



Open Library of Humanities



Part of the Ubiquity
Partner Network

Digital Studies /
Le champ numérique

Research

How to Cite: Laramée, François Dominic. 2019. "How To Extract Good Knowledge from Bad Data: An Experiment with Eighteenth Century French Texts." *Digital Studies/Le champ numérique* 9(1): 2, pp. 1–25. DOI: <https://doi.org/10.16995/dscn.299>

Published: 30 January 2019

Peer Review:

This is a peer-reviewed article in *Digital Studies/Le champ numérique*, a journal published by the Open Library of Humanities.

Copyright:

© 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Open Access:

Digital Studies/Le champ numérique is a peer-reviewed open access journal.

Digital Preservation:

The Open Library of Humanities and all its journals are digitally preserved in the CLOCKSS scholarly archive service.

RESEARCH

How to Extract Good Knowