocessed digital sources present significant issues to privacy, security, and practical usability. Consequently, access to authentic born-digital sources for research is problematic, and it is normal for archives to preserve most of their digital sources in closed or 'dark' archives, only releasing a fraction of their collection through graduated or controlled access (Kirsch, 2009; The National Archives (UK), 2017).

Reborn-digital sources address some of these issues, being altered from born-digital originals to aid preservation, accessibility or analysis (Brügger, 2018). Like the digitisation of analogue sources, they are therefore representations of authentic material that has been subject to additional selection and processing. For instance, a computer file might be converted to mitigate format obsolescence (the non-continuation of contingent software or hardware), or a piece of digital media incorporated into web-based database. During this process, appearance or functionality may be altered, removed or presented differently to their original counterpart. At a basic level, a PDF of a spreadsheet can serve as a basic record of digital content, but it would not capture its original functionally (e.g., reordering columns). However, many reborn-digital sources do retain good deal of a source's intrinsic digital character, as seen with the *AvocadoIT Collection*, a reborn-digital dataset of emails from a failed dot-com company (Oard et al., 2015). Created from back-up discs, the new dataset addressed several security and privacy issues and presented the contents in a readily analysable format (structured text files rather than email archive files). Accordingly, while reborn-digital sources lack the unedited format authenticity of the originals, they provide a practical route to greater accessibility in a way that retains many original born-digital characteristics.

Given that the alteration of analogue and born-digital sources is increasingly normal, reflection on format authenticity allows for greater methodological transparency around related issues of selection and access. Currently, the digitisation of analogue sources is largely treated as a methodologically uncomplicated route to enhanced access, requiring little critique or reflection. However, this process adds an additional layer to the selective judgements archivists make and requires consideration, despite the benefits it presents (Schwarzkopf, 2012). Furthermore, the improved searchability and other computational functionality gained, fundamentally alters the way historians interact with such previously analogue material, changing how they are interrogated and interpreted (Lansdall-Welfare & Cristianini, 2020;

Putnam, 2016). For intrinsically digital material, drawing a distinction between born and reborn digital makes explicit any retrospective influence on a source's authenticity and provides a basis for greater appreciation of the process underlying its preservation. Moreover, business historians will increasingly work with representations of sources rather than authentic originals, perhaps to the point where they become the default for historical enquiry into the digital era. Accordingly, the use of digital sources is likely to have methodological implications for the histories of both the pre and post-digital past.

The next sections show how we worked with different digital representations of original sources materials in our research projects (see Table 1). In the first of these, we utilised reborn-digital communication records collected from the US energy company Enron and published by the Federal Energy Regulatory Commission (FERC) as part of a fraud investigation. Here, email correspondence and telephone recordings were originally created digitally, but altered before their public release. Such alternations include redaction, transcription of audio files, or in some cases, re-digitisation of printed documents. In the second project, journalistic interviews from the turn of the twenty-first century were used to explore entrepreneurial narratives. These sources were accessed through the library database *BusinessSource Complete*, and primarily contained articles from the print and digital magazine media. As such they constitute edited representation of authentic analogue and born-digital originals, aggregated, and formatted for conducive preservation and access. We used these projects to highlight two alternative approaches to working with digital sources, reflecting on their methodological significance as we do so. While these cases do not provide an exhaustive view of the possibilities or issues here, they offer contrasting empirical examples of how business history might actually be undertaken using them.

[Table 1]

Enron and the use of digital era company records

While there are several areas where digital sources have methodological implications for business history, the production, preservation and access to company (and related public) records is among the most significant. We first engaged with such sources while researching the US company Enron and its involvement in the manipulation of California's deregulated energy markets. Sources relevant to our interests were published by the FERC's e-library, an online public records information system providing access to items filed as part of federal energy proceedings (FERC, 2014). Drawing on our use of these materials we show that digital sources like telephone and email records can provide new insight into parts of organisational life seldom recorded. However, we also highlight that working with such records presents particular challenges (notably the quantity of information) and these require us to consider new analytical processes. Finally, we show the value of automated metadata (information *about* sources) as means to contextualising related digital sources and assessing their validity and authenticity.

New insight into organisational pasts

Company archives tend to prioritise records concerning the executive functions within organisations, providing access to the correspondence of top managers or the minutes taken at high-level meetings (Decker, 2013). What they less typically preserve is ephemeral daily life, as seen through middle-management decisions or the activities of non-managers. As such, the routines, emotions and other interconnected elements of specific work practices can be hard for historians to untangle (cf. Lipartito, 2013). As we found though, digital interactivity affords a valuable insight into these parts of organisational existence. Indeed, our understanding of Enron was primarily obtained through traces of everyday communication between energy traders. While technologies like telecommunications do not typically leave preservable traces for historical enquiry, trading-floor conversations are commonly audio recorded for legal

purposes, as was the case within Enron. While these normally remain the property of the organisation (and are routinely destroyed after a short period), federal proceedings meant that Enron's recordings were transcribed, and these subsequently became the basis for our analysis (Crowley, 2005). These sources, with their immediacy to key events and incidental nature of production, provided a fragmented but highly primary record from which to observe the everyday realities of fraudulent trading practices (Megill et al., 2007; Rowlinson et al., 2014).

During our analysis of the telephone conversations, we regularly encountered traces of the past, too fleeting or informal to have survived as more traditional sources. Indeed, despite the serendipitous circumstances of their survival and their novelty as a historical source, they came to form the central evidence for our analysis, providing valuable and unexpected glimpses of corrupt organisational practice that would have otherwise been lacking. For instance, through silences created when traders moved between monitored and unmonitored platforms, we were able to see how routines of illicit practice were developed and maintained (Decker, 2013). Spoken dialogue also provides a rich view into the decision-making processes and sensemaking of actors that better represents the organic nature of events. While this is something oral history has long appreciated in its use of retrospective interviews (Keulen & Kroeze, 2012; Kroeze & Vervloet, 2019), the immediacy of telephone conversations takes this one step further, placing the researcher as a 'fly-on-the-wall' in the trading room as events unfolded. In this way, digital communications are particularly effective in preserving the interactive social content of normal organisational life.

Unlike telephone calls, many other everyday interactions of modern organisational life routinely leave a trace that survives the actual moment of interaction. Perhaps the most significant example of this is email, which has become a ubiquitous and integral part of organisational communication and can provide incredibly detailed insight into historical events. In our case, we accessed the *Enron Email Dataset* (EED), which had been processed to

remove some personal information (e.g., social security numbers) and all email attachments but was otherwise a complete copy of the original files taken from Enron.[1] The EED contains a set of 150 sub-folders, one for each of the original 'custodians'. While 150 employees out of a company of nearly 30,000 is far from exhaustive, the 126-gigabit collection nonetheless contains over 500,000 emails and is well beyond the scope of standard historical or qualitative analysis. Moreover, the abundance of emails unconnected to our research questions (or indeed the organisation) and a lack of any curation or categorising information made accessing potential insight from the corpus a significant challenge (Fellman & Popp, 2013). Accordingly, while email still had the capacity to provide valuable and novel glimpses, it also provides an excess of 'spam mail' that makes any such insights hard to obtain.

