Critical Questions for Archives as (Big) Data[1]

DEVON MORDELL

ABSTRACT We may observe a growing preoccupation in archival literature with characterizing digital archives as *big data* – a term that suitably captures both their scale and their potential for manipulation through the application of computational methods and techniques for the purposes of discovering new insights. The possibilities for working with digital archives as data, from supporting archival arrangement and description tasks to promoting the use of digital archives as data sets by researchers, are indeed encouraging. But what are digital archives becoming when they are reframed as data, big or otherwise? What consequences might such a conceptualization have for the ways archival professionals imagine their role and their work? To the four archival paradigms of evidence, memory, identity, and community theorized by Terry Cook, a fifth may now be poised to emerge: an archives-as-data paradigm. In this article, I begin to map out what an archives-as-data paradigm could entail by exploring how the conceptual and practical dimensions of applying computational methods to digital archives may work conservatively to revivify notions of archival neutrality. For an archives-as-data paradigm to realize the more liberatory aims of which it is capable, an active and ongoing commitment to recognizing and calling out these tendencies is necessary.

1 The title of the article pays homage to danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* 15, no. 5 (2012): 662–679.

RÉSUMÉ On peut dénoter, dans la littérature archivistique, une préoccupation croissante quant à la qualification des archives numériques comme données massives – une terminologie qui souligne bien à la fois leur taille et leur potentiel à être manipulées par l'utilisation de méthodes et de techniques informatiques pour en extraire de nouvelles idées. Les possibilités d'utiliser les archives numériques comme données – de l'aide apportée aux tâches de classification et de description archivistique à la promotion de l'utilisation par les chercheurs des archives numériques en tant qu'ensemble de données – sont, en effet, encourageantes. Mais que deviennent les archives numériques lorsqu'elles sont redéfinies comme données, massives ou autres? Quelles conséquences une telle conceptualisation peut-elle avoir sur la façon dont les professionnels des archives pensent leur rôle et leur travail? Aux quatre paradigmes archivistiques de preuve, de mémoire, d'identité et de communauté, théorisés par Terry Cook, pourrait s'ajouter un cinquième: le paradigme de l'archive comme donnée (archives-as-data). Dans cet article je tente d'abord d'établir ce que pourrait signifier un paradigme d'archive comme donnée en explorant la façon dont les dimensions conceptuelles et pratiques de l'emploi de méthodes informatiques sur les archives numériques peuvent servir, à tout le moins, à raviver les notions de neutralité des archives. Pour qu'un paradigme d'archive comme donnée soit en mesure de produire, dans toute son ampleur, l'effet libérateur recherché, un engagement actif et soutenu à repérer et désigner ces tendances est nécessaire.

Now, however, the archivist is beginning to appreciate the applicability of some new techniques to his problem. He is beginning to see the computer as a boon . . . to the advancement of specific manipulation of large bodies of data to meet the individual needs of the historian or other researcher.[2]

The prescience of Rhoads' address – before, it must be remembered, the widespread commercial availability of personal computers – is astounding: in addition to recognizing its utility for clerical purposes, Rhoads envisages a "cybernetic approach" in which the computer becomes an "extension of the researcher himself," an intelligent system able to identify archives related to a concept or topic.[3] Through a database, he proposes, "*all* the information currently known about the archives' records" – including finding aids, lists, and catalogues – could be compiled to generate a special topic-based guide on the fly from user-provided search terms.[4] He predicts the replacement of static printed inventories with "a fluid body of data, ready to be retrieved according to the plan of the researcher," anticipating the delivery of archival description through dynamic, database-driven websites. Though Rhoads' account does not include the contents of the records themselves as an information resource at the computer's disposal, his vision of an encyclopedic retrieval system may well have accommodated them, given tools like document scanners, optical character recognition, and natural language processing algorithms. He concludes by praising the emancipatory possibilities of tailored access to archival description via computer – access enabling "a state of individual research freedom impossible to attain in a manual system."[5]

2 James Rhoads, "The Historian and the New Technology," *American Archivist* 32, no. 3 (1969): 211 (emphasis added); see also Anne J. Gilliland, "Archival Description and Descriptive Metadata in a Networked World," in *Conceptualizing 21st-Century Archives* (Chicago, IL: Society of American Archivists, 2014), 110–11. Gilliland's chapter directed me to Rhoads, of the US National Records and Archives Service, and this address at the annual luncheon of the Society of American Archivists on 17 April 1969, and provided gratefully appreciated contextual information for it.

3 Rhoads, "The Historian and the New Technology," 212. With remarkable foresight, Rhoads' "cybernetic approach" approximates topic modelling, a natural language processing technique that has been used experimentally in archival arrangement and description.

4 Ibid. (emphasis in original).

5 Ibid., 213. Rhoads continues emphatically: "It will be the computer that will liberate researchers and enhance the role of the individual in his attempt to reconstruct the past by giving each researcher the opportunity to ask for

