

Historiography's Two Voices: Data Infrastructure and History at Scale in the *Oxford Dictionary of National Biography* (ODNB)

Christopher N. Warren

11.22.18

Peer-Reviewed By: Matt Erlin

Clusters: Data, Infrastructure

Article DOI: 10.22148/16.028

Dataverse DOI: 10.7910/DVN/D3KFLP

Journal ISSN: 2371-4549

Cite: Christopher N. Warren, "Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB)," Journal of Cultural Analytics. November 22, 2018.

On its release in 2004, the *Oxford Dictionary of National Biography* was called "the greatest book ever" and "a more enthralling read than all the novels ever entered for the Booker Prize put together."¹² The tabloid *The Daily Mail*, where these giddy pronouncements appeared, is not known for understatement, but more cautious academic researchers have long held the ODNB in similarly high esteem. Stefan Collini, writing in the *London Review of Books*, found himself "experiencing a rare, and wholly unironic, feeling that mixes pride and humility with a dash of wonder" when he considered "generations to come making use of this vast consolidation of scholarly accuracy for purposes of their own which may be

barely imaginable to us now.”¹ Taking into account both the hardbound version and what most assume is its digital doppelgänger at oxforddnb.com, Noel Malcolm in *The Sunday Telegraph* called the ODNB “an astonishing piece of work: a colossal, beautiful, fully functional and utterly user-friendly engine of enlightenment.”² Reviewers’ initial responses—awe and astonishment—have, until recently, arguably been the responses most appropriate to the ODNB considered in its entirety. The enormous scope of ODNB, which is the work of roughly 10,000 scholars, runs to 60 volumes in print, and is made up of more than 62 million words, quickly defeats the capacities of even those most eager to praise it.

Awe and astonishment have been the most reasonable scholarly responses to the ODNB, that is, until the new possibilities afforded by several key fields that, in combination, form a critical engine suitable for this “engine of enlightenment.” In what follows, I combine insights from information history, digital history, sociology of knowledge, media archeology, history of archives, distant reading, and data visualization to gain further purchase on the ODNB. These fields—which for shorthand I’ll call “digital humanities”—mitigate some of the challenges of studying the ODNB as such. As another early reviewer complained, “If you were to read one life in the new DNB every day you would take 137 years to finish it. So reviewing it is like exploring a continent by rowing boat.”³ However, the reasons to dedicate critical and computational power to study the ODNB in this way are not limited to the slightly imperialist-sounding work of “exploring a continent.” As former ODNB Senior Research and Publication editor Philip Carter has argued, we now have “the ability to use national biographies both as written collections and as data to make connections and trace patterns that could not be identified without the existence of collective biography in digital form.”⁴ Some historical trends and latent ideologies, in other words, only become visible by reading historiography at scale.

¹Stefan Collini, “Our Island Story,” *London Review of Books*, January 20, 2005; Stefan Collini, “National Lives: The Oxford Dictionary of National Biography,” in *Common Reading: Critics, Historians, Publics* (Oxford: Oxford University Press, 2008), 315.

²Noel Malcolm, “Not Quite Your Average Book: The Newly Updated Dictionary of National Biography Runs to 62 Million Words and, Says Noel Malcolm, Is a Thing of Wonder,” *The Sunday Telegraph*, October 17, 2004, sec. Review.

³Christopher Howse, “Robin Hood Made Real: The Gargantuan New Dictionary of National Biography Explodes Some Myths, and Gives Fresh Life to Others,” *Daily Telegraph*, September 25, 2004, sec. Books.

⁴Philip Carter, “What Is National Biography for? Dictionaries and Digital History,” in *True Biographies of Nations? Cultural Journeys of Dictionaries of National Biography*, ed. Karen Knox (ANU Press, forthcoming), 82-103.

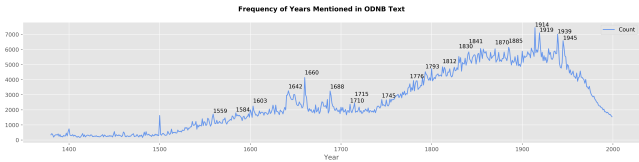


Figure 1. Frequency of Years Mentioned in *ODNB* Text.

Consider Figure 1, which shows how frequently each year is mentioned in the full text of the *ODNB*. This data alone tells us about the outsized importance of certain dates in Britain’s historiographical imaginary. Dates mentioned more frequently that seem to rise above their immediate context include 1603 (accession of James VI and I), 1642 (Civil War and Revolution), 1660 (Restoration), 1688 (Glorious Revolution), and 1793 (the French Reign of Terror). Data of this kind represent what I’ll consider the first voice of the *ODNB* at scale—pertinent and possibly valuable indexes for the past as such. This we might call the Rankean voice, the voice that speaks from the totality of biographies about the past *wie es eigentlich gewesen*, as it really happened. From historiography’s Rankean voice, we learn, for example, that the mothers of *ODNB* subjects have most frequently been actresses, teachers, and noblewomen while their fathers have been landowners, army officers, clergymen, or merchants (Figure 2). We learn too that 49 *ODNB* subjects were born in Hungary and died in England and that 1914, 1919, 1939, and 1945 were highly significant years in the lives of Britons.

Rank	Women’s Mothers	Men’s Mothers	Men’s Fathers	Women’s Fathers
1	actress	gentlewoman	landowner	landowner
2	noblewoman	noblewoman	Church of England clergyman	army officer
3	teacher	schoolteacher	merchant	Church of England clergyman
4	schoolteacher	teacher	army officer	merchant
5	gentlewoman	actress	farmer	solicitor
6	dressmaker	domestic servant	gentleman	gentleman
7	queen	nurse	solicitor	farmer
8	singer	farmer	naval officer	naval officer
9	nurse	schoolmistress	Church of England priest	physician
10	courtier	shopkeeper	surgeon	barrister

Figure 2.

But once we notice that some years are mentioned more frequently than others—that 1914 is the year mentioned most often in the *ODNB* overall, that nine of the top ten years mentioned most frequently are in the 20th century, and that women’s mothers are quite frequently queens (!)—we can begin to anticipate my main argument, which is that historiography of the kind represented by the *ODNB* in aggregate speaks with a double voice. The *ODNB* cannot help but give us information about things as they happened—who can deny that 1914 was a significant year?—but that testimony is accompanied at every point with equally

pertinent testimony concerning the ODNB's own contingent making.

It is this second voice—the testimony about how a key piece of our historiographical infrastructure got made—that may make an article such as this most urgent, particularly as scholars across the humanities grapple with the effects and implications of big data. Beyond the many troubling cases of mass data collection and algorithmic bias, the digital revolution has profoundly changed the way scholars find and deploy evidence.⁵ Scholars regularly query massive digital archives using pre-defined search terms and have, in many cases, received field-changing evidence for very little cost in time or effort.⁶ Among early reviewers of the ODNB, it became something of a set piece to narrate one's searches of the online edition's "ocean of data."⁷ To illustrate that the ODNB could be "searched and cross-referenced in an exciting variety of ways," Piers Brendon in *The Independent* advised, "Tap in the word 'moustache', and you get 390 entries."⁸ Noel Malcolm found that "entering versions of the phrase 'did not suffer fools gladly' produces no fewer than 88 people."⁹ Kevin Whelan informed readers of *History Ireland* that "'Fenian' will throw up 150 entries, and 'bastard' generates 227."¹⁰ In his Leslie Stephen special lecture of 2004, Sir Keith Thomas observed, "It is hard to think of any aspect of the British past which will not be illuminated by running a word search of this colossal database."¹¹

But the very databases, data structures, and algorithms that produce these useful query results remain for many scholars an unexamined black box beyond the reach of critical analysis. Contemporary scholars—ever-more reliant on digital prosthetics—have a pressing if rarely acknowledged need to know more about

⁵ Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2016); Mirko Tobias Schäfer and Karin Van Es, eds., *The Datafied Society: Studying Culture through Data* (Amsterdam: Amsterdam University Press, 2017); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York, NY: St. Martin's Press, 2018); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).

⁶ I refer to Putnam's observation that "technology radically reduced the cost of discovering information about people, places, and processes outside the borders of one's prior knowledge." Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *The American Historical Review* 121, no. 2 (April 1, 2016): 383.