Making sense of digital company records

The management of abundant material is nothing new to historians (Decker, 2013; McNeill, 1986). Nonetheless - and as we found with the EED - the digital shift will inevitably lead to the preservation of exponentially more material. In our case, to allow for a traditional historical reading of the email, we navigated this challenge by creating a targeted sub-set of the overall dataset that was specific to our research focus. Firstly, we restricted the sample to 18 accounts specific to the trading division we were investigating, extracting those emails held within the 'sent-mail' folder. Our rationale here was that users are closer to the production of sent mail, actively authoring emails and replies. Additionally, unsolicited marketing, newsletters and automated confirmations were overwhelmingly held in inboxes, so this also provided a more relevant collection of emails to work from. While this process reduced our sample dramatically (to 4,160 emails), it was effective in allowing us to manually search the remaining emails using more established methods as opposed to computational alternatives (Kipping et al., 2014;

[1] The dataset was also originally available via a FERC; however, it is now only available via third parties. We used a version made available via Carnegie Mellon University (https://www.cs.cmu.edu/~enron/).

Rowlinson, 2004). Here, qualitative coding allowed patterns to emerge from the remaining emails, which in turn provided a further basis for our interpretation that would not have been possible through keyword searches (Glaser & Strauss, 1967). Reading the emails (and telephone transcripts) in this way maximised the chance of finding information that we were not explicitly looking for or indeed expecting.

In contrast to our approach, Gavin Benke engaged with Enron's emails via a now inaccessible online version managed by the FERC (Benke, 2018). To mitigate its size, he started by running keyword searches on the inboxes of key executives (like chairman, Ken Lay), using the results to lead him onto other potentially interesting accounts. As he notes, the networked nature of email meant that certain accounts "served as hubs through which a good proportion of relevant material passed" (Benke, 2018:229). While his keyword-based approach limited scope for new discoveries, it retained access to the whole collection and thus the potential for obtaining complementary insight on specific points of interest. This provided an effective method of source triangulation, with the emails largely corroborating and elaborating the insight gained from other sources more primary to his analysis (Kipping et al., 2014). Thus, while different from our own, Benke's approach also dealt effectively with the issues of informational "swamping" that research into more recent business histories can present (Fellman & Popp, 2013:218).

Somewhat paradoxically, digital sources also suffer from issues of scarcity; meaning there is simultaneously too much information to manage, and a greater risk that important records are lost, damaged or destroyed (Milligan, 2019; Rosenzweig, 2003). This digital entropy is particularly problematic within organisations themselves, where historical significance is often not the primary reason for record-keeping (Kirsch, 2009). While some companies maintain formalised procedures for recording digital information, limited adoption of such practices represents a significant threat to the future survival of potentially important

historical sources (Moss, 2012). Pertinently, it was only through federal intervention that we were able to obtain so much primary material on Enron, and the sources available on other companies involved are far less plentiful. Even in the comparatively abundant Enron records, gaps are readily evident. For instance, although there were emails on the Enron server, there does not appear to be any longer-term retention strategy. Thus, the number of surviving emails reduces steadily the further they get to the point of collection (see Figure 2).[2] From a research perspective this created a challenge, as the period we were most interested in (the peak of the crisis) corresponded to a time with fewer surviving emails. Furthermore, reliable access to those sources that do survive can be frustratingly ephemeral, as seen with the loss of access to the original FERC email database that Benke (2018) used in his analysis.

For their part, archivists have already acknowledged that digital technologies enable – and indeed, require – a different outlook to preservation (Dallas, 2016; Waugh et al., 2016). For some researchers, the 'noise' of digital information risks obscuring more salient details about the past (Nicholas & Clark, 2015), while others are already showing the insight that more complete historical datasets can provide (Tumbe, 2019). In meeting these diverse needs of users, archives may need to maintain multiple reborn-digital versions of a source, each set up to afford a different form of interrogation. Alternatively, historians may increasingly prefer to make their own selection decisions on comparatively unedited collections, as we did with the EED. While this is often seen within organisation studies (for instance, Aven (2015) and Byun & Kirsch (2020) use email records to investigate organisational communication norms), unrealised potential certainly exists for more historically focused questions.

[2] The number of emails also reduced after the company's bankruptcy in late-2001; however, it is likely that this was because fewer emails were being sent in the first place.

Leveraging automated metadata

For historians, entry into an archive involves the study of a collection based on the archive's catalogue information (or metadata). While automation of digital metadata can create issues for archival processing in term of accuracy and format consistency (Gollins & Bayne, 2015), it provides new contextual information that analogue sources rarely afford. For instance, because digital files generally contain embedded temporal information, a digital source often provides the researcher with a highly precise understanding of when it was created. For pre-digital history, the dating and sequencing of sources are often limited to the day of creation, with letters and meeting minutes customarily including the date as part of the preamble to any actual content. This represents a limited but vital form of temporal metadata, which historians rely upon to place sources within their historical and intertextual context. For many born-digital sources, the increased accuracy provided by automated timing offers a far greater degree of temporal specificity and mitigates the higher frequency of communication within genres such an email or instant messaging. Within our research, telecommunications recordings offered a suitable example of this specificity.

When digital recordings of traders' telephone conversations were originally created, the date and time the call began and concluded were captured to the centisecond. Taking the metadata of '20001206-8521801-9090564 (16:47)' as an example, we would, therefore, know that a 16 minute and 47 second call took place on 6th December 2000 between 08:52 (and 18.1 seconds) and 09:09 (and 5.64 seconds). This information proved vital in developing a diachronic appreciation of events, illuminating the causal interconnections between the various dispersed and fragmented interactions (Lipartito, 2014). During a given trading period, it was common for traders to make and receive numerous calls to various contacts in a very short space of time (e.g., 15 minutes). Using the digital time stamps, we were able to maintain an understanding of the sequence of these calls without having to infer it from their contents (cf.

Mink, 1966). Through this minute-by-minute reconstruction of events, we were able to produce a highly detailed analysis that uncovered a chain of critical incidents that occurred over a matter of minutes and hours rather than days. In this way, automated metadata affords historians the ability to interpret tightly spaced interactions, lifting small and seemingly ordinary moments and showing their collective significance to historical events. The trading tapes are by no means an exception in relation to metadata of this sort, and it is possible to take similar time stamps from email, websites, and even documents. Moreover, the geotagging of images means it is highly likely that future historians will not only know when a photograph was taken but also exactly where (Hernández-Ramírez, 2013).