"Data" features prominently in Rhoads' address: data banks that store search queries, exposing both patterns within requests and their responses; databases that emulate the conceptual framework of archivists; reference tools *as* data. It is not merely data that Rhoads describes, however; he seeks to manipulate "large bodies of data" and "*all* the information currently known," albeit on a scale that might seem quaint by contemporary standards. Rhoads explicitly connects the quantity of data to the quality of the service provided to the researcher, a gesture that aligns him with the present-day enthusiasm for big data.[6] Indeed, his address is a fitting preface to the growing preoccupation in the archival field with conceptualizing digital – and even non-digital – archives as (big) data.[7] The term *big data* captures both the metaphorical and the actual dimensions of many born-digital fonds. As data, textual records can be mined to facilitate archival arrangement and description or interpreted through information visualization techniques to allow for novel approaches to access and discovery.[8] The transformation of digitized archival materials into data that researchers can manipulate through computational operations offers the tantalizing promise of engaging new audiences and an assurance that the archival profession will remain relevant in an age of big data. The question of whether there is much to be gained from conceptualizing and working with digital archives as data is not at issue here, nor is the importance of preparing archival professionals to support such an endeavour. What I mean to explore, however, are the latent assumptions and biases that hitch a ride when an archives-as-data conflation takes place and that undermine the liberatory potential of such a development. In their analysis of big data as a socio-technical phenomenon, danah boyd and Kate Crawford assert that "how we handle the emergence of an era of Big Data is critical";[9] within the

information in the form and to the extent that suits his personal needs best. That is freedom."

6 In their "Critical Questions for Big Data," boyd and Crawford address the pervasive but erroneous belief about the innate superiority of "bigger data" (see pages 663, 668–70).

7 The term *digital archives*, as I will use it throughout the article, is intended to encompass both born-digital archives and the digitized surrogates of non-digital archives.

8 Though there are some fledgling tools for working with audiovisual records, like automated transcription services and computer vision techniques, textual records continue to be privileged by the considerably more mature and abundant text-based computational methods available for archival arrangement and description.

9 boyd and Crawford, "Critical Questions for Big Data," 664. The authors capitalize the term *Big Data* to distinguish it as a "cultural, technological, and scholarly phenomenon"; that is, "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross reference large data sets" (ibid., 663).

archival field, it requires an attentiveness to the more regressive dimensions of positioning digital archives as data.

In this article, I will examine the reframing of digital archives as (big) data, symbolically achieved by establishing a discursive relationship between archives and data, and its operationalization through the application of computational methods to support archival arrangement and description.[10] Certainly digital archives are comprised of digital data, as is evident to those who are charged with their preservation. But data is a slippery term, which also signifies an amenability to computation for analytical purposes. The proposition that digital archives should be imagined and interacted with as data – that is, as computationally tractable – is a more recent phenomenon in the archival field. Examples of digital archives being discussed as data increasingly abound, from the Library of Congress "Collections as Data" meetings (2016 and 2017) to several posts on the Society of American Archivists' *Bloggers!* website and numerous conference papers within the information sciences.[11] What I will foreground is the notion that digital archives are not already data by virtue of being digital but *become* data – are datafied – through the various acts of preparing them for manipulation by computational means.[12] When an instance of an electronic record is

10 Throughout the article, I use both *data* and *big data* as the context warrants, with an acknowledgement that the two terms are not semantically identical. They do, however, overlap considerably and for the purposes of my argument are indistinguishable, aside from the scale that *big data* more precisely expresses. In either case, I am referring to data or big data as cultural constructs – amalgams of meanings and associations – rather than as specific (albeit abstract) things.

11 For an illustrative sample, see Laurie Allen and Stewart Varner, "Archival Collections as Data for Digital Scholarship," *Bloggers!* (blog), 19 December 2017, accessed 28 January 2018, https://saaers.wordpress.com/2017/12/19/archival-collections-as-data-for-digital-scholarship/; Glen McAninch, "Big Data and Big Challenges for Archives," *Bloggers!* (blog), 5 March 2015, accessed 28 January 2018, https://saaers.wordpress.com/2015/03/05/big-data-and-big-challenges-for-archives/; Maria Esteva, Jeffrey Felix Tang, Weijia Xu, and Karthik Anantha Padmanabhan, "Data Mining for 'Big Archives' Analysis: A Case Study," *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–10; Luis Martinez-Uribe, "Digital Archives as Big Data," *Mathematical Population Studies*, 16 February 2018, 1–11, doi:10.1080/08898480.2017.1418116. The plenary session at the 2018 Association for Canadian Archivists' conference, "New Modalities of Archival Exploration," similarly took up the datafication of archives. A recent article in *Archivaria* also tackles the theme: see Michael Moss, David Thomas, and Tim Gollins, "The Reconfiguration of the Archive as Data to Be Mined," *Archivaria* 86 (Fall 2018): 118–51.

12 "To datafy a phenomenon is to put it in a quantified format so that it can be tabulated and analyzed." Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston, MA: Houghton Mifflin Harcourt, 2013), 78. Mayer-Schönberger and Cukier make an important distinction between datafication and *digitization*, which they describe as the process of converting analog information into a format that can be reproduced by a computer. Digitized materials may become datafied

reproduced on a screen, or when steps are taken to assure its long-term accessibility, it remains a record – that is, intelligible to the duelling archival teleologies of evidence and memory. The use of digital archives by archivists as a data set to support computationally enabled modes of discovery or by archival users for another computational purpose, on the other hand, represents a paradigmatic shift.[13] Datafied, digital archives exist within but also beyond evidence and memory;[14] I am suggesting, then, that *archives as data* may be poised to emerge as a fifth archival paradigm, beyond the four originally proposed by Terry Cook, and that the ideological dimensions of an archives-as-data paradigm can be unpacked and scrutinized even in its formative stages.

Tracing the development of archival thinking and practice over the past 150 years in his "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms," Cook identifies evidence, memory, identity, and community as evolving and historically situated "ways of imagining archives and archiving," which find their expression in the ways archivists interpret their roles and responsibilities.[15] The evidence paradigm, for example – which Cook summarizes as "the custodian-archivist guards the juridical legacy" – constitutes an identity framework distinct from memory, wherein "the historian-archivist selects the archive."[16] The various mindsets, as Cook eloquently outlines in his article, are products of the historical, social, and material conditions of the eras to which he assigns them, but they are neither bound by chronology nor mutually exclusive. So while the archival profession may be headed towards an archives-as-community paradigm, as Cook suggested in 2013, it could conceivably accommodate an

when, for example, optical character recognition is used to convert image data into machine-readable text that can then be parsed to determine how frequently a given word appears within it.

13 To borrow an analogy from the digital humanities, approaching a literary text as data – in order to visualize which words appear most frequently within it, for example – presumes an analytical stance that is distinct from that taken in approaching it as a narrative.

14 Central to the four paradigms that Cook identifies is the "creative tension" between memory and evidence; even in later paradigms, evidence and memory still occupy a prominent role in scaffolding archival identity as "each era interprets anew evidence and memory." See Terry Cook, "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms," *Archival Science* 13, no. 2–3 (2013): 102, 118.

15 Cook, "Evidence, Memory, Identity, and Community," 97. Cook comments on his reticence to use the word *paradigm* and suggests that a more apt term would be *frameworks* or *mindsets*.

16 Ibid., 106–107.

alternative – if overlapping – identity framework.[17] While archives-as-data is perhaps less rhetorically attractive than an archives-as-community, it better encapsulates the reframing of digital archives as data and the attendant implications for archival identity; that is, supporting community-oriented goals is not an inherent feature of datafication, though datafied digital archives may be used for community-building purposes consistent with Cook's paradigm.[18] But if *archives as data* does indeed describe a surfacing mindset within the archival profession, what roles do archivists and other recordkeepers play within it? What are their responsibilities? How do they imagine, work with, and make digital archives accessible for use by others?

To explore the kinds of transformations suggested by an archives-as-data paradigm and its possible consequences for archivy, I begin by taking up a pair of connotative meanings associated with data to demonstrate how they may work conservatively to reconnect digital archives with past interpretations of the archival role – as revenant myths, of sorts. I then focus on natural language processing, a machine learning approach used in the arrangement and description of digital archives, to highlight the invisibilization of human decision-making within the often-opaque interfaces of computational tools, which enables users to view them as neutral rather than historically determined. While I am discussing the concept of datafied archives and its practical application separately, they should be understood as jointly constitutive. I close by offering several possible sites through which to resist the reactionary tendencies of an archives-as-data paradigm, making recommendations for the emerging (inter) discipline of computational archival science, and proposing directions in which an archives-as-data paradigm might be further elaborated. Though there are many opportunities for opening up and enriching access to digital archives through computational operations once they are datafied, the same techniques and ways of thinking also bring their own histories and disciplinary assumptions;

17 Cook himself notes, "Of course, there is overlap. Strands from all four mindsets are interwoven. This discussion is about emphasis, not rigid definition" (ibid., 117).

18 Many computational archival science projects exist that demonstrate the application of computational methods to digital archives in ways consistent with an archives-as-community paradigm; these include Mapping Inequality: Redlining in the US and Human Face of Big Data, undertaken by the Digital Curation Innovation Center at the University of Maryland. See "Projects," *Digital Curation Innovation Center* (blog), accessed 30 March 2018, http://dcic.umd.edu/research/projects/. Nonetheless, an archives-as-data paradigm premised upon using computational thinking and methods with digital archives does not intrinsically embody the same principles as an archives-as-community paradigm.

so while datafication may enable new avenues for considering and working with archives, it can simultaneously foreclose on others. Most notably, as computational methods begin to constitute a larger part of archival work, the efforts of the archival profession over the past three decades to pluralize the archival endeavour and to introduce a social justice orientation – incorporating critical race theory, feminist theory, queer theory, and other overtly politicized modes of inquiry – into archivy may be stifled or even undone. If an archives-as-data paradigm is genuinely taking shape, then the archival profession has a responsibility to ensure that a social justice critique is maintained within it.

Imagining Archives as Data: Revenant Myths from an Archival Past

To see collections as data begins with reframing all digital objects as data. Data are defined as ordered information, stored digitally, that are amenable to computation. Wax cylinders, reel to reel tape, vellum manuscripts, websites, masterworks, musical scores, social media, code and software in digital collections are being brought onto the same field of consideration.[19]

What are archives becoming when they are reframed as data? Padilla's definition of data – as "ordered information, stored digitally, that are amenable to computation" – is both succinct and helpful, but it conceals a vast trove of historically contingent meaning in its simplicity.[20] He hints at the conceptual homogenization taking place when collections become data: what was once disparate is now made up of the same stuff. But what of the stuff? In the section that follows, I will elaborate on data as a discursive construct – not what it is, but how it is imagined and talked about. From a veritable constellation of meanings, I shall draw on two in particular: the positioning of data in relation to neutrality and objectivity and

19 Thomas Padilla, "On a Collections as Data Imperative," in *Collections as Data: Stewardship and Use Models to Enhance Access* (paper presented at Library of Congress Digital Preservation Meeting, Library of Congress, Washington, DC, 27 September 2016), 1, accessed 30 March 2018, http://digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf. Padilla's definition of data as "being amenable to computation" parallels Mayer-Schönberger and Cukier's distinction between datafying and digitizing.

20 That said, Padilla's report clearly acknowledges the ethical obligations of approaching collections as data and exemplifies the crossing of a transdisciplinary boundary between social justice and technology. Expanding on the various connotations of data is simply outside of the scope of the report.

the framing of data as raw material. Both are particularly relevant to conceptualizing digital archives as data – as opposed to collections as data more generally – because of their potential to reinvigorate deeply entrenched beliefs regarding the impartiality of the archivist. In an archives-as-data paradigm, then, these meanings could have a profound influence over the ways archival professionals view their roles and responsibilities and the ways they communicate their work to the public.