⁷ Kevin Whelan, "Surfing an Ocean of Data: Oxford Dictionary of National Biography," *History Ireland* 13, no. 3 (2005): 42-47.

⁸ Piers Brendon, "The Oxford Dictionary of National Biography: Fresh Air in a Hall of Fame," *The Independent*, September 24, 2004.

⁹ Malcolm, "Not Quite Your Average Book."

¹⁰ Whelan, "Surfing an Ocean of Data," 47.

¹¹ Keith Thomas, *Changing Conceptions of National Biography: The Oxford DNB in Historical Perspective*, The Leslie Stephen Lecture (Cambridge: Cambridge University Press, 2005), 35.

the digital infrastructure increasingly unpinning humanistic scholarship. This is where media archeology, digital history, and the history of information have important roles to play. As scholars including Ted Underwood, Jim Mussell, and Katherine Bode have argued, humanists are often guilty of taking their digital searches for granted, with little knowledge of or interest in the ways digital archives have been composed.¹² All too often, the “editorial and bibliographical consciousness” that characterizes much humanistic scholarship falters in the face of data-rich digital objects.¹³ Matthew Kirschenbaum has written on what he calls a “medial ideology of electronic text” that in treating digital objects adopts misleading metaphors of “light, reason, and energy unleashed in the electric empyrean” at the expense of the material realities of “inscription, mechanism, sweat of the brow...and cramp of the hand.”¹⁴ For a fuller understanding of the ODNB’s role in knowledge-making, it takes an approach capable of considering the ODNB as data *and* as historical artifact *at the same time*. In her study of ProQuest’s *Early English Books Online*, Bonnie Mak shows how “infrastructures of knowledge-making” are quietly effaced by digital publishers.¹⁵ Even though search results are almost always a function of what Mak calls “intersecting temporalities,” in which older media have been transformed or remediated into machine-readable data, such transformations often escape scrutiny or comment. “[D]igital information provenance does not tend to feature in historiography discussions,” Toni Weller observes.¹⁶ Datasets and digital infrastructures tend to be considered primarily as conditions of possibility for historical analysis rather than objects potentially subject to analysis in their own right, and as such they occupy a status not unlike that of *data* in its premodern, Latin sense—as something *given*, a precondition, something prior to controversy and thus uncontroversially understood.¹⁷ In reality, however, digital interfaces that present some things and omit others should challenge the presumed stability of “the archive” and prompt further inquiry into the data at their source. Any researcher making claims on the basis of an archive like Google Books, for example, should be attuned to the ways certain types of materials were systematically excluded

¹²Ted Underwood, “Theorizing Research Practices We Forgot to Theorize Twenty Years Ago,” *Representations* 127, no. 1 (August 1, 2014): 64-72; Jim Mussell, “Doing and Making: History as Digital Practice,” in *History in the Digital Age*, ed. Toni Weller (Routledge, 2013), 79-94.

¹³Katherine Bode, “The Equivalence of ‘Close’ And ‘Distant’ Reading: Or, toward a New Object for Data-Rich Literary History,” *Modern Language Quarterly* 78, no. 1 (March 1, 2017): 94.

¹⁴Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge, MA: MIT Press, 2008), 39.

¹⁵Bonnie Mak, “Archaeology of a Digitization,” *Journal of the Association for Information Science and Technology* 65, no. 8 (August 1, 2014): 1519.

¹⁶Toni Weller, “Introduction,” in *History in the Digital Age*, ed. Toni Weller (Routledge, 2013), 11.

¹⁷Mak, “Archaeology of a Digitization”; Daniel Rosenberg, “Data before the Fact,” in *Raw Data: Is an Oxymoron*, ed. Lisa Gitelman (Cambridge: The MIT Press, 2013), 15-40.

from the scanning process such as very large, very small, or very brittle books or books with tight bindings or foldout maps.¹⁸ So too with the *ODNB*.

Most researchers consulting a given *ODNB* entry have been dividing, as it were, without a denominator. How typical or unusual is this entry? Where does it fit into the *ODNB*'s broader patterns of representation? It should matter, for example, that editors declared early on that no biography would surpass 15,000 words "without special case being made" but currently 72 biographies are longer than 15,000 words, and 15,000 is actually lower than the mean for rulers, royalty, and aristocracy.¹⁹ Scholars should be aware that among more typical professions—or at least typical for the *ODNB*—magnates, air force officers, and philosophers get the lengthiest treatments. Nor should, say, historians of antiquarianism, religious historians, or art historians remain unaware that Antiquaries, Jesuits, and engravers get shorter shrift. Analysis at this scale, then, expands the affordances of biography to tell us something about what scholars have valued when they wrote about particular kinds of figures. To take another example, all activities are situated somewhere but locations are more fundamental to certain kinds of biographical writing than to others. While all the *ODNB*'s musicians, engravers, actors, and journalists did their work in some places and not others, their biographies nevertheless privilege people far more than places (Figure 3).²⁰ And while biographies of explorers and spies unsurprisingly mention foreign places more frequently than domestic places, it may be more notable that science, business, and scholarship further fill out the international end of the spectrum while agriculture, royalty, and religion stay closest to home (Figure 4). Such features of the different kinds of biographies are interesting enough in their own right and can help readers contextualize any given biography, but—critically—they also point toward the infrastructure that makes such knowledge possible.

¹⁸Melissa K. Chalmers and Paul N. Edwards, "Producing 'One Vast Index': Google Book Search as an Algorithmic System," *Big Data & Society* 4, no. 2 (December 1, 2017): 6.

¹⁹H. C. G. Matthew, *Leslie Stephen and the New Dictionary of National Biography* (Cambridge; New York: Cambridge University Press, 1997), 25.

²⁰On the other side of the spectrum, the lives of air force officers, geologists, and building engineers are constituted by locations, at the expense of mentioning names.



Figure 3 and 4.

“Historians,” Mussell argues, are obliged to “account for the transformation of the evidence base in their analysis, and this necessitates understanding the methodologies and technologies responsible for these transformations.”²¹ Mussell proposes “Using digital resources against the grain of their interfaces in order to access the data they contain. It is a shift that depends on defamiliarization, on recognizing what is distinct about digital media and technologies and then exploiting this difference for scholarly ends.”²² Behind the *ODNB* is a deep substrate that includes a vast data management system, highly detailed SGML markup conventions, extensive international labor, and the enormous cultural weight of the Victorian-era *DNB*. I argue that it is only by investigating components of our historiographical infrastructure like the *ODNB* (1.) *in their entirety* and (2.) as *historically contingent digital artifacts* that we can fully access their double voice, and it is only then that we can understand the knowledge they make available and the knowledge-making they constrain.

★ ★

²¹Mussell, “Doing and Making,” 87.
²²Mussell, 81.

From Paper Knowledge to Data

With my argument thus laid out, I turn now to the *ODNB*, whose editor, Colin Matthew, it is said, “seized on computer-compilation with an almost apocalyptic fervour.”²³ In reality, much of the digital infrastructure underpinning the *ODNB* was conceived and overseen by OUP’s Robert Faber and Rosemary Roberts while Matthew himself admitted that the editorial process—“fully computerized from the start”—required his own “re-tooling.”²⁴ But the point remains that creating the *ODNB* was an experience, in onetime *ODNB* data manager Rupert Mann’s words, “of converting information into data.”²⁵ The many necessary transformations make the *ODNB* online a perfect exemplar of Mak’s “intersecting temporalities” and offer a window into the politics—and geo-politics—of scholarly data at the turn of the 21st-century. Like the print edition, the *ODNB* bears the marks of Leslie Stephen’s original *Dictionary of National Biography*, begun in 1882 at the behest of publisher George Smith, and finally published in revised form in 1908 by Stephen’s successor Sidney Lee. “Along with the *Oxford English dictionary*, the *Encyclopaedia Britannica*, *Grove’s dictionary*, *Fowler’s modern English usage*, *Cruden’s concordance*, and the annual volumes of *Who’s who*, *Burke’s peerage*, *Crockford* (and, for some, *Wisden*),” James Raven has written, “the *DNB* formed part of the scaffolding of civic knowledge in Britain.”²⁶ It also formed the scaffolding of the *ODNB*. After Colin Matthew was appointed in 1992 as editor of what was then to be called the “New Dictionary of National Biography,” Matthew decided that this major collaboration between Oxford University Press and the British Academy would not start wholly from scratch but instead include all subjects already included in the original *DNB*. Entries might have to be revised or rewritten, but if someone merited inclusion in the Victorian-age *DNB*, he or she would appear in the New *DNB*. The decision was then made to key in the entire text of the 1908 *DNB* and its 20th-century supplements.