In addition to its analytical usefulness, the presence of automated metadata is important for source criticism, as it records details about the authorship of a source and the circumstance and context of its production (Kipping et al., 2014). However, born-digital sources exist in an editable state, and their content (or metadata) can be altered by the original author or a third-party (Gollins & Bayne, 2015; Schwarzkopf, 2012). This is often a perfectly legitimate condition of its original use, as seen by online news articles, which often undergo multiple changes as a news story develops (Cohen & Rosenzweig, 2006). However, the recent prominence of 'fake news' is a timely reminder that malign intent cannot be ruled out. In many cases, metadata does record these alterations; however, this is by no means a certainty and, when combined with easy duplication, a definitive 'original' of some digital sources simply will not exist. Outstanding questions also remain about the nature of authorship itself, and digital technologies like intelligent writing aids (Grammarly), autoreply functions (Gmail) and collaboration platforms (Slack) all pose methodological questions that we are yet to answer. Nonetheless, automated metadata provides at least a potential basis for new forms of source criticism, providing an ability to appraise sources and identify the most historiographically valid and credible option.

[Figure 2]

For future business historians, company records have the potential to illuminate new aspects of organisational pasts, as they did for us in relation to Enron. However, effective usage of such source requires that we adapt our practice to accommodate their characteristics and the nature of their preservation as abundant, but fragile. This is particularly true of the approaches used to organise and search digital records, and the choice between traditional reading of a limited sample, or the application of search queries on an entire corpus. While the former retains a closeness to individual sources valued by historians, the latter presents opportunities for coverage well beyond what historians can process manually. Similarly, the digital information embedded in sources has added a new aspect to the traces that organisational life leaves behind and using this new form of metadata allows historians to sequence, contextualise, and critique sources in a way previously not possible. For some sources, particularly digital correspondence, an appreciation of this information will increasingly become an unavoidable feature of future historical enquiry.

Searching digital media archives for entrepreneurial pasts

While organisational digital archives may not always be accessible to historical researchers, published media such as newspapers have long been a key resource, though perhaps less typically within business history (Bowie, 2019; Heller & Rowlinson, 2020). It differs from research in company archives, which usually contain a variety of different documents, not just one 'genre'. In the past, such research was facilitated by places like the former British Newspaper Library in Collingwood, which maintained original print and micro-fiche copies that have now become obsolete. Magazines, such as *The Economist*, have a fully digitised historical archive available online as PDF files, meaning that historical research on newspapers has mostly become digital already. Other subscription-only databases such as *Nexis UK* or *BusinessSource Complete* aggregate print media (and sometimes summary or text of radio

broadcasts) worldwide. They are in effect genre-based digital archives and give us a sense of the kind of breadth and granularity that will become the main features of these collections of the future. Most importantly, they provide us with an insight into the affordances of full text and metadata search facilities, and how this is likely to impact on the future practice of archival research in the digital sphere.

Clear search algorithms by which to identify relevant documents in a much larger, not necessarily subject-specific collection obviously have the potential to speed up archival research significantly. Gone the quiet, introspective and (probably not so) dusty weeks in archives that historians such as Steedman (2002) references (see also Czarniawska & Löfgren, 2013; Fellman & Popp, 2013). With a paper archive that is too large to exhaustively search, historians have had little choice but to strategically dive into parts of the collections to maximise both coverage and serendipitous finds (Decker, 2013). This general sense of context is obviously lost by relying on the more targeted digital search, but arguably digital content is too vast to really allow researchers to comprehend the entirety of a collection without search functions. The results of this search usually still require sifting and excluding items that were brought up by mistakes – ultimately this second step of manual assessment of results is necessary and it allows more exhaustive, if perhaps vaguer, search terms.

However, as a quicker approach to finding relevant sources, it also reduces the general familiarity with the material that comes from manual searching and skim-reading. This limits the opportunity for serendipitous finds but by selectively reading full sources some of this may be regained. Moreover, search terms are, by necessity, more likely to be descriptive, describing categories of people (entrepreneurs, women), events or activities (mergers, interviews) rather than analytical terms (gender, inequality). Familiar to historians is also that this requires paying attention to potential synonyms, particularly as the relevant terminology may change over time (Koselleck, 1982). Online databases such as *BusinessSource Complete* and others allow

Boolean search phrases – terms such as AND, OR, AND NOT – which allow a narrowing-down of results beyond what a normal paper-based index can achieve. In practice it takes some time to understand the syntax and adjust it to the type of data one is searching. This is one obvious example where the digital nature of sources does not just afford different ways of discovery, but also shapes historical practice in new ways that have not been fully described or explored. Digital sources offer better techniques to deal with an overwhelming abundance of primary sources (McNeill, 1986), which would make it impossible to 'read everything'; they also require more transparent description of sampling strategies and analytical techniques. That, however, is not commonly done in qualitative historical research at present.

Finding sources in a digital archive

Boolean and full-text search also changes how researchers work with the resulting digital sources, as they can now be understood as a body of text that can be analysed with computer-aided tools, and potentially without reading the relevant text in its entirety. We explored the opportunities afforded by this type of source and the use of programmes designed to facilitate computer-aided discourse analysis (CADA) in a research project focused on how entrepreneurs use the past in media interviews.[3] The data was collected from *BusinessSource Complete* and analysed with CADA software,[4] used widely by corpus linguistics, a specialist approach within linguistics that investigates language use through large quantities of naturally occurring text (Baker, 2006; Baker et al., 2012). By corpus, we mean a large body of text in which the contents are equivalent and represent a theme or genre. This approach made sense as the literature on uses of the past focuses on the use of rhetoric and narrative (Anteby & Molnar, 2012; Foster, Coraiola, Suddaby, Kroezen, & Chandler, 2017; Mordhorst & Schwarzkopf, 2017; Oertel &

[3] This research project was conducted by the second author with a different team of collaborators.
[4] There are a number of packages available: Voyant Tools offers the widest range of features: https://voyant-tools.org/. AntConc is freeware: https://www.laurenceanthony.net/software/antconc/. WordSmith and WMatrix cost £50 per licence: https://www.lexically.net/wordsmith/ and http://ucrel.lancs.ac.uk/wmatrix/. We used WMatrix for this project as it facilitates semantic analysis.

Thommes, 2015; Suddaby, Foster, & Trank, 2010; Wadhwani, Suddaby, Mordhorst, & Popp, 2018; Zundel, Holt, & Popp, 2016), which are core themes of linguistics.