Though data, as a concept, has become nearly inseparable from the digital environment, the term itself was in circulation long before the invention of the modern computer. Daniel Rosenberg, in his early history of the concept of data, notes its first recorded appearance in the English language during the 17th century.[21] Initially used in the domains of mathematics and theology to indicate respectively facts or principles already given in an equation or argument, both the context and meaning of *data* had shifted by the end of the 18th century. More commonly used at the time in empirical disciplines like medicine, finance, natural history, and geography, the term had come to refer to "facts in evidence determined by experiment, experience or collection," a development upon which – according to Rosenberg – our current understanding of data as information in numerical form relies.[22] The redefinition of data as what is sought through scientific inquiry may also have provided the necessary conditions for its connections to positivistic notions of neutrality and objectivity, which I will take up shortly. Following its semantic stabilization in the late 1700s, the term *data* underwent a period of cultural latency until it resurfaced in the 20th century; though it was by then "a well-established concept . . . it remained largely without connotative baggage," allowing the term to accumulate new associations while retaining its underlying meanings from the 18th century.[23] Although Rosenberg's analysis regrettably ends before the present day, it nonetheless situates the concept within an Enlightenment-era empiricist epistemological milieu.

To unravel a more recent cultural conception of data, the first strand to tease out concerns its relationship to objectivity and neutrality. The phenomenon of

21 The etymological roots of the term *data* can be traced to the Classical Latin word *datum* (also used as the singular of data in English), which is derived from the verb *dare* (to give). See Daniel Rosenberg, "Data before the Fact," in *"Raw Data" Is an Oxymoron*, ed. Lisa Gitelman, Infrastructures Series (Cambridge, MA: MIT Press, 2013), 18: "A 'datum' in English . . . is something given in an argument, something taken for granted."

22 Ibid., 32–33.

23 Ibid., 34.

"Big Data" is defined by boyd and Jackson as an interplay of technology, analysis, and mythology, where the latter is summed up as "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity and accuracy."[24] For archival scholars and practitioners who have worked assiduously to denaturalize and challenge the assumption that the archivist is a neutral guardian of records, imagining and working with archives as data may give cause to be wary about its resurgence: big data, boyd and Crawford maintain, "offers the humanistic disciplines a new way to claim the status of quantitative science and objective method."[25] The title of a *Bloggers!* blog post on the use of named-entity recognition systems for archival processing, "Let the Entities Describe Themselves," conjures up a persistent image of the impartial craftsperson addressed by Duff and Harris in "Stories and Names: Archival Description as Narrating Records and Constructing Meanings" – now aided in descriptive tasks by a kindred, disinterested algorithm.[26] And yet neither the archivist nor the algorithm are without histories, assumptions, or biases shaped by the structures of unequal power relations; both are engaged in description as an interpretive act. Reframing digital archives as data, then, may allow claims of objectivity or neutrality to be reasserted in datafied contexts and to work conservatively to undermine settled debates about the extent to which recordkeepers influence users' encounters with archives.[27]

A second, related thread – also drawing on the conceptual legacies of positivism, scientific progress, and natural history – is the discursive construction of data as raw material. The characterization of data as a (natural) resource

24 boyd and Crawford, "Critical Questions for Big Data," 663.

25 Ibid., 667.

26 Josh Schneider and Peter Chan, "Let the Entities Describe Themselves," *Bloggers!* (blog), 3 May 2016, accessed 9 February 2018, https://saaers.wordpress.com/2016/05/03/let-the-entities-describe-themselves/; Wendy M. Duff and Verne Harris, "Stories and Names: Archival Description as Narrating Records and Constructing Meanings," *Archival Science* 2, no. 3–4 (2002): 264. Duff and Harris note the resilience of the belief that "the archivist's role in relation to records is to reveal their meaning and significance – not to participate in the construction of meanings."

27 Randall Jimerson makes a well-reasoned case for differentiating between objectivity and neutrality, but I draw here on Gitelman and Jackson's definition of objectivity as "the abnegation, neutrality, or irrelevance of the observing self," which is "situated and historically specific; it comes from somewhere and is the result of ongoing changes to the conditions of inquiry." See Randall Jimerson, "Archives for All: Professional Responsibility and Social Justice," *American Archivist* 70, no. 2 (2007): 270–72; Lisa Gitelman and Virginia Jackson, "Introduction," in Gitelman, ed., *"Raw Data" Is an Oxymoron*, 4.

for exploitation is metaphorically borne out in the use of language within the discipline, through references to *entity extraction* and *data mining*, and more explicitly in the expression *raw data* itself.[28] Geoffrey Bowker expands on the notion of raw data by invoking Lévi-Strauss's use of the term *raw* in *The Raw and the Cooked*, where the raw is mapped onto the natural world in binary opposition to the cooked/social.[29] Bowker, who first asserted that "'Raw Data' is an oxymoron," goes on to clarify that data are never raw; like archival description and arrangement, they are always representational and the result of interpretive acts.[30] The implications for conceptualizing archives as data again suggest a reactionary return to earlier ideologies: the notion of data as raw material parallels what Lara Moore identifies as the organicist ideas that shaped 19th-century French archival theory around arrangement, wherein the fonds was thought to contain a "natural" order that the archivist simply discerned or restored.[31] In Muller, Feith, and Fruin's 1898 *Manual for the Arrangement and Description of Archives*, the order of the fonds is similarly likened to the skeleton of "a prehistoric animal," with the archivist playing the role of the paleontologist who reassembles it.[32] Cook also notes the influence of Darwinian metaphors on 19th-century characterizations of the principles of provenance and respect des fonds, in references to the "*natural* accumulation of records" and the "*organic* character of archives."[33] Reinforcing the conceptual relationship between data and objectivity, an archives-as-data-as-raw-material frame may likewise serve to minimize the archivist's agency in arrangement and description: rather than constructing meaning, the archivist merely discovers what is innate in the