The story of the *DNB*’s 1990’s transformation from dictionary to digital data is one not just of intersecting temporalities, but of intersecting geographies as well. Critics have charged that Colin Matthew and his successor, Brian Harrison, were able to publish the *ODNB* within the relatively short span of twelve years only by employing an overly mechanistic “engine of compilation,” but when an unnamed

²³Nicolas Barker, “The Biographers’ Tales: Who’s in, Who’s Out and Who’s Writing,” *The Times Literary Supplement*, no. 5306 (December 10, 2004): 7.

²⁴Matthew, *Leslie Stephen and the New Dictionary of National Biography*, 33–34.

²⁵Rupert Mann, “Searching the Oxford Dictionary of National Biography,” *The Indexer* 25, no. 1 (April 1, 2006): 17.

²⁶James Raven, “The Oxford Dictionary Of National Biography: Dictionary Or Encyclopaedia?,” *The Historical Journal* 50, no. 4 (December 2007): 993.

referent in this process is low-cost Indian labor, dehumanizing metaphors are especially ill-advised.²⁷ To “capture” the Victorian *DNB*’s roughly 38,000 subjects and 33 million words—“capture” was the project’s internal term—OUP turned to a company in Pondicherry, India, the Alliance Photosetting Company, whose contract work keying in the *DNB* in the short term allowed Harrison’s team to assess the *DNB*’s broadest trends of inclusion and exclusion and provided the base text for contributors revising outdated *DNB* bios.²⁸ In the longer term, digitizing the *DNB* permitted the *ODNB* to add the 1908 *DNB* biographies to its website, where they can still be found and compared to their 2004 successors.²⁹ The Alliance Photosetting Company headquartered in the former French-colonial outpost of Pondicherry is in fact something of a hero in the digital history of the *ODNB*: its employees’ work transforming documents into data also stands behind every search for a person on the *ODNB* website and behind the meticulously-tagged markup of the full *ODNB* text. With each of their entries, *ODNB* contributors were required to complete a so-called “Profile” or “Information Sheet,” which concisely summarized details such as variant names, aristocratic titles, sex, dates of birth and/or baptism, parents’ names and occupations, education, religious affiliations, and geographic and cultural associations. Often completed on distinctive green paper, these questionnaires (see Figure 5) were meant to capture “events and experiences... that... would be shared by most of the subjects in the dictionary.”³⁰ Contributors were advised, “Frequently in literary memoirs—often for excellent reasons of style and conciseness—facts are given allusively or collectively in a way unsuitable for computer searching. You are therefore asked to itemize some of the factual basis of your article on this sheet.”³¹ The sheet’s facts were intended for digitization, but only at the appropriate time, for the process depended upon the affordances of variously colored paper and pens at different stages: “Do *not* send your article or accompanying sheets by fax; the original paper copy is needed for editing,” contributors were warned.³²

²⁷ Barker, “The Biographers’ Tales,” 7.

²⁸ Brian Harrison, “Introduction,” in *The Oxford Dictionary of National Biography: From the Earliest Times to the Year 2000*, ed. H. C. G. Matthew and Brian Harrison, vol. 1 (Oxford: Oxford University Press, in association with The British Academy, 2004), xv.

²⁹ Matthew, *Leslie Stephen and the New Dictionary of National Biography*, 34.

³⁰ Mann, “Searching the Oxford Dictionary of National Biography,” 17. “If you are using the paper version of this form, please write legibly OR type your responses and paste them into the relevant section,” the form instructed.

³¹ *New Dictionary of National Biography: Notes for Contributors*, Repr. (Oxford: Oxford University Press, 1998), 23.

³² *New Dictionary of National Biography*, 29. Emphasis in original. For a thoughtful discussion of fillable paper forms in relation to structured knowledge, see Lisa Gitelman, “A Short History of _____,” in *Paper Knowledge: Toward a Media History of Documents* (Durham: Duke University Press, 2014), 21–52.

Revised version: CMR, 3 June 2004 (see also new sheets)

BIRTH AND/OR BAPTISM (if birth date not known)

Birth
Date: day / month / year _____ Place of birth (full address if known; otherwise street / town or city / county / country) _____

Baptism
Date: day / month / year _____ Place of baptism (full address if known; otherwise institution / town or city / county / country) _____

Source of data _____
Comments _____

or FLORUIT (use only if neither birth/baptism nor death dates can be given)

fl. _____
Single year or start year- end year _____

Source of data _____
Comments _____

FATHER In Oxford DNB?
Office use only

Main name _____ Surname _____ Forenames Subsidary name: epithet, toponym or patronymic _____

Variants of Main names _____ Surname _____ Forenames Subsidary name: epithet, toponym or patronymic _____

Alternative names (former name, pseudonym, nickname, etc.: please indicate the type of each name) _____

Titles _____

Birth date _____ Day / month / year _____ Death date _____ day / month / year _____ OR Floruit _____

Principal occupation(s) _____

MOTHER In Oxford DNB?
Office use only

Main name _____ Surname _____ Forenames Subsidary name: epithet, toponym or patronymic _____

Variants of Main names _____ Surname _____ Forenames Subsidary name: epithet, toponym or patronymic _____

Alternative names (former name, later name, pseudonym, nickname, etc.: please indicate the type of each name) _____

Titles _____

Birth date _____ Day / month / year _____ Death date _____ day / month / year _____ OR Floruit _____

Principal occupation(s) _____

Figure 5. Extract (Page 1) from ODNB Profile Sheet requested from contributors.

As Mann has described the process: “The contributor wrote something on the profile sheet; our research editor would then edit it so that it provided the information we needed; a copy-editor would then edit it so that it accorded with our conventions, and supplement it with other information gutted from the article text. And finally a keyboarder would enter it into the database using the template that permitted only a small repertory of values.”³³ “Keyboarder” may imply the modern information economy’s most unglamorous drudge-work, but it is important to honor and emphasize the enormous manual labor by Alliance Photosetting in conjunction with part-time workers in Oxford that went into digitizing the profile sheets and further enriching the ODNB biographies with 7 million tags.³⁴ Alliance Photosetting employees responsible for the tagging received a 285-page document of markup instructions, with detailed plans for capturing

³³ Mann, “Searching the Oxford Dictionary of National Biography,” 18.

³⁴ Harrison, “Introduction,” xv.

dates, places, variant names, legal cases, religious denominations, and much else. The results exhibit the inevitable human errors, but miraculously few.

What scholars can also easily miss about the online *ODNB* if they ignore its distinctively digital medium is that the data side of the project seems to have harbored a quiet internationalist politics. Victorian publisher George Smith's original idea for the *DNB* had been for a *universal* rather than *national* biographical dictionary, but Leslie Stephen convinced Smith that the global scope of such a project would be impractical. Both the "national" in the successor *ODNB*'s name and its nearly 4 million pounds of funding from the British Academy similarly ensured that the *ODNB* would never veer overly far from the United Kingdom (broadly defined). However, the emphasis on data always reflected broader historiographical horizons—horizons perhaps commonplace a decade ago but newly salient in post-Brexit Britain. In a 1996 essay, Colin Matthew prophesied, "Who can doubt that in the course of the next century, as nationality in Europe gives way to European Union, so national reference works, at least in Europe, will do so also."³⁵ "Just as the computer is collapsing national library catalogues in a single world-wide series, so I am sure that in the course of the next fifty years we will see the gradual aggregation of our various dictionaries of national biography. We will be much blamed by our users if we do not!"³⁶ Matthew included very similar remarks in his 1997 Sir Leslie Stephen lecture in Cambridge: "It will be remarkable if in the course of the next century—and perhaps quite early in it—the many dictionaries of national biography do not become electronically linked, either in a single great publication, or more likely in an associated series of computer-held texts. Posterity will think us negligent if we do not make what provision we can for this development."³⁷ Matthew had reportedly opposed British entry into the European Economic Community in 1973, but his data internationalism was a different matter.³⁸ The shortest way to a pan-European historiography ran through Pondicherry.