We collected interviews with entrepreneurs by linking these as search terms through the Boolean operator AND, starting our search in the late 1990s when these interviews were likely to be available in machine-readable formats. Through a Boolean search and subsequent manual checking of all items for relevance, 327 interviews with entrepreneurs were identified from publicly available sources such as magazines and other media outlets, published between 1996 and 2015. The total text amounts to nearly 800 single-spaced pages. However, neither the way they are stored in the database nor the format in which we chose to download them (TXT files) was the same format in which they were initially published. To facilitate textual analysis, the files were edited to remove header and footer information other than the title, author and date, as otherwise, every single file would have contained the words 'Business', 'Source', and 'Complete', as well as newspaper and magazine names, which would have skewed our analysis by highlighting the high density of such words in the analysis. Due to the requirements of the analysis software, the individual interviews were then merged into one TXT file. Thus, the degree of processing of the downloaded files means that they represent files in the reborn-digital category. This is because they were not considered in the format in which they were originally created, but rather converted into a format that made their content easier to analyse. We anticipate that this level of processing of digital sources is likely to become more common in order to exploit the affordances of CADA-type software for historical research. However, this implies a significant challenge to existing notions of authenticity common to historical research and archival practice, such as working with the original document. Ultimately,

authenticity is likely to be maintained by archival practices that are still evolving, such as the use of blockchain in the ARCHANGEL project (ARCHANGEL, n.d.).[5]

While this extraction of files from a database may seem substantially different from ordinary archival research, the process by which a subset of sources is selected from a larger archive to create a historian's personal research collection is similar with analogue sources. While some researchers may take handwritten notes, increasingly historians are taking notes directly in word files, either as summaries or verbatim quotations. Since the advent of digital photography, analogue historical files can easily be digitised as image files (GIF, TIFF, JPEG, PDF etc.) while in the archive, allowing researchers to selectively digitise files for their private databases. So, whether this processing of digital archival files into reborn-digital formats will ultimately present any greater challenge to assuring authenticity than current practices, remains to be seen. Most likely, the role of the archivist in assuring the authenticity of the original digital file of record will become more prominent, with the potential for hyperlinking directly into digital archives where they are open to the public, though this is unlikely for private company archives (Smith & Umemura, 2019).

Giving new voice to historical sources

The digital text file(s) that result from this processing can be searched for key terms and themes, which permits faster, more selective reading as well as standardised and replicable coding techniques. In the research project on the uses of the past by entrepreneurs, these affordances enabled us to look at historical sources in a completely new way. This approach offered significant advantages over a standard full-text search, as the software identified themes based on a pre-existing semantic dictionary (see figure 3) rather than just highlighting individual words (Wilson & Thomas, 1997). Other features include, for instance, the analysis of the

[5] Project ARCHANGEL seeks to safeguard the future of digital records by creating assurances of digital record integrity through distributed ledger (blockchain) technology. The aims are to protect data against tampering and restore trust in digital records.

frequency and co-occurrence of certain words, as well as the statistical significance of certain word choices compared to another reference corpus that represents 'standard' use of language (for example the British National Corpus, BNC).[6] By comparing these texts to a reference corpus, words and themes that are used more frequently reveal what this particular selection of texts is 'about'.

[Figure 3]

The processed text was analysed with the aid of a semantic software developed for CADA, WMatrix (Rayson, 2008, 2009). Semantics is a branch of linguistics concerned with word meanings and relationships between these meanings. The web-hosted software used automatically annotates text based on a standardised system of semantic domains (Wilson & Thomas A, 1997).[7] Even though these software packages were originally developed to identify lexical choices and linguistic patterns, researchers have used them to study the content and meaning of documents (Noel & Erskine, 2013; Pollach, 2012). The advantage of this programme is that it identifies words that are equivalent in meaning, e.g. forever, never-ending, and permanent. Like other CADA tools, the software linked the analysis directly back to the original text, so that researchers can easily go back to the full source and appreciate the context of the conversation that gave rise to dominant themes in the text. This also allows a consideration of the document in terms of historical source analysis (Dobson & Ziemann, 2008; Howell & Prevenier, 2001). WMatrix works especially well when researchers do not know what terms they are looking for because the initial analysis presents an aggregate and quantifiable picture of what a large amount of text is 'about' before commencing in-depth reading. While such tools do not preclude historians unexpectedly gaining insight from

[6] For an introduction to the BNC, see http://www.natcorp.ox.ac.uk/ .
[7] WMatrix is about 91 percent accurate in identifying semantic domains (Rayson, Archer, Piao, & McEnery, 2004) so we excluded any errors in identifying the correct semantic domain from our analysis.

happenchance ‘flashes’ in the archive (Stoler, 2010), they do provide a way to identify themes and discover relevant sources in the absence of catalogue information or archival guidance.

The topics or themes that are identified through the software can be explored either because of their relevance to the research question or because they stand out as particularly prominent in the particular selection of texts investigated. Linguistic software packages designed for CADA afford some advantages over alternative approaches as they facilitate easy access from the extracted passage back to the full document. This means that once themes are identified, they can be explored qualitatively and manually and allows researchers to narrow down their reading to relevant sections or documents within a large amount of text. Here the software can be used as much as a search aid than an analytical tool, depending on the researcher’s preferences. Hence digital sources that offer full-text search do not just add a quantitative component to text analysis, but also support a more focused qualitative analysis by identifying relevant passages, even those that researchers may not have been aware of before analysing the text. Ultimately this achieves greater replicability in terms of selection of archival materials but still leaves the interpretation of the results to the historical researcher.

Other disciplines are already leveraging the significant amount of social text being made available on the internet and social media platforms for research, such as analysing news reporting on controversial issues (Baker et al., 2012; Grundmann & Krishnamurthy, 2010) or language use on Twitter (Huang et al., 2016; Nini et al., 2017). The software programmes used are very much tailored for linguistic use, and admittedly their relevance for historians is somewhat limited. Designed to analyse language rather than content, they nevertheless offer interesting and ready-made approaches to using full-text search and identifying themes in the sources that may not have been immediately apparent, thus offering new directions to research. These tools are not yet widely used by historians, but for sources that are digital and digitised like newspaper collections they offer new ways of exploring the material (Tumbe, 2019).

Ultimately, we chose WMatrix to investigate the meaning-making of over 300 entrepreneurs as they presented themselves in interviews in the media, in which they retrospectively narrated the story of their success to journalists. While our initial interest in history and the past overlapped neatly with a pre-coded semantic domain in the software, the analysis showed that entrepreneurs did not refer to the past more often than one would expect in 'normal' language use (here we used a specialist sub-corpus of the BNC that focused on business communications, as well as the entire BNC, with similar results). More intriguingly, they referred even less frequently than expected to the future, but overall, their use of temporal semantic domains was much higher than in normal speech. Rather than employing static categories like past or future, their use of temporal vocabulary was more dynamic, focusing on growth and longevity (Garud & Giuliani, 2013). This nuance became apparent because of the fixed semantic domains already used by the software, which allowed us to break down language use in ways that we would not have considered initially, and which afforded us a comparative perspective to 'normal' language use through the BNC. The literature on the uses of the past, while focusing on rhetoric and narrative, has not really investigated how entrepreneurs and organisations think about temporality in ways that are strikingly different from our simplistic past, present and future categories (Wadhwani et al., 2018). Thus, the CADA approach offers a different analytical angle through large scale comparative analysis from traditional source analysis.

Business historians can benefit significantly from engaging more with digital history and exploring the new types of search and analysis these new formats and tools offer. This requires a greater incorporation of computer-aided methods into historical research, as this will better allow researchers to capitalise on the characteristics of digital sources. Existing software can be used as search tools that produce a tailored index for researchers and thus help with dealing with too much material. They can also facilitate interpretation by highlighting

interesting aspects through large scale comparisons that would not have been possible with analogue sources. Yet current software options are tailored towards the needs of other disciplines, and are not always suitable for historical analysis. While archivists are exploring and developing solutions for historical collections, there is still relatively little collaboration with historical researchers who, for the most part, have not started investigating these new resources. To this end, existing methods, including those of other fields such as corpus linguistics, hold potential value for historians and archivists as they explore the possibilities of digital collections.