28 To be clear, I am not suggesting that such language should be avoided – it is the terminology of the discipline and necessary for communicating with other practitioners – just that it is also important for fraught terms not to become naturalized so as to appear invisible and elude critical analysis.

29 Geoffrey C. Bowker, "Data Flakes: An Afterword to *'Raw Data' Is an Oxymoron*," in Gitelman, ed., *"Raw Data" Is an Oxymoron*, 168.

30 Ibid.

31 Lara Jennifer Moore, "Putting the Past in Order: Archival Classification and the Ecole Des Chartes in the Late July Monarchy," in *Restoring Order: The École Des Chartes and the Organization of Archives and Libraries in France, 1820–1870* (Duluth, MN: Litwin Books, 2008), 120–21.

32 Samuel Müller, Johan Adriaan Feith, and R. Fruin, "Chapter II: The Arrangement of Archival Documents," in *Manual for the Arrangement and Description of Archives* (Chicago, IL: Society of American Archivists, 2003), 69–71. Thanks and acknowledgement to Jennifer Douglas for drawing attention to and elaborating the organicist metaphor in historical texts on archival arrangement.

33 Cook, "Evidence, Memory, Identity, and Community," 103 (emphasis in original).

records. Taken together, these two associations – with objectivity and with raw material – have considerable potential to reshape how archivists envision and describe the work they perform.

First across the Devil's Bridge? The Use of Natural Language Processing in Archival Arrangement and Description

Devil's Bridge mythologies are based in part on the conviction that human hands alone could not possibly have undertaken such state-of-the-art technical structures and that their very accomplishment must have exceeded human capacity at a number of levels.[34]

Verhoeven anchors her sublimely poetic meditation on digital research infrastructure in the parable of the Devil's Bridge – a motif from European fairy tales of the Middle Ages – to evoke the wondrous qualities of new technologies that function to efface the material conditions of their production.[35] It also fittingly captures the use of computational methods in archival arrangement and description, afforded by the datafication of digital archives, as an expression of that infrastructure.[36] The demonstrated capacity of these tools to tackle difficult or labour-intensive problems for archivists – for example, by instantly creating item-level inventories or automatically generating access points based on the contents of the fonds – may lead us to marvel at the anonymous feats of computational ingenuity going on behind the interface. In an archival context, the human cunning is in remaining attentive to the snares inherent in such

34 Deb Verhoeven, "As Luck Would Have It: Serendipity and Solace in Digital Research Infrastructure," *Feminist Media Histories* 2, no. 1 (2016): 8.

35 Verhoeven summarizes the standard plot of the Devil's Bridge fairy tale, relating the "infrastructure predicament" of building an architecturally daunting bridge, as follows:

> The bridge builders face an insurmountable difficulty caused by undue urgency, a unique environmental challenge, or sheer technical ambition, which is solved after a serendipitous encounter with the Devil, who agrees to complete the bridge in exchange for the first soul to cross it. The stories always conclude with the Devil being cheated by a shrewd human who sends a hapless animal (usually a dog, goat, or rooster) across the span instead.

Ibid., 7.

36 In speaking of digital research infrastructure, Verhoeven refers primarily to the management of digital archives.

a pact: Devil's Bridges, numerous examples of which still exist throughout Europe, perform the sleight of hand of rendering the physical labour behind the structure invisible. A similar disappearing act takes place when the tools of big data become unmoored from the specifics of their origin stories: it is easy to forget that software and algorithms are designed by people and shaped in ways that are often hidden to the end-user.

The application of machine learning techniques to archival description offers an opportunity to support the claim that computational tools are far from neutral. Natural language processing (NLP), wherein computer algorithms are trained to identify and classify elements of natural language, has been used extensively in the field of digital humanities to analyze literary texts and historical documents. It is also beginning to be incorporated into archival arrangement and description practices through projects like the University of California, Berkeley's ArchExtract web application, the Interactive Topic Model and Metadata Visualization (TOME) project at the Georgia Institute of Technology, and fondz, "a command line tool for auto-generating an 'archival description' from a [digital preservation] bag or series of bags," developed by Ed Summers at the Maryland Institute for Technology in the Humanities (MITH).[37] Efforts are likewise being made to integrate NLP tools into open source projects like BitCurator and ePADD to support the processing of large-scale digital archives within the application environment. Using NLP, the contents of a machine-readable fonds can serve as a corpus, or body of text, for analysis; techniques like named-entity recognition – which are designed to distinguish proper names of people, organizations, and places that can then be used as candidates for access points – offer significant opportunities for archivists to provide increased access through more comprehensive description.[38]

The potential for NLP techniques to assist archivists in processing high volumes of born-digital materials – greatly enhancing the discoverability of their holdings – should, of course, be met with excitement. But as part of the

37 Mary W. Elings, "Using NLP to Support Dynamic Arrangement, Description, and Discovery of Born Digital Collections: The ArchExtract Experiment," *Bloggers!* (blog), 24 May 2016, accessed 4 February 2018, https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/; Ed Summers, "Fondz," GitHub, accessed 14 April 2018, https://github.com/edsu/fondz.