³⁵H. C. G. Matthew, "Dictionaries of National Biography," in *National Biographies & National Identity: A Critical Approach to Theory and Editorial Practice*, ed. Iain McCalman, Jodi Parvey, and Misty Cook (Canberra: Humanities Research Centre, Australian National University, 1996), 17.

³⁶Matthew, "Dictionaries of National Biography," 17.

³⁷Matthew, *Leslie Stephen and the New Dictionary of National Biography*, 35.

³⁸Lawrence Goldman, "A Monument to the Victorian Age? Continuity and Discontinuity in the Dictionaries of National Biography 1882-2004," *Journal of Victorian Culture* 11, no. 1 (January 1, 2006): 115.



Figure 6.

As I turn now to examine the people, places, and professions of the ODNB at scale, it should be evident that the SGML markup that makes it possible is a scholarly artifact from multiple times and places (see Figure 6). It includes entries added to the ODNB through early 2015, and it is most certainly *not* the work of disembodied algorithms. Rather, it is the product of meticulous, judicious, flesh and blood humans whose unglamorous but absolutely central prior work in creating the markup schema, tagging the text, and editing the results fundamentally underpins the research I present here. As if to confirm Jo Guldi and David Armitage's remark in *The Historians' Manifesto* (2014) that historians turning their attention to big data can "simultaneously pioneer new frontiers of data manipulation and make historical questions relevant to modern concerns," it is possible to create valuable, structured datasets related to place names, people, contributors, religious denominations, educational institutions, bibliographical references, legal cases, and more by parsing SGML tags with the Python library BeautifulSoup.³⁹

³⁹Jo Guldi and David Armitage, *The History Manifesto* (Cambridge: Cambridge University Press, 2014). See further Franco Moretti, *Distant Reading* (London: Verso, 2013); Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscopic* (Imperial College Press, 2015).

<Missing>

In order to illustrate how investigating the *ODNB* (1.) *in its entirety* and (2.) *as an historically contingent digital artifact* offers wider purchase on the historical knowledge it makes available and the historical knowledge-making it constrains, I want to recall Mussell's admonition to exploit the differences between print and digital media for scholarly ends. In the SGML file of aggregated profile sheets, there is a tag whose contents are unavailable to users of the *ODNB* online. Nearly a third of profiles—roughly 16,000 of them—include text tagged as <missing>. <Missing> tags demarcate contributors' responses when asked to list what standard factual components they were unable to provide. Examples are mostly straightforward declarations of known unknowns (parents' occupations, precise dates of birth or death, etc.) but they can sometimes be quite colorful indications of historiography in the making. Seen through the lens of the <missing> tag, the *ODNB* exhibits more of its collective mania, exhaustion, frustration, and fun than its more buttoned-up official presentation. Missing, one contributor wrote: "Father's dates. Wife's birth date. Please DO NOT ask me for further information as it has taken me 30 years of effort to accumulate these facts, and I can do no more." Another contributor confessed—rather boldly for someone contributing to a dictionary of biography—"I have been unable to find any biographical information." Versions of "I have not (consulted|checked|looked|seen)" appear 159 times. Of one eighteenth-century courtesan, the contributor notes that a "satirical verse in which she's mentioned is *lively in the extreme*, but adds nothing to what is given in the *Memoirs* concerning her life and times." "Perhaps more could be discovered" about the subject, another contributor wrote tartly, but "For such an extremely minor figure, I believe we know enough."

If such unguarded remarks in the *ODNB*'s digital materials give access to aspects of the *ODNB* that were understandably withheld from public view, they also give renewed salience to Colin Matthew's comments about the apparent solidity of the old *DNB*—remarks that remain applicable in certain respects to the new one: "We should be cautious about erecting major cultural interpretations—as many have done—on the basis of Old *DNB* coverage. Many less elevated factors" contributed to the composition of the dictionary.⁴⁰ Applied to the *ODNB* on the large scale, the caution that the *ODNB* is ultimately historiography rather than history *tout court* is certainly apt.

While the *ODNB* purports to be a *national* biographical dictionary, investigating the *ODNB* (1.) *in its entirety* and (2.) *as an historically contingent digital*

⁴⁰Matthew, *Leslie Stephen and the New Dictionary of National Biography*, 14.

artifact allows us to put pressure on its own self-representation. Here, it is important to recall the *ODNB*'s quiet data internationalism, which comes to the fore most strongly in what, in the SGML markup, is tagged <geocult>. In one respect, this data testifies to how poorly any single national frame serves most *ODNB* lives as lived. Captured between these tags are contributors' responses to the invitation to list any "geographical/ethnic" associations of their subjects. While the tag's name itself (geocult) suggests some internal tension within the *ODNB* between so-called "cultural" and "ethnic" associations, profiles of figures like King Edward VII, traveller William Lithgow (1582-1645?), and explorer Sir Henry Morton Stanley (1841-1904) are annotated with over 25 unique countries apiece. In total, 44,364 lives are attached in one way or another with 810 historical and contemporary countries. Linking the place names with subjects' life dates gives unique access to one of the most intriguing features of the *ODNB*'s historiographical imaginary.

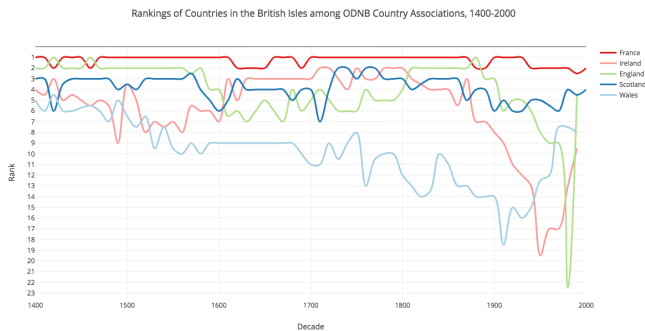


Figure 7. *ODNB* Geographic and Cultural Associations, 1450-2000 Height relative to other lines represents rank of country association for a given decade.

Latent in thousands of biographies is something both more and less illuminating than the normal stories of rise and fall of nations, less illuminating perhaps because dependent on the peculiar collection of lives registered in the *ODNB*, less too because it at best provides access to Britain's subjective sense of the rise and fall of nations—the entrances and exits of the world stage as viewed all-too-partially from Britain, not Brazil or Bengal (Figure 7).

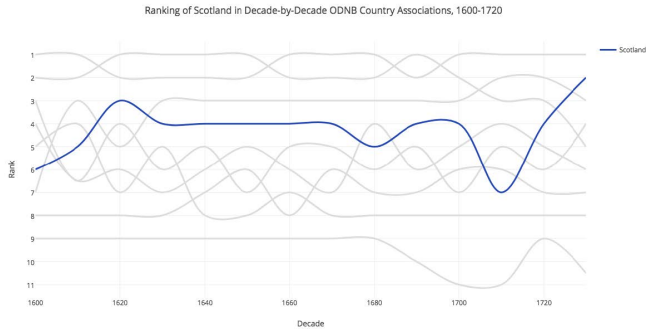


Figure 8.

But there's also something enthralling about data that yields such stories almost organically, accidentally as it were, for no one set out to show Scotland rise precipitously during the reign of King James VI and I and following the Act of Union (Figure 8), nor is it plausible that illustrating post-war American hegemony was the concerted aim of *ODNB* subjects and co-conspiring biographers (Figure 9).

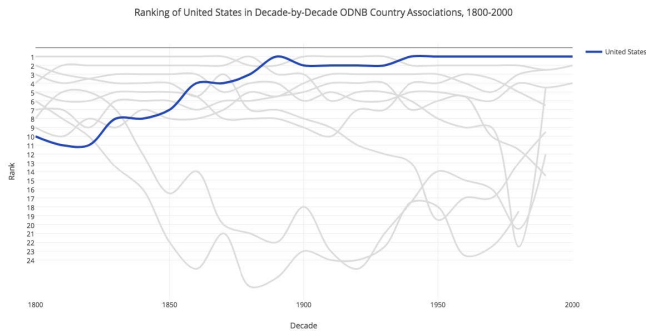


Figure 9.