Conclusion: The future of business history?

Science fiction authors such as Adrian Tchaikovsky have already re-imagined the future research practice of historians, portraying them sifting through damaged ancient databases, and running sophisticated translation algorithms to decipher defunct languages and dialects from the deep past (Tchaikovsky, 2015). Albeit a piece of fiction, this creative re-imagining of a profession currently associated with dusty paper-based archives may well be closer to the truth of what research will look like for contemporary historians in only a decade or two. As our own research has shown, the digital shift will pose challenges to historical practice; however, we believe it also represents a significant opportunity. In particular, digital formats can provide new voice to historical sources and insight into organisational life that would traditionally be hard to preserve. Additionally, many of the basic tools needed to engage within digital sources have already been established in fields like computing and linguistics and adapting these for historical research presents encouraging avenues for methodological development. However, such adaption does not call for significant deviation from underlying historical methodology, which remains fundamental to the analysis and interpretation of digital sources. Rather, engagement with digital sources and new approaches means that methodological reflexivity,

source criticism and archival theory will be as important as ever for historical knowledge generation.

Today, the number of historians engaging actively with digital history is still small, though this is very likely to increase significantly as more sources are digitised and (re)born-digital records from the late twentieth and early twenty-first century become available (Cohen & Rosenzweig, 2006; Milligan, 2018). Business historians are to varying degrees aware of issues surrounding digitisation (Laurell et al., 2019; Schwarzkopf, 2012), digital company record keeping (Kirsch, 2009; Moss, 2009), historical analysis (Tumbe, 2019); however, digital sources and the information they afford have largely been absent from methodological debates. Among the considerations that will need to be addressed are the algorithms, protocols and technologies active in the (re)production of sources, and which need to be understood in order to appreciate their provenance and pertinence, and ultimately enable effective source criticism. Similarly, the type of insight digital sources provide historians is different, either by merit of the traces they capture, or the approaches used to investigate them, as our research shows in relation to everyday trading activities and the use of semantic tagging. Accordingly, while such issues need not preoccupy our attention to the detriment of other methodological questions, we need to reflect more explicitly on how we analyse digital sources. In connecting business history to the ideas and debates of digital history, we have sought to stimulate this reflection, but much more will be needed to meet the challenges and opportunities that digital sources present.

Some way ahead of historians, archives professionals are already adapting to born-digital accessions and the problems of archiving records when many are encoded in unique proprietary software, which may become obsolete through progress or business failures. Ensuring that data remains accessible 10, 20 or 30 years in the future is difficult to plan for and implement, though technical registries such as PRONOM do provide some support for archives

(The National Archives (UK), n.d.). Initiatives such as the *Task Force on Technical Approaches for Email Archives* are also making significant headway in addressing the archival practicalities of born-digital material and their particular characteristics (Prom et al., 2018). A key question that historians can help answer is how sources should be preserved to best meet their needs as users (Green & Lee, 2020). Maintaining file formats in their originally unedited form while providing a research-conducive resource is not easily achieved. For instance, without processing, advanced machine-readability is limited, making tools such as CADA less applicable. Even using more traditional methods, an abundant but uncurated collection often looks more like a mass data dump than a usable historical resource, even when compared to the most disorganised of company archives. Moreover, while archivists are currently dealing with these fundamental changes to their professional practices, training and skills, business historians have hardly begun to address their digital future, and what it will require of them in terms of new skills and methodological practices.

In this paper, we have explored the use of digital sources as basis for historical research, and in doing so bridge the gap between business history and established engagement with digital history. To do this, we engaged with digital and web history, offering a framework for business historians to understanding the characteristics of different digital sources and their interrelationship. Here we considered two practical examples: the use of digital-era company records and the searching of digital media archives. In elaborating how these issues impact on business history research, we have shared some of our practical responses and the results they yielded, offering tentative avenues for new methodological practices. To some extent, our experience shows that traditional methods can be employed to account for digital sources, with new, *digital* considerations complementing the established criteria for source validity and credibility. However, we also show that new approaches, like the analytical tools developed by linguists, hold the potential to enhance the professional practice of historians as well as

archivists. In either case, while the future of business history will likely remain embedded in the principles underlying its past, digital sources represent a material change for historical research, the significance of which we are only just starting to confront.

References

Abbott, A. (2014). *Digital paper: A manual for research and writing with library and internet materials*. University of Chicago Press.

Amos, J. (2020, March 31). Self-isolation proves a boon to rainfall project. *BBC Online*. https://www.bbc.co.uk/news/science-environment-52040825

Anteby, M., & Molnar, V. (2012). Collective memory meets organizational identity: Remembering to forget in a firm's rhetorical history. *Academy of Management Journal*, *55*(3), 515–540.

ARCHANGEL. (n.d.). *ARCHANGEL: Trusted Digital Archives.* Retrieved May 1, 2020, from https://www.archangel.ac.uk/

Aven, B. (2015). The Paradox of Corrupt Networks: An Analysis of Organizational Crime at Enron. *Organization Science*, *26*(4), 980–996.

Ayers, E. L. (2001). The pasts and futures of digital history. *History News*, *56*(4), 5.

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Baker, P., Gabrielatos, C., & McEnery, T. (2012). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998–2009. *Applied Linguistics*, *34*(3), 255–278.

Benke, G. (2018). *Risk and Ruin: Enron and the Culture of American Capitalism*. University of Pennsylvania Press.

Blevins, C. (2019). A Tour of the Virtual Stacks. *Modern American History*, *2*(2), 265–268.

Bolick, C. M. (2006). Digital archives: Democratizing the doing of history. *International Journal of Social Education*, *21*(1), 122–134.

Boss, K., & Broussard, M. (2017). Challenges of archiving and preserving born-digital news applications. *Ifla Journal-International Federation of Library Associations*, *43*(2), 150–157. https://doi.org/10.1177/0340035216686355

Bowie, D. (2019). Contextual analysis and newspaper archives in management history research. *Journal of Management History*, *25*(4), 516–532.

Breckenridge, K. (2014). The Politics of the Parallel Archive: Digital Imperialism and the Future of Record-Keeping in the Age of Digital Reproduction. *Journal of Southern African Studies*, *40*(3), 499–519.

Bressers, B., & Hume, J. (2012). Message Boards, Public Discourse, and Historical Meaning: An Online Community Reacts to September 11. *American Journalism*, *29*(4), 9–33.

Brügger, N. (2012). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research/Historische Sozialforschung*, *37*(4), 102–117.

Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.

Brügger, N., & Milligan, I. (2018). *The SAGE Handbook of Web History*. SAGE Publications Limited.