38 For an example of how named-entity recognition is being experimented with in archival description, see Elings, "Using NLP."

work of adapting or developing these tools within an archival setting, we must also acknowledge how they privilege certain records and records creators over others, and how assumptions based on race, gender, gender identity and expression, sexuality, ability, class, and other categories of difference may be encoded into their design and use. What may appear to be an apolitical act of computationally analyzing a collection of born-digital archives to automate archival arrangement and description may actually involve a complex series of interpretive decisions on the part of both the archivist and the developer(s) of the algorithms and tools. Though a more sustained critique is warranted, I will briefly outline the ways that acts of marginalization might operate in the processing of datafied digital archives.

Current NLP systems are not programmed with the ability to understand natural language; rather, they are "trained" on large data sets of sample text. The quality and accuracy of the results obtained, then, depend on what is contained in the system's training corpus and the instructions provided as part of its learning. Illustrating the "brittleness" of natural language processing systems – that is, their non-transferability to contexts other than those for which they have specifically been trained – Maciej Cegłowski explains:

> If you go to Google translate and paste in an Arabic-language article about terrorism or the war in Syria, you get something that reads like it was written by a native speaker of English. If you type in a kid's letter from camp, or an extract from a novel, the English text reads like it was written by the Frankenstein monster.
>
> This isn't because Google's algorithm is a gung-ho war machine, but reflects the corpus of data it was trained on.[39]

A lack of transparency regarding the system's training process may thus make it difficult for archivists to assess or critique what occurs between the input and output of data. Schneider and Chan's remark on the lack of "open source NER tools broadly tuned towards the diverse variety of other textual content collected and shared by cultural heritage institutions" may be read as a call for archival

39 Maciej Cegłowski, "Deep-Fried Data," *Idle Words* (blog), accessed 31 March 2018, http://idlewords.com/talks/deep_fried_data.htm.

professionals to become more involved in the development of these systems.[40] Otherwise, the use of out-of-the-box NLP systems is likely to privilege records that are more computationally legible to those particular systems – which is not, perhaps, an ideal criteria for providing access to archives.

To expand on Schneider and Chan's comment, if the training set is skewed towards textual content created primarily by white, English-speaking settler/colonizer authors, then there may be consequences for the ability of the system to work with and provide access to the materials of Indigenous or racialized creators. An obvious limitation is the Anglo-centric bias of many natural language processing tools, which marginalizes fonds or collections that consist of records created in languages other than English. Though not referring to natural language processing, Elvia Arroyo-Ramirez's account of processing the born-digital, Spanish-language records of Juan Gelman and of the politics of "cleaning" diacritics from data in order to render them computationally tractable is nonetheless an pertinent example.[41] It surfaces the multiple layers of hidden barriers that emerge as a result of both deliberate and inadvertent decisions made in the design of these systems: barriers to encoding languages other than English so that they can be adequately read by the computer, and barriers to training NLP systems on other languages' grammar, syntax, and semantics. While it is tempting, then, to focus on using computational methods to support the arrangement and description of digital collections that are easy to parse – the "quick wins" – it is also important to remember that it is not by accident that certain kinds of data transformations are easier than others. As the profession explores the opportunities in working with archives as data, a concerted effort to develop tools that can represent and provide access to a wider and more inclusive range of digital archives is also necessary.[42]

To return to Verhoeven's metaphor, the Devil's Bridge constitutes a Faustian bargain only if the material and intellectual contributions by humans to its

40 Schneider and Chan, "Let the Entities Describe Themselves." The authors, who are affiliated with the ePADD project, indicate that they are actively working on improving the browsing accuracy of the tool according to the needs of libraries, museums, and universities.

41 Elvia Arroyo-Ramirez, "Invisible Defaults and Perceived Limitations: Processing the Juan Gelman Files," *Medium* (blog), 30 October 2016, accessed 12 February 2018, https://medium.com/on-archivy/invisible-defaults-and-perceived-limitations-processing-the-juan-gelman-files-4187fdd36759.

42 The continued privileging of textual records over non-textual formats is another consideration in the use of natural language processing techniques.

construction go unrecognized, unacknowledged, and unconsidered. That is, the opacity of our interactions with the computational methods of datafied archives may risk perpetuating the very power differentials that the archival profession is slowly working towards addressing in its current practices and institutions. Just as the various processes of archival selection, arrangement, and description are naturalized and presented as the authoritative record rather than as one of myriad options, so too do computational methods appear, as Wendy Hui Kyong Chun suggests, to be "always already there" and not the outcome of a historically specific trajectory of development.[43] In the archival profession's efforts to render its practices more visible to the people who use or could use archives, leveraging computational methods for arrangement and descriptive tasks introduces another layer of abstraction that must be accounted for. To realize the access-enhancing and labour-saving potential of approaching archives as data, then, requires archival professionals to take an active role in supporting and becoming involved in the development of these systems, critically questioning their design and providing plain language explanations of what they do.[44]

Provocations: Critical Directions for Computational Archival Science

Those of us who are interested in seeing more robust cultural critique need to be more specific about where the intervention might most productively take place. It's not only about shifting the focus of projects so that they feature marginalized communities more prominently; it's about ripping apart and rebuilding the machinery of the archive and database so that it doesn't reproduce the logic that got us here in the first place.[45]

43 Wendy Hui Kyong Chun, *Programmed Visions: Software and Memory* (Cambridge, MA: MIT Press, 2011), 104–106.

44 Many of the NLP projects being developed for archival arrangement and description are the result of the dedicated efforts of a small number of people; in those under-resourced conditions, it is often challenging to translate principles into product – so an archivist-programmer building an arrangement and description tool may be considering the very questions I am raising here but also needs to produce results to fulfill grant obligations, etc.