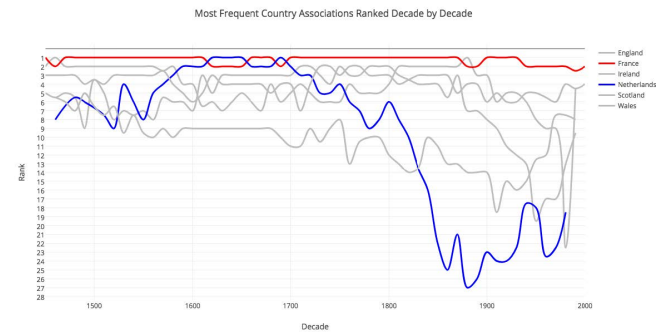


Figure 10.

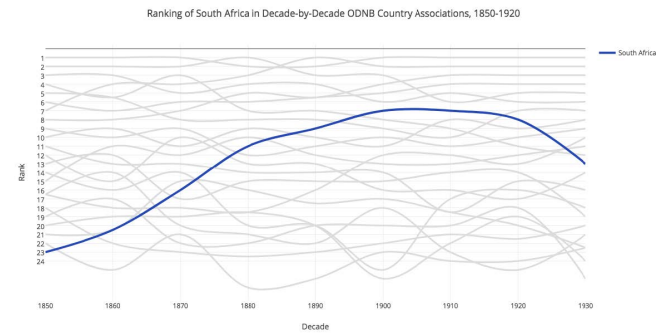


Figure 11.

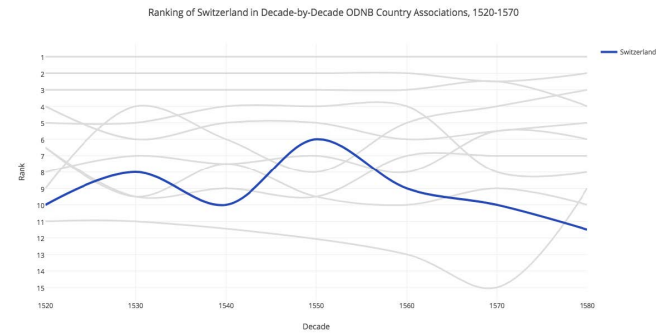


Figure 12.

The Dutch Golden Age (Figure 10) and the Boer Wars (Figure 11) emerge from the *ODNB*'s lives-turned-data as if drawn up by an invisible hand. What's captured in the aggregate are not stories about others but stories about selves—selves made by and in a world of nations. France dominates *ODNB* lives for nearly 400 years between 1400-1800 not only because of such frequent Anglo-French trade, travel, and warfare but because *ODNB* subjects so often claim Norman descent. Amidst the mid-sixteenth century Reformation, Switzerland rapidly becomes part of British selfhood (Figure 11); India becomes part of British selfhood more slowly but also far more durably over the 18th-century (Figure 12).

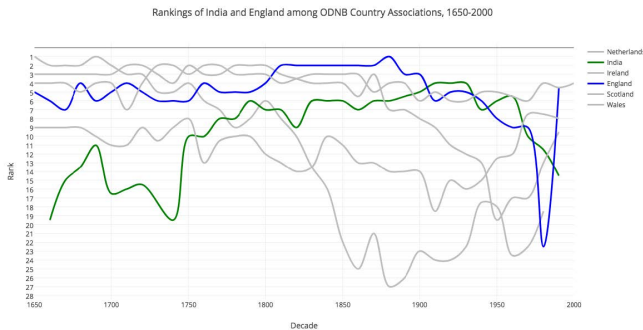


Figure 13.

If these seem like historical truths (Reformation! Empire!) that have somehow floated free from their historiographical infrastructure, historiography's second voice—the voice of the *ODNB*'s own making—speaks here as well. The selves that emerge from *ODNB* data considered in aggregate remain inescapably mediated by the idiosyncratic documents and workflows of the mid-1990s. An investigator looking at Figures 7 and 13 expecting total access to global history's hidden currents would do well to observe that England, surprisingly, is only occasionally the *ODNB*'s most prominent country. The referent of “national” in “national biography” had always been left deliberately underdetermined, but how can it be that between 1450 and 2000, France, the Netherlands, and the United States of America each supersede England, Scotland, Ireland, and Wales in importance—this in Britain's own dictionary of national biography?

Paradoxically, the answer has to do with the *ODNB*'s predominant Englishness. England is the water in which the *ODNB* swims. It was on page five of the five-page Profile Sheet that *ODNB* contributors were asked to complete associations by descent (“Family/cultural origins”) and by association (“Political, professional, Major landholding, travel interests etc.”) (Figure 14). When it

came time on the Profile Sheet's final question to list countries of descent and association, it seems that England was simply assumed. France could emerge from this data as the *ODNB*'s prevailing country because unlike England, it was an *extra* place requiring explicit mention. At the same time, Ireland, Scotland (and to a lesser extent Wales), could be seen as co-equals with England in the *ODNB*'s geo-historical imaginary even though biographical subjects' countries of birth and death leave little doubt that England is the quiet center of the *ODNB* (Figure 15). Ironically, it is the very weight of Englishness as mediated by the Profile Sheets contributors completed that artificially makes France seem more important than England—and Ireland, Scotland, and Wales equally constitutive of the *ODNB*.

For office use only

GEOGRAPHICAL/ETHNIC ASSOCIATIONS

By descent _____
Family / cultural origins

By association _____
Political, professional,
Major landholding,
Travel interests etc.

Additional fields of interest _____

Figure 14.

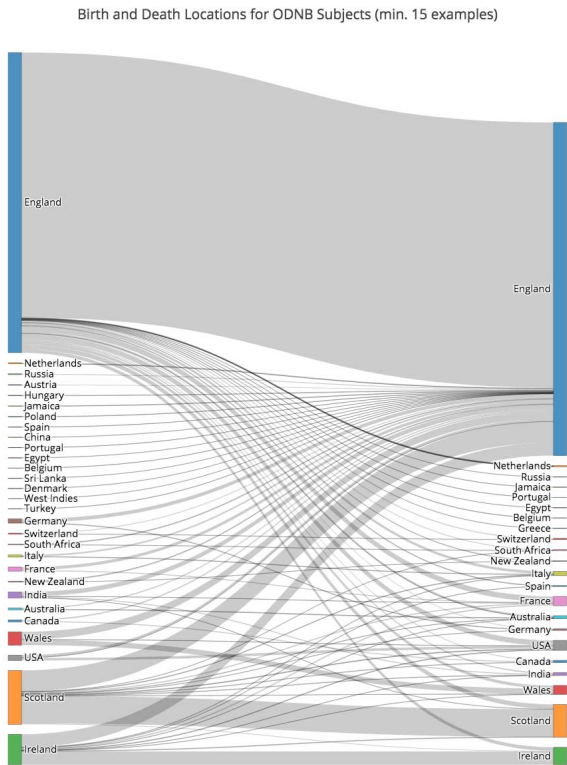


Figure 15.

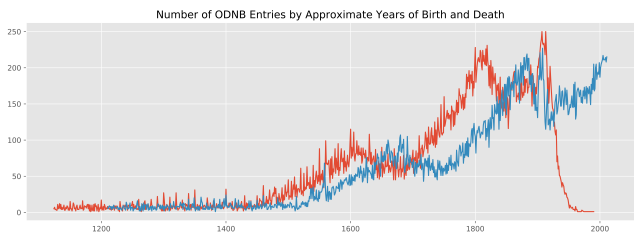


Figure 16.

Considering the ODNB (1.) *in its entirety* and (2.) *as an historically contingent digital artifact* yields a fuller understanding of its temporal and gender contours

as well. While the *ODNB* boasts it is “the national record of men and women who have shaped British history and culture, worldwide from the Romans to the 21st century,” in numerical terms, a key date to understand its composition is 1785, for half of the *ODNB*’s words are dedicated to people born before 1785 and half dedicated to those born after. A subject picked at random, in other words, is just about equally likely to have been born before the French Revolution as after. Medievalists and early modernists like myself might fault the *ODNB* for setting its balancing point so far into the timeline of British history, but it is worth noting that the equivalent fulcrum point for the British population overall is in fact considerably later, likely around 1935.⁴¹ One reason for the *ODNB*’s particular distribution of lives is of course the *DNB*, whose roughly 38,000 lives obviously are almost all pre-1900. For the 14,169 biographies revised but not rewritten for the 2004 *ODNB*—all but 535 of them men—the number of subjects alive peaks at 1833.