Bui, T., Cooper, D., Collomosse, J., Bell, M., Green, A., Sheridan, J., Higgins, J., Das, A., Keller, J., & Thereaux, O. (2019). ARCHANGEL: Tamper-proofing Video Archives using Temporal Content Hashes on the Blockchain. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0.

Byun, H., & Kirsch, D. (2020). The morning inbox problem: Email reply priorities and organizational timing norms. *Academy of Management Discoveries*, *In Press*.

Cassell, C., & Symon, G. (2004). *Essential guide to qualitative methods in organizational*

research. Sage. internal-pdf://235.199.119.66/Ess Guide to Qual Res.pdf

Cohen, D., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., Thomas, W., & Turkel, W. (2008). Interchange: The promise of digital history. *The Journal of American History*, *95*(2), 452–491.

Cohen, D., & Rosenzweig, R. (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. University of Pennsylvania Press.

Coopersmith, J. (2015). *Faxed: The Rise and Fall of the Fax Machine*. JHU Press.

Corrado, E., & Sandy, H. (2017). *Digital preservation for libraries, archives, and museums*. Rowman & Littlefield.

Crowley, P. (2005). *Initial Audio Tape Testimony of Patrick R. Crowley* (EL03-180, S-84).

Cummings, S., Bridgman, T., & Hassard, J. (2017). *A new history of management*. Cambridge University Press.

Czarniawska, B., & Löfgren, O. (2013). *Coping with excess: how organizations, communities and individuals manage overflows*. Edward Elgar Publishing.

Dallas, C. (2016). Digital curation beyond the "wild frontier": a pragmatic approach. *Archival Science*, *16*(4), 421–457.

DCDC. (2019). Navigating the Digital Shift. *Discovering Collections, Discovering Communities Conference Programme*. https://dcdcconference.com/wp-content/uploads/2019/11/DCDC19-Programme-WEB-FINAL.pdf

Decker, S. (2013). The silence of the archives: Business history, post-colonialism and archival ethnography. *Management & Organizational History*, *8*(2), 155–173.

Decker, S., Kipping, M., & Wadhwani, R. D. (2015). New business histories! Plurality in

business history research methods. *Business History*, *57*(1), 30–40.

Decker, S., Rowlinson, M., & Hassard, J. (n.d.). Rethinking History and Memory in Organization Studies: The Case for Historiographical Reflexivity. *Human Relations*, *forthcoming*, 1–50.

Dobson, M., & Ziemann, B. (2008). *Reading Primary Sources: The Interpretation of Texts from Nineteenth and Twentieth Century History*. Routledge.

Donnelly, M., & Norton, C. (2012). *Doing history*. Routledge.

Dougherty, J., & Nawrotzki, K. (2013). *Writing history in the digital age*. University of Michigan Press.

Dunn, S., & Hedges, M. (2016). How the crowd can surprise us: Humanities crowd-sourcing and the creation of knowledge. In M. Ridge (Ed.), *Crowdsourcing our Cultural Heritage* (pp. 231–246). Routledge.

Duranti, L. (2001). The impact of digital technology on archival science. *Archival Science*, *1*(1), 39–55.

Fayard, A.-L., & Weeks, J. (2007). Photocopiers and Water-coolers: The Affordances of Informal Interaction. *Organization Studies*, *28*(5), 605–634.

Fellman, S., & Popp, A. (2013). Lost in the archive: the business historian in distress. In B. Czarniawska & O. Löfgren (Eds.), *Coping with excess. How organizations, communities and individuals manage overflows.* Edward Elgar.

FERC. (2014). *Your Guide to Electronic Information at FERC*. FERC. https://www.ferc.gov/docs-filing/elec-info-guide.pdf?csrt=16938571389058315840

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N.

(2013). RESEARCH METHODS IN THE AGE OF DIGITAL JOURNALISM. *Digital Journalism*, *1*(1), 102–116. https://doi.org/10.1080/21670811.2012.714928

Foster, W. M., Coraiola, D. M., Suddaby, R., Kroezen, J., & Chandler, D. (2017). The strategic use of historical narratives: A theoretical framework. *Business History*, *59*(8), 1176–1200.

Garud, R., & Giuliani, A. P. (2013). A narrative perspective on entrepreneurial opportunities. *Academy of Management Review*, *38*(1), 157–160.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Aldine Publishing.

Gollins, T., & Bayne, E. (2015). Finding archived records in a digital age. In M. Moss & B. Endicott-Popovsky (Eds.), *Is Digital Different?: How information creation, capture, preservation and discovery are being transformed* (p. 129). Facet Publishing.

Graham, S., Weingart, S., & Milligan, I. (2012). Getting Started with Topic Modeling and MALLET. *The Programming Historian*. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Green, A. R., & Lee, E. (2020). From transaction to collaboration: redefining the academic-archivist relationship in business collections. *Archives and Records*, *41*(1), 32–51.

Grundmann, R., & Krishnamurthy, R. (2010). The discourse of climate change: A corpus-based approach. *Critical Approaches to Discourse Analysis across Disciplines*, *4*(2), 125–146.

Harvey, C., & Jones, G. (1990). BUSINESS HISTORY IN BRITAIN INTO THE 1990s. *Business History*, *32*(1), 5.

Heller, M., & Rowlinson, M. (2020). Imagined corporate communities: Historical sources and discourses. *British Journal of Management*, *31*(4), 752–768.

Hernández-Ramírez, R. (2013). Visualising photography: The photographic image and its

ontological status after the information revolution. *In Cultural Technologies and Media Arts Conference*.

Howell, M. C., & Prevenier, W. (2001). *From reliable sources: An introduction to historical methods*. Cornell University Press.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255.

Jaillant, L. (2019). After the digital revolution: working with emails and born-digital records in literary and publishers' archives. *Archives and Manuscripts*, *47*(3), 285–304. https://doi.org/10.1080/01576895.2019.1640555

Keulen, S., & Kroeze, R. (2012). Back to business: A next step in the field of oral history—the usefulness of oral history for leadership and organizational research. *The Oral History Review*, *39*(1), 15–36.

Kipping, M., Wadhwani, R. D., & Bucheli, M. (2014). Analyzing and interpreting historical sources: A basic methodology. In M. W. Bucheli RD (Ed.), *Organizations in Time: History, Theory, Methods* (pp. 305–330).

Kirsch, D. (2009). The record of business and the future of business history: Establishing a public interest in private business records. *Library Trends*, *57*(3), 352–370.

Koselleck, R. (1982). Begriffsgeschichte and social history. *Economy and Society*, *11*(4), 409–427.

Kroeze, R., & Vervloet, J. (2019). A life at the company: oral history and sense making. *Enterprise & Society*, *20*(1), 33–46.

Lansdall-Welfare, T., & Cristianini, N. (2020). History playground: A tool for discovering

temporal trends in massive textual corpora. *Digital Scholarship in the Humanities*, *35*(2), 328–341.

Laurell, C., Sandström, C., Eriksson, K., & Nykvist, R. (2019). Digitalization and the future of Management Learning: New technology as an enabler of historical, practice-oriented, and critical perspectives in management research and learning. *Management Learning*, *51*(1), 89–108.