45 Miriam Posner, "The Radical Potential of the Digital Humanities: The Most Challenging Computing Problem Is the Interrogation of Power," *LSE Impact Blog*, 12 August 2015, accessed 9 February 2018, http://blogs.lse.ac.uk/impactofsocialsciences/2015/08/12/the-radical-unrealized-potential-of-digital-humanities/.

There are undeniable benefits to automating descriptive tasks in order to increase the discoverability and accessibility of digital archives, and it is not my intent to suggest that archival professionals should not take advantage of the computational methods afforded by the datafication of archives. Nor do I mean to imply that there are not computing practitioners, within archival studies and more broadly, who approach their work critically and call out the tendencies of the discipline to cordon off intersectional critiques of its work – although I would argue that analyzing the larger systemic issues of power and privilege within computing is not a preoccupation within the field as a whole.[46] I also recognize that it is difficult for archival workers who are themselves disempowered by precarious labour contexts to raise questions about dominant narratives of technological progress in their workplace. I am not, then, writing out of a Luddite impulse to preserve archival practice in the amber of its pre-digital state; in fact, I am deeply committed to seeing the fulfillment of computational methods' emancipatory promises: to open up archives to a broader, more diverse audience and to ameliorate the conditions of the profession as a whole. However, to ensure such an outcome is achieved – and to avoid the soul-trapping fate alluded to in Verhoeven's Devil's Bridge metaphor – requires active participation and critical discourse.

One of the most conspicuous sites in which archives are being datafied and computational thinking and methods are being integrated into the archival profession is the emerging discipline of computational archival science (CAS), defined by its proponents as a

> transdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation,

46 Tara McPherson's analysis of the historical development of the UNIX operating system in the United States makes this very point with respect to the cultural interplay between the reductivist "rule of modularity" within UNIX philosophy and the encoding of new, more covert modes of racism in the organization of social life. See Tara McPherson, "U.S. Operating Systems at Mid-Century: The Intertwining of Race and UNIX," in *Race After the Internet*, ed. Lisa Nakamura and Peter A. Chow-White (New York: Routledge, 2012), 26–31.

> and access decisions. . . . This suggests that computational archival science is a blend of computational and archival thinking.[47]

Corralling various computational practices used with archival materials under a more formalized domain, CAS includes computational linguistics, graph analytics, computational finding aids, and digital curation within its scope. The authors deliberately frame their definition in nascent, evolving terms, offering an opportunity to include a broader interpretation of archival theory and practice within it.

As part of their rationale for a CAS transdiscipline, Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad appeal to the importance of examining the computational "theories and methods that dominate records practices."[48] The authors do not intend for the exchange of knowledge to be one way, however: "there is a need to fundamentally transform both disciplines in order to infuse archival theories, principles, and methods with the computational, and equally, *to infuse the computational with archival conceptualizations and theories of the record*."[49] In terms of potential contributions to computer science from archivy, they highlight the attention paid to the record and aggregates of the record, but we may also include the profession's deep appreciation for and emphasis on context – the contexts of records in relation to each other but also their broader societal context. Beyond rigorous theorizations of authenticity and provenance, then, mapping the conceptual terrain of CAS also involves honouring and incorporating the profession's engagement with questions of power, privilege, and oppression. CAS thus represents a crucial juncture for uniting two trajectories of archival theory and practice that have hitherto operated on separate planes, focusing on issues of social justice and technology, respectively. As the use of computational tools begins to shape a greater extent of archival work, it becomes

47 Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad, "Chapter 9: Archival Records and Training in the Age of Big Data," in *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, ed. Johnna Percell, Lindsay C. Sarin, Paul T. Jaeger, and John Carlo Bertot, Advances in Librarianship, vol. 44B (Bingley, UK: Emerald Publishing, 2018), 181.

48 Ibid.

49 Ibid., 181 (emphasis added). The authors clarify that they are not suggesting that archival science and computer science merge with each other, but that the transformation they are referring to takes place in the transdisciplinary space of CAS.

increasingly important for the two intellectual traditions to come together and recognize areas of mutual concern. The following sections propose a few sites of convergence.

Pluralizing Computational Archival Science

If both the archival and computational professions are characterized by an underrepresentation of Indigenous people and people of colour, then a transdiscipline that combines the two fields could conceivably exacerbate the excluded status of these groups. Although projects that endeavour to represent the interests of marginalized communities are also important, pluralizing CAS goes beyond this to involve members of those communities within the field as collaborators, principal investigators, and project leads.[50] The systemic nature of racialized barriers to participation does not preclude attempts to name and confront such barriers wherever possible;[51] for example, the code of conduct for the Collections as Data initiative, supported by the Institute of Library and Museum Services, makes the commitment to inclusivity explicit.[52] Pluralizing CAS also requires increasing access to computational tools, techniques, and knowledge beyond the academic environment as an extension of boyd and Crawford's argument about the concentration of data-wealth within the university system.[53]

Supporting and Encouraging the Continued Visibility of Decision-Making Processes in Computational Methods

Existing strengths in the computational domain are the extent of available open source software for archival processing, the standard practice of commenting

50 It is not, however, incumbent on professionals from marginalized communities to perform the lion's share of institutional critique, which should also be undertaken by colleagues with the privilege of not having their views attributed to their subject position.