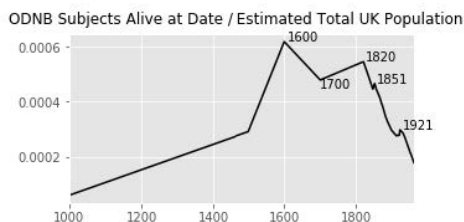


Figure 17.

In another respect, however, the *ODNB* may be particularly valuable to early modernists, considering *ODNB* representation as a ratio of total UK population. Pre-19th century population data is patchy, but Angus Maddison’s estimates for populations at 1500, 1600, and 1700 and year-by-year numbers beginning in 1820 suggest that, all else being equal, a man born around 1600 would have the best chance of being immortalized in the *ODNB*.⁴² Of all individuals alive in 1500 or 1900, about 3 out of every 10,000 has an *ODNB* entry. For those born around 1600, the number is twice that (Figure 17).

⁴¹I use “balancing point” and “fulcrum” interchangeably to refer to the year before which and after which the cumulative populations of those who have lived are equivalent.

⁴²Angus Maddison, “Historical Statistics of the World Economy: 1-2008 AD,” accessed March 21, 2017.

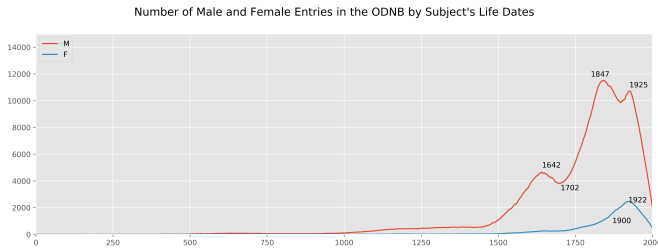


Figure 18. *ODNB* subjects alive at date (male and female).

Women are another story altogether (Figure 18). For commentators including Virginia Woolf, Gillian Fenwick, Allison Booth, Jane Garnett, and others, the representation of women in the national biographical dictionaries has been an important and longstanding concern.⁴³ Due in no small part to their efforts, the *ODNB*'s 10% of female lives improves measurably on the 3.5% of women in the initial *DNB*, and the *ODNB* number has been growing steadily with recent supplements. In raw numbers and proportion of total UK population alike, women are best represented in the early 20th century, with the years around the suffragette movement—late 1910s and early 1920s—offering the highest numbers. Alas, even at its highest the proportion of women in the *ODNB* relative to the total UK population is little better .005% or 1 out of every 20,000 (Figures 19 and 20).

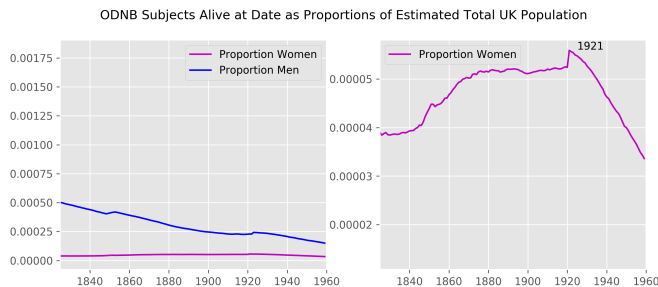


Figure 19 and 20.

Inescapably, *ODNB* subjects' dates of death again tell the twin stories of Britain and the *ODNB*'s own making. The patterns are mostly smooth rather than jagged, but we do see some interesting spikes in the late-1560s and the 1640s, which

⁴³ Gillian Fenwick, *Women and the Dictionary of National Biography: A Guide to DNB Volumes 1885-1985 and Missing Persons* (Scolar Press, 1994); Alison Booth, "Focus on the Oxford DNB: Fighting for Lives in the *ODNB*, or Taking Prosopography Personally," *Journal of Victorian Culture* 10, no. 2 (2005).

correspond with known tumultuous periods with many prominent, public deaths (Figure 21).

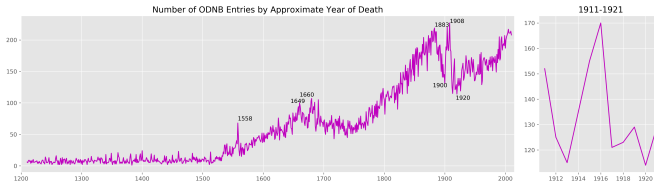


Figure 21 and 22.

In the 1558 spike, for example, the *ODNB* bears the marks of the notoriously fraught transition from Catholic Mary to Protestant Elizabeth. The local peaks in 1883 and 1908, however, once again remind us to attend to the data infrastructure. Rather than marking some hitherto unknown plague afflicting the Victorian aristocracy, 1883 marks the point at which contemporaneous deaths ceased to be meaningful to Stephen, his deputy Sidney Lee, and their collaborators. By 1883, Leslie Stephen and the original *DNB* contributors had turned their attention to those already dead: deaths fall off after that year simply because *DNB* editors had little room for or interest in accommodating those who died while the *DNB* was in process. For that sharp rise in deaths in the first decade of the twentieth century, it's a similarly material story involving historiographical infrastructure. Neither the result of a war nor a pandemic, the spike in deaths in that decade suggests something more mundane and yet fundamental to the project of understanding the historical and institutional underpinnings of the *ODNB* data.⁴⁴

In 1912, eleven years after the initial publication of the *DNB*, Sidney Lee published his three-volume second Supplement, dedicated to some 1,660 people who died between 1901-1911. The number of biographies in this particular supplement might not raise eyebrows now, but it certainly did in the years following 1917, when Oxford University Press formally took over the project and deemed it too much. It is clear from the "Prefatory Note" to the *ODNB*'s first supplement (1927) that the Press had little interest in continuing Lee's pace, which amounted to a "bold and attractive experiment" but one that would, if continued throughout the 20th-century, "add about 15,000 lives (and nearly 20,000 pages of print) to the main work, which (with the three supplementary volumes published in 1901) contains a little more than 30,000 substantive articles."⁴⁵ OUP's

⁴⁴I owe this point to conversation and subsequent e-mail exchange with Mark Curthoys, who very helpfully pointed me toward the OUP 1927 Supplement's "Prefatory Note."

⁴⁵H.W.C. Davis and J.R.H. Weaver, eds., "Prefatory Note," in *Dictionary of National Biography: Founded in 1882 by George Smith, 1912-1921* (London: Oxford University Press, 1927), v.

DNB would instead be one planned along “less ample lines” than the Smith-Lee biographical bonanza, and it is this editorial and financial decision that explains the 1908 bump.⁴⁶ Understood as an aftereffect of early 20th-century editorial choices, then, the true anomaly in terms of *ODNB* representation is the period between 1901-1911, years whose number of deaths in many cases even surpass those of the First World War. Yet, once we learn to look past the publishing effect, we can indeed see the higher number of deaths that we might expect for the years 1914-1916 (see Figure 22).

My argument thus far has been that the *ODNB*, considered in its entirety as an historically contingent artifact, tells two simultaneous stories—the history of Britain in a world of nations and the story of its own making. Perhaps nowhere is the latter more evident than in the bibliographical underpinnings of the *ODNB*. Consider the hidden connection between the Victorian naval historian Sir John Laughton and contemporary scholar Rory Muir. Biographical research never occurs in a historiographical vacuum and almost always relies on existing archives, scholarship, obituaries, diaries, indexes, finding aids, and earlier biographical dictionaries. J.H. Hexter once termed these “the infrastructure, the bone and gristle” of the historian’s trade.⁴⁷ For every *ONDB* biography, contributors included a “source sheet” listing archives and references, which was then digitized by the *ODNB*. To look at the list of the most frequently-cited book authors resulting from this process is to see a confection of historical personages and modern historians appropriate to this project of intersecting temporalities.

Joseph Foster, the great Victorian editor of *Alumni Oxonienses*, tops the list that also includes Horace Walpole, 17th century biographer Anthony Wood, Reformation martyrologist John Foxe, cataloguer of ejected 17th-century ministers Edmund Calumny, and modern historian of the Elizabethan-era Patrick Collinson (Figure 23).