Lipartito, K. (2013). Connecting the cultural and the material in business history. *Enterprise & Society*, *14*(4), 686–704.

Lipartito, K. (2014). Historical sources and data. In M. Bucheli & R. D. Wadhwani (Eds.), *Organizations in Time: History, Theory, Methods* (pp. 284–304). Oxford University Press.

Maclean, M., Harvey, C., & Clegg, S. (2017). Organization theory in business and management history: Present status and future prospects. *Business History Review*, *91*(3), 457–481.

McNeill, W. H. (1986). Mythistory, or truth, myth, history, and historians. *The American Historical Review*, *91*(1), 1–10.

Megill, A., Shepard, S., & Honenberger, P. (2007). *Historical knowledge, historical error: A contemporary guide to practice*. University of Chicago Press.

Milligan, I. (2018). Historiography and the Web. In N. Brugger & I. Milligan (Eds.), *The SAGE Handbook of Web History*. SAGE Publications Ltd.

Milligan, I. (2019). *History in the Age of Abundance?: How the Web Is Transforming Historical Research*. McGill-Queen's University Press.

Mills, A. J., Weatherbee, T. G., & Durepos, G. (2014). Reassembling Weber to reveal the-past-as-history in management and organization studies. *Organization*, *21*(2), 225–243.

Mink, L. O. (1966). The autonomy of historical understanding. *History and Theory*, *5*(1), 24–47.

Mordhorst, M., & Schwarzkopf, S. (2017). Theorising narrative in business history. *Business History*, *59*(8), 1155–1175.

Moss, M. (2009). Archival research in organizations in a digital age. *The SAGE Handbook of Organizational Research Methods, London: SAGE Publications Ltd*, 395–408.

Moss, M. (2012). Where have all the files gone? Lost in action points every one? *Journal of Contemporary History*, *47*(4), 860–875.

Moss, M. (2015). What is the same and what is different. In M. Moss & B. Endicott-Popovsky (Eds.), *Is digital different?: How information creation, capture, preservation and discovery are being transformed* (pp. 1–17). Facet Publishing.

Mullen, L. (2014). Using Metadata and Maps to Teach the History of Religion. *Transformations: The Journal of Inclusive Scholarship and Pedagogy*, *25*(1), 112–118.

Nicholas, D., & Clark, D. (2015). Finding stuff. In M. Moss & B. Endicott-Popovsky (Eds.), *Is Digital Different?: How information creation, capture, preservation and discovery are being transformed* (pp. 19–34). Facet Publishing.

Nicholson, B. (2013). THE DIGITAL TURN. *Media History*, *19*(1), 59–73. https://doi.org/10.1080/13688804.2012.752963

Nini, A., Corradini, C., Guo, D., & Grieve, J. (2017). *The application of growth curve modeling for the analysis of diachronic corpora*. *7*(1), 102.

Noel, T., & Erskine, L. (2013). The Silent Story: Using Computer-Aided Text Analysis to Predict Entrepreneurial Performance. *The Journal of Entrepreneurship*, *22*(1), 1–14.

Norton, C., & Donnelly, M. (2018). *Liberating Histories*. Routledge.

Oard, D., Webber, W., Kirsch, D., & Golitsynskiy, S. (2015). *Avocado Research Email Collection* (LDC2015T03). Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2015T03

Oertel, S., & Thommes, K. (2015). Making history: Sources of organizational history and its rhetorical construction. *Scandinavian Journal of Management*, *31*(4), 549–560.

Onaga, L., & Shell, H. R. (2016). Digital histories of disasters: history of technology through social media. *Technology and Culture*, *57*(1), 225–230.

Pollach, I. (2012). Taming textual data: the contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, *15*(2), 263–287.

Prom, C. (2016). *Digital Preservation Essentials*. Society of American Archivists.

Prom, C., Murray, K., Baker, F., Connelly, M., & Gogel, W. (2018). *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives*. https://www.clir.org/pubs/reports/pub175/

Putnam, L. (2016). The Transnational and the Text-Searchable: Digitized Sources and the Shadows They CastThe Transnational and the Text-Searchable. *The American Historical Review*, *121*(2), 377–402.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549.

Rayson, P. (2009). Wmatrix: A Web-Based Corpus Processing Environment. *Lancaster*. http://ucrel.lancs.ac.uk/wmatrix/.

Rosenzweig, R. (2003). Scarcity or abundance? Preserving the past in a digital era. *The*

American Historical Review, *108*(3), 735–762.

Rowlinson, M. (2004). Historical Anelysis of Company Documents. In C. S. Cassell Gillian (Ed.), *Essential guide to qualitative methods in organizational research*. Sage.

Rowlinson, M., Hassard, J., & Decker, S. (2014). Research strategies for organizational history: A dialogue between historical theory and organization theory. *Academy of Management Review*, *39*(3), 250–274.

Schwarzkopf, S. (2012). What is an archive - and where is it? Why business historians need a constructive theory of the archive. *Business Archives: Sources and History*, *105*, 1–9.

Schwarzkopf, S. (2013). Why business historians need a constructive theory of the archive. *Business Archives*, *105*, 1–9.

Seefeldt, D., & Thomas, W. (2009, May). What is Digital History ? A Look at Some Exemplar Projects A Look at Some. *Intersections: History and New Media*.

Smith, A., & Umemura, M. (2019). Prospects for a transparency revolution in the field of business history. *Business History*, *61*(6), 919–941.

Spencer, R. (2017). Binary trees? Automatically identifying the links between born-digital records. *Archives and Manuscripts*, *45*(2), 77–99.

Steedman, C. (2002). *Dust: The archive and cultural history*. Rutgers University Press.

Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search, and metadata as archival processes for assessing historical contextualization. *The American Archivist*, *74*(2), 544–575.

Stoler, A. L. (2010). *Along the archival grain: Epistemic anxieties and colonial common sense*. Princeton University Press.

Suddaby, R., Foster, W. M., & Trank, C. Q. (2010). Rhetorical history as a source of competitive advantage. *Advances in Strategic Management*, *27*(2010), 147–173.

Tchaikovsky, A. (2015). *Children of Time*. PanMacmillan.

The National Archives (UK). (n.d.). *PRONOM technical registry*. Retrieved May 1, 2020, from https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The National Archives (UK). (2017). *Digital Strategy*. https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf

Thomas, W. (2016). The Promise of the Digital Humanities and the Contested Nature of Digital Scholarship. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities* (2nd ed.). Wiley-Blackwell.

Thompson, N. (2017). Hey DJ, don't stop the music: Institutional work and record pooling practices in the United States' music industry. *Business History*, *60*(5), 677–698.

Trouillot, M.-R. (1995). *Silencing the past: Power and the production of history*. Beacon Press.

Tumbe, C. (2019). Corpus linguistics, newspaper archives and historical research methods. *Journal of Management History*.