51 The guide to identifying and dismantling white supremacy in archives developed by Michelle Caswell and her students offers a point of departure but may need to be further elaborated for a datafied archives/CAS context. See Michelle Caswell, "Teaching to Dismantle White Supremacy in Archives," *Library Quarterly* 87, no. 3 (2017): 222–35. The guide acknowledges that it is "an incomplete list."

52 "Always Already Computational," *Always Already Computational – Collections as Data* (blog), accessed 2 April 2018, https://collectionsasdata.github.io/.

53 boyd and Crawford, "Critical Questions for Big Data," 674: "It is also important to recognize that the class of the Big Data rich is reinforced through the university system: top-tier, well-resourced universities will be able to buy access to data, and students from the top universities are the ones most likely to be invited to work within large social media companies. Those from the periphery are less likely to get those invitations and develop their skills. The result is that the divisions between scholars will widen significantly."

code, and the culture of transparency evidenced in scholarly literature that, for example, includes the mathematical models for machine-learning algorithms. To question how a function or algorithm might be approached differently, however, more archival professionals will need to develop both an understanding of what is happening at the level of the code or the model and the ability to articulate computational concepts in an accessible manner as part of that critique, which we may take to be the objective of computational archival science.

Incorporating Data Provenance into Documentation Practices

Another modality through which to render the actions of the practitioner visible is data provenance, accounting for where the data came from and what actions have been performed on it. As part of their documentation practices – wherein archivists may record details about the particular order they used in arranging collections and their reasons for doing so – they can also indicate any data "cleaning" actions and other interpretive decisions they made about the use of computational tools, communicating these to people who are accessing the archives.

Reviewing Archival Codes of Ethics

It is also imperative to consider how datafied archives interface with the ethical standards of the profession within an archives-as-data paradigm. To return to the earlier theme of data as raw material, might such a conceptualization have consequences for the care of archives? Is there a potential for archives to be used as data sets in ways that are not in the best interests of their creators? How might our ethical standards need to change to address the use of datafied archives for different purposes?

Exploring Scholarly Collaborations with Allied Disciplines

An intersectional critique that would examine the computational thinking and methods undergirding an archives-as-data paradigm may require archival scholarship to extend beyond its current boundaries. Inasmuch as writing on the use of technology in archival practice has tended not to discuss questions of structural power and privilege, archival literature aligned with social justice issues has been equally remiss in its analysis of technology. We may then turn to allied disciplines like digital humanities, critical information studies, and more specialized fields like critical data studies, ancestral and decolonial computing,

and software studies to identify a body of literature to support computational archival science and theorize an archives-as-data paradigm.

Engaging critically with the concept and practice of archives as data – nudging our practices towards more pluralistic and inclusive outcomes – is not needed only in CAS, but CAS is a key site where it can occur. If an archives-as-data paradigm is taking shape, then it is in its initial stages, before practices and beliefs become too firmly entrenched, when transformative critique may be most achievable. As computational methods begin to inhabit a growing number of functions within archival work, the way we represent archives as data will have a significant impact both on how we encounter archives in our work and on how they are perceived by the people who use them.

Conclusion: Enriching an Archives-as-Data Paradigm

Paradigms can be destructive or enabling.. . . [W]e can view our paradigms and mythologies as bastions of identity, in which case we become defensive and they rigidly destructive, or we can see them as liberating, authorizing us to develop new directions in light of the astonishing challenges to archiving today from theory, technology, and society, and the expectations and demands each occasions.[54]

The tendency, when reframing archives as data, to re-inscribe earlier notions of neutrality and objectivity upon archival identity represents a small fragment of the larger archives-as-data paradigm. Clearly, much theorizing remains to be done, and the work that should be celebrated as exemplifying an enabling paradigm must be acknowledged. Padilla's politicized definition of a collections-as-data imperative, for instance, evidences a humanistic lens and underscores the importance of being "critically attuned to the possibilities *and* perils that come with [data's] use."[55] The computational thinking and methods that animate an archives-as-data paradigm – produced through a complex set of practices, discourses, and ideologies that are rarely apparent to the end-user – constitute another area that requires closer inspection. Likewise, there is ground yet to cover in articulating a more nuanced interpretation of the ways data is mobilized

54 Cook, "Evidence, Memory, Identity, and Community," 116–17.

55 Padilla, "On a Collections as Data Imperative," 2 (emphasis in original).

discursively in order to explore its conceptual cross-pollinations – emancipatory *and* restrictive – with archival identities. An archives-as-data paradigm has the capacity to be liberating in the sense that Cook intends: to help us imagine archives and archiving in new directions – directions that we ourselves can determine. As a profession, we can embrace the use of computational methods and tools without sequestering our practices or failing to consider how they are enmeshed in systemic structures of power and privilege; indeed, an archives-as-data paradigm is perhaps one of the most important sites where this work will be performed.

BIOGRAPHY Devon Mordell is the Digital Scholarship and Archiving Librarian at the University of Windsor Leddy Library, located on the traditional territory of the Three Fires Confederacy of First Nations, which includes the Ojibwa, the Odawa, and the Potawatomie. Prone to bouts of academic wanderlust, she holds a BFA in Visual Arts (University of Windsor), an MA in Cultural Studies and Critical Theory (McMaster University) and most recently, an MAS from the University of British Columbia. She is grateful to the incredible cohort of SLAIS graduate students and faculty with whom she had the privilege of brewing big ideas about archivy. Her research practice examines the use of digital technologies in providing access to archives through a critical lens with a utopian tint.