⁴⁶ Davis and Weaver, “Prefatory Note,” v.

⁴⁷ J. H. Hexter, “The Rhetoric of History,” *History and Theory* 6, no. 1 (1967): 4.

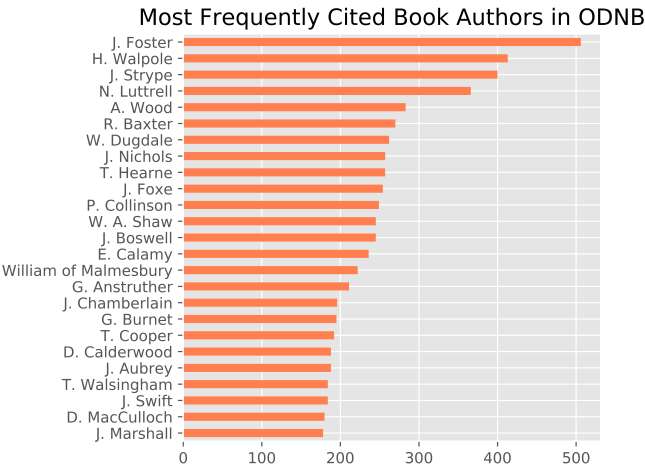


Figure 23.

Collinson—a prolific authority on Elizabethan puritanism who as associate editor for the *ODNB* oversaw the lives of the *ODNB*’s 97 post-1560 godly divines, non-conformists, and radicals—in fact can lay claim to the being the *ODNB*’s single most-cited modern scholar, something he achieved in part by citing himself over thirty times. Reformation Church historian Dairmand MacCulloch’s impressive 175 citations benefit from no such self-help (MacCulloch did not write any *ODNB* biographies).

Who else does the *ODNB* cite? Well, itself, at least insofar as the *DNB* is the most-cited work overall (Figure 24).

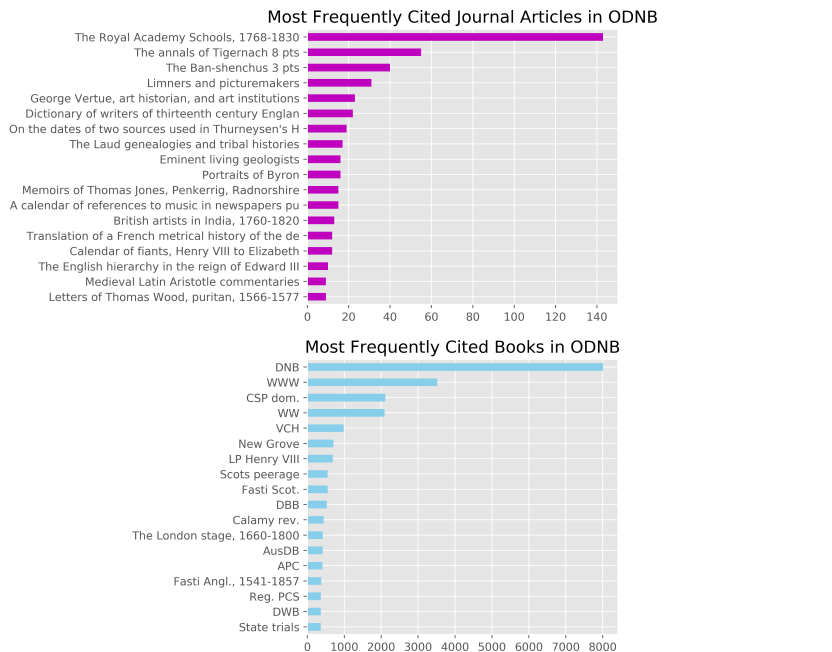


Figure 24.

Sidney C. Hutchison’s 1960 article “The Royal Academy of Schools, 1768-1830,” from the *Volume of the Walpole Society* is cited twice as often as the next most-frequently cited article, but the post-war monograph most-cited overall is Rory Muir’s 1996 book *Britain and the Defeat of Napoleon* (Figures 24 and 25).

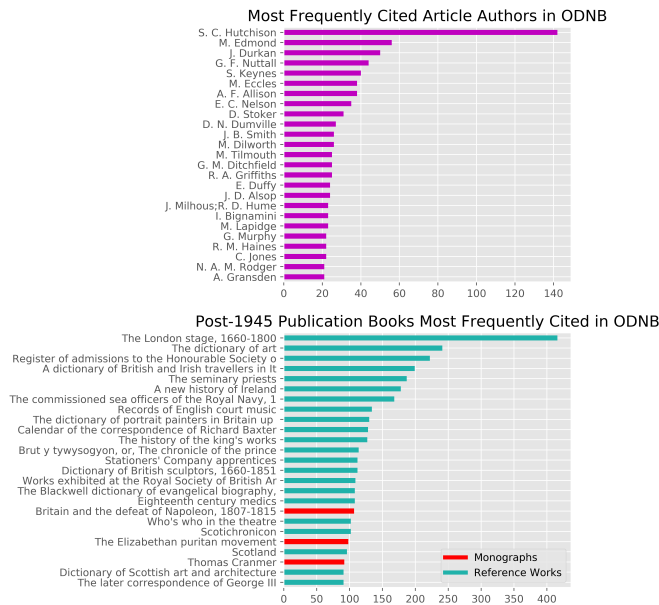


Figure 25.

The raises the question, is *Britain and the Defeat of Napoleon* cited in over 100 biographies because it's the most import monograph in and on British history? Maybe, maybe not, but how do we account for the astonishing prominence of a book published as recently as 1996? One place to begin is by considering the distribution of professions in the ODNB. Profession is a slightly fuzzy category in the ODNB, but the ODNB data offers two broad ways to access its patterns of representation, first through the ODNB's own rubric of "areas of renown," which assigned each biography to one of 25 broad categories such as "Art," "Scholarship and Research," or "Exploration" (Figures 26 and 27). The second is in the data field known as "historical significance," which is the contributors' own account of what makes their subject important for the life of the nation. Such descriptions usually, but not always, include something we might call a profession—though they're quite loose and somewhat difficult to work with, not least because there's no controlled vocabulary. There's little surefire way to ensure kindred occupations like "writer," "journalist," and "pamphleteer" end up in the same bucket. Even so, the recurrence of many professions is illuminating: together, those known simply as army officers, naval officers, politicians, Church of England clergymen, judges and poets account for nearly a tenth of the ODNB. For present purposes, it isn't immediately obvious to a student of British history that

“naval officer” would be the third most frequent profession in the *ODNB*, but here we once again encounter our intersecting temporalities.

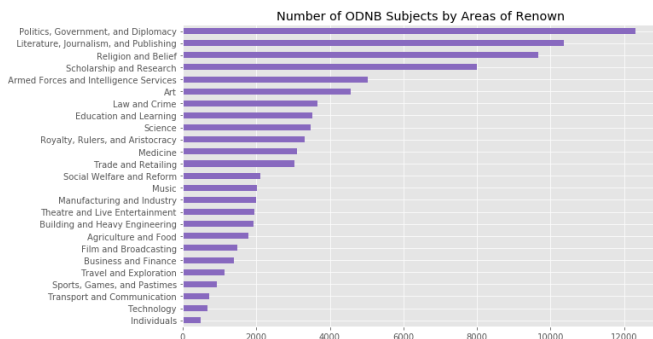


Figure 26.

One of the most prolific Victorian contributors to the *DNB* was a naval historian named Sir John Laughton (1830-1915), who was responsible for an astonishing 1000 biographies of naval figures, roughly 1 out of every 38 *DNB* entries. Because of the editorial decision to include all *DNB* figures in the *ODNB*, Laughton's legacy required massive research into British naval history, especially in the period of the Napoleonic Wars. The temporal distribution of entries in the Armed Forces and Intelligence Services grouping, which rises relative to other areas of renown in the late 18th century, confirms this point. As the *ODNB* got underway in the mid-1990s, it ultimately fell to more than 65 *ODNB* contributors to revise Laughton's entries. Those responsible included Andrew Lambert, now the Laughton Professor of Naval History at King's College London who revised the most at 221, and Roger Moriss, who revised 74 Laughton biographies, second-most. It is here that Rory Muir's *Britain and the Defeat of Napoleon* (1996) comes in. For these modern revisers, *Britain and the Defeat of Napoleon* (1996) was a timely resource indeed. That Rory Muir's 1996 book on the Napoleonic Wars became the most-cited academic monograph in the whole of the *ODNB* is due in no small way to the preternatural historical labors of Sir John Laughton more than a century before. Roughly a quarter of all of the Muir citations occur in Roger Moriss' Laughton revisions alone.