Turner, M. (1978). THERE IS NO FUTURE FOR BUSINESS HISTORY! *Business History*, *20*(2), 235.

Vallejo Pousada, R., & Larrinaga, C. (2020). Travel agencies in Spain during the first third of the 20th century. A tourism business in the making. *Business History*, 1–20.

Wadhwani, R. D., Suddaby, R., Mordhorst, M., & Popp, A. (2018). *History as organizing: Uses of the past in organization studies*. SAGE Publications Sage.

Waugh, D., Roke, E. R., & Farr, E. (2016). Flexible processing and diverse collections: a tiered approach to delivering born digital archives. *Archives and Records-the Journal of the Archives and Records Association*, *37*(1), 3–19.

Wilson, A., & Thomas A, J. (1997). Semantic Annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 53–65). Longman.

Yates, J. (2005). *Structuring the information age: Life insurance and technology in the twentieth century*. JHU Press.

Zundel, M., Holt, R., & Popp, A. (2016). Using history in the creation of organizational identity. *Management & Organizational History*, *11*(2), 211–235.

Figure 1 – The Digital Characteristics of Historical Sources

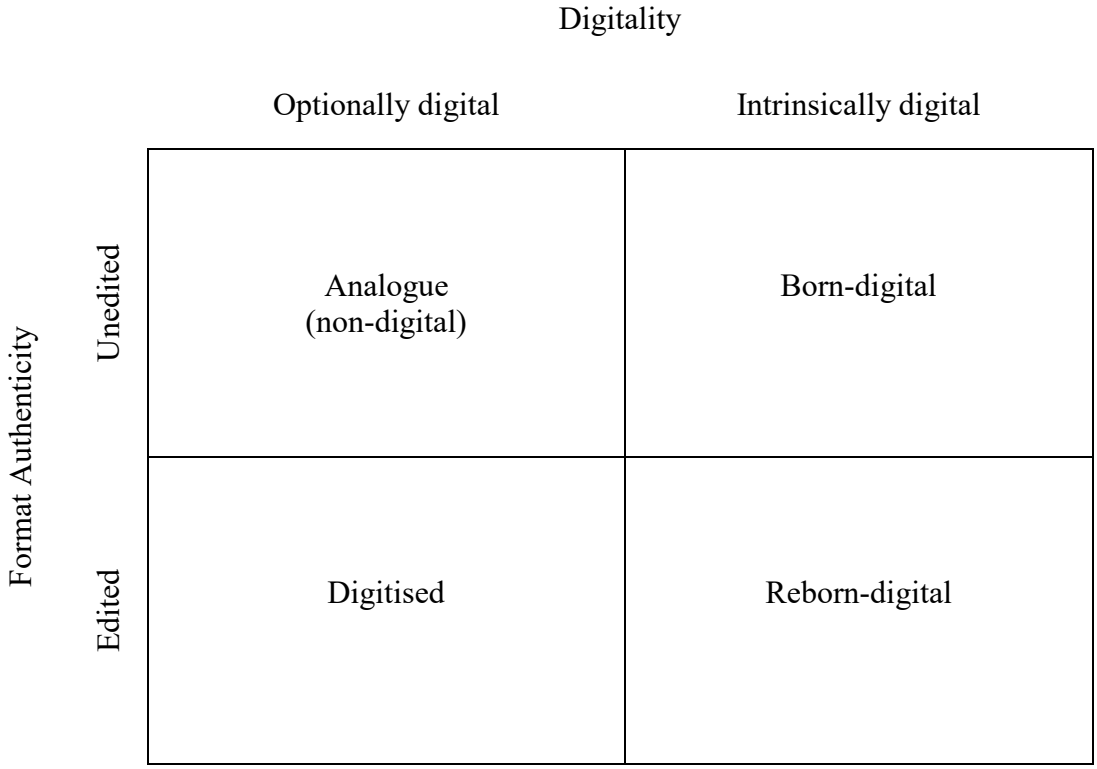

Table 1 Summary of digital sources for the Enron and Entrepreneurial Narratives projects

Source Example/Name	Available Source Format	Details	Notes
Enron Trader Tapes	Typed transcripts (PDF) taken from digital audio tapes (DATs).	380 individual telephone conversations, internal and external	Social Document (spontaneous dialogue, mostly one to one)
Enron Email Dataset	TXT files held within original email folder structure	Purpose built corpus of 4,160 emails from 20 Enron West Trading employees	Social Document (written communication, often one to many)
International media interviews with entrepreneurs	Online database (HTML), downloaded as TXT files	327 publicly available interviews with entrepreneurs in magazines or similar (1996 - 2015); c. 800 single-spaced pages.	Edited and negotiated articles based on journalistic interviews Questions may have been pre-arranged, and answers may have been edited

Figure 2 Email sent dates over time

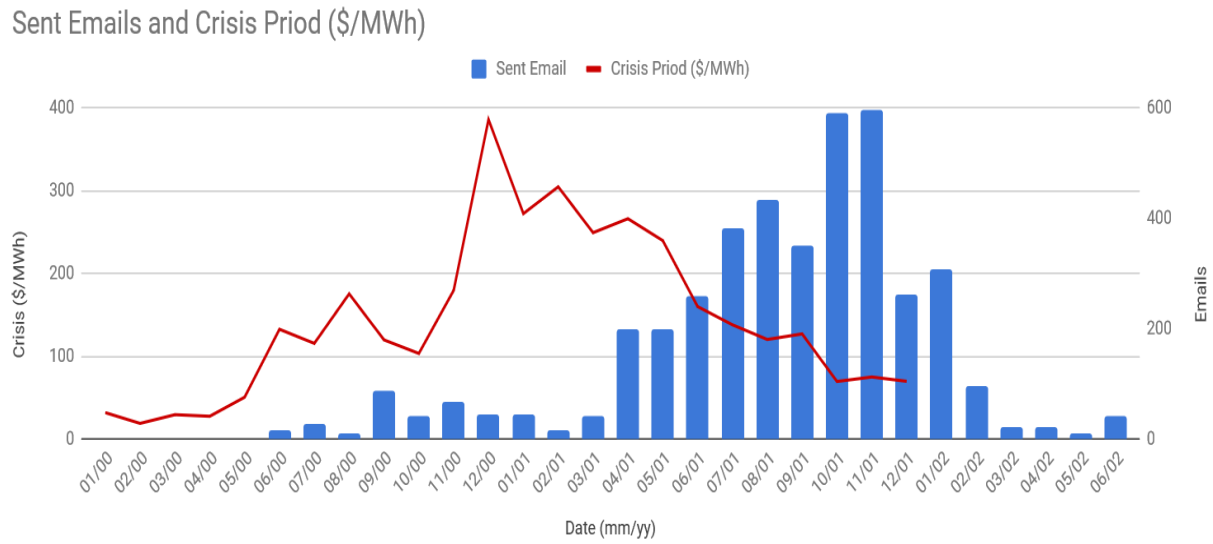

Figure 3 Overview of broad semantic domains pre-coded in WMatrix

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			