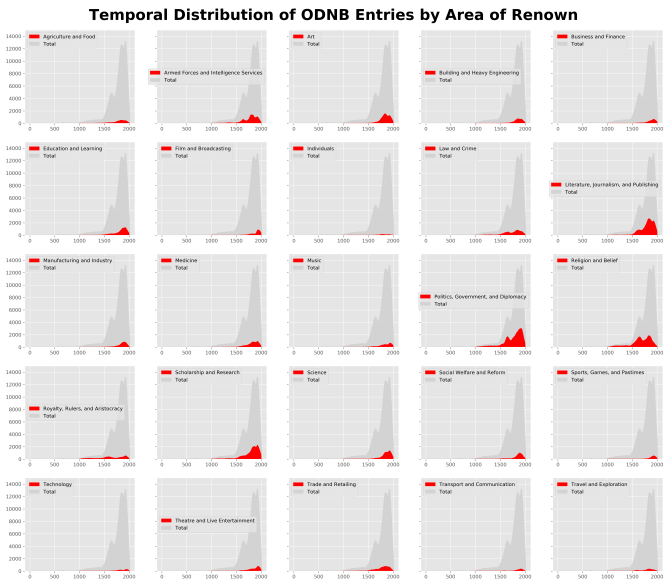


Figure 27.

Investigating the *ODNB* through the lens of professions and areas of renown doesn't only help answer questions about the most the most frequently cited archives, books, and articles. By grouping the biographies of people who share a historical significance, it also becomes possible to observe the emergence and sometimes decline of certain activities over time. The results are particularly revealing in the case of female subjects. Of groups with 15 or more examples, “author” is both the earliest and most enduring historical significance for women (Figure 28). The last *ODNB* woman known principally as a “noblewoman,” by contrast, died in 1821.

Lest we make too much of such claims, however, a finer resolution can be salutary, for no one would claim that 1821 was the death of British nobility.

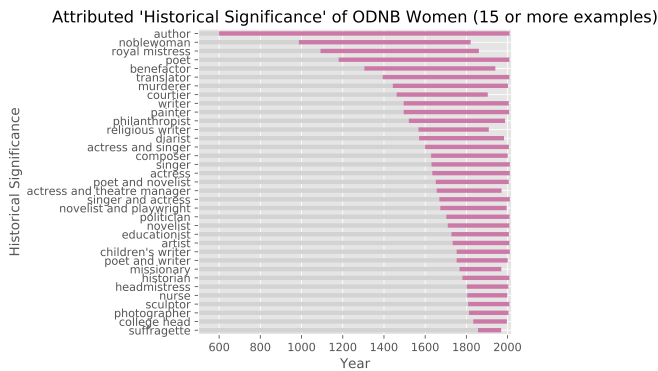


Figure 28.

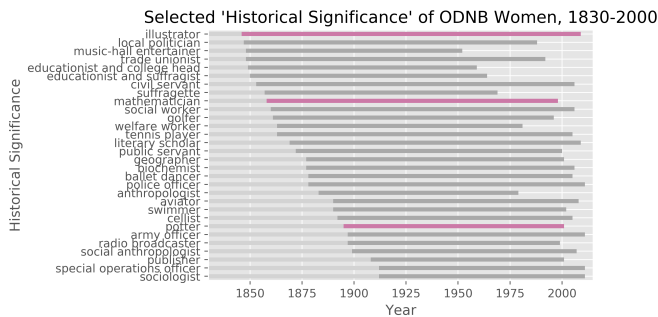


Figure 29.

Neither did pottery, for example, only become available as a female activity at the turn of the 20th century (Figure 29). Instead, it was only in the 20th century that such an activity achieved sufficient status to merit inclusion in a dictionary of national biography. Far from indicating something about Britain's “last” noblewoman or its “first” female potters, illustrators, or mathematicians, this data reveals the contingency of first-ness and last-ness, the cultural and historiographical factors that make certain historical activities legible or illegible at various points in time.

* *

Conclusion

Whether in the case of the *ODNB*'s most frequently mentioned years, the rise and fall of nations as seen through its lens, its most frequent citations, annual numbers of deaths, and representations of women and their historical significance, the *ODNB* has simultaneously testified to Britain's past as such and about the particularities of the Dictionary's own making.

I have emphasized these dual aspects of historiography in part because many discussions of historical data at scale emphasize one but not the other. On the one hand, data at scale is sometimes treated as a pure pipeline to the past. This was one of the pitfalls of so-called "cliometrics" in the 1970s, and it has found renewed prominence in some strands of digital history and cultural analytics. So-called "macroanalyses" that "place[...] the emphasis on the systematic examination of data, on the quantifiable methodology" while "demphasiz[ing] the more interpretive act of reading," are intended as "a more objectively determined exploration of facts."⁴⁸ "With enough data, the numbers speak for themselves," Chris Anderson has written.⁴⁹ The result—ostensibly desirable—is "history as it is told by...robots."⁵⁰

On the other hand, the tendency toward contingency and specificity, dubbed by some the "idiographic" impulse, frequently roars back against data enthusiasts' "nomothetic" urge for larger scale generalities and quasi-robotic historical laws.⁵¹ Some scholars view historical data at scale as a poisoned fruit, irreparably tainted by the means of collection or the positivist ideology of data itself. Ethan Kleinberg, Joan Wallach Scott, and Gary Wilder, the authors of the recent "Theses on Theory and History," for example, critique mainstream academic historians' uncritical "commitment to empirical data that serves as a false floor to hold up the assertion that past events are objectively available for discovery, description, and interpretation."⁵² Yet it is quite possible, even desirable, to hear both of these voices from within the digital *ODNB* considered in its entirety. We can agree both with Thomas Piketty, whose *Capital in the Twenty-First Century* credits its major contributions in economic history to "advances in computer technology

⁴⁸Matthew Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013), 25, 29.

⁴⁹Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2008.

⁵⁰Erez Aiden and Jean-Baptiste Michel, *Uncharted: Big Data as a Lens on Human Culture*, 1st edition (New York: Riverhead Books, 2013), 23.

⁵¹Scott Weingart, "Lessons From Digital History's Antecedents," *The Scottbot Irregular* (blog), October 26, 2016.

⁵²Ethan Kleinberg, Joan Wallach Scott, and Gary Wilder, "Theses on Theory and History," accessed September 3, 2018.

[that] have made it much easier to collect and process large amounts of historical data,” and also with Mimi Onuoha, who insists that “data sets are the results of their means of collection” and that they frequently “outlive the rationale for their collection.”⁵³ The *ODNB* confirms both things at once.

And the point where these perspectives meet offers one of the strongest rationales for studying historiography at scale. While big historiographical data might seem anathema to a traditional humanist ideographic epistemology invested in specificity and contingency, the rewards of such an approach can be appreciated according to those very same standards of knowledge. Considering digital resources in their entirety and as historically contingent artifacts demystifies the means of collection and more clearly delimits the knowledge-making that a particular data infrastructure allows and constrains.

Ultimately, the online *ODNB* is now an indispensable scholarly resource for just about every field touching on British history and culture, but apart from a brief flurry of scholarship surrounding its 2004 publication, it has rarely itself been the object of scholarly investigation. Whether researchers have considered it or not, those who use the *ODNB* almost daily have been like the fabled blind men gathered around an elephant who, each feeling a different part, likened it, variously, to a wall, a spear, a snake, and a tree. The digital *ODNB* is a complex artifact of intersecting temporalities, characterized like any digital archive by selections, assumptions, and transformations. But studying such archives in their entirety and as historically contingent digital artifacts is necessary for measuring and apprehending the whole beast.



Unless otherwise specified, all work in this journal is licensed under a Creative Commons Attribution 4.0 International License.

⁵³Thomas Piketty, *Capital in the Twenty-First Century* (Harvard University Press, 2017), 26; Mimi Onuoha, “The Point of Collection,” *Data & Society: Points*, February 10, 2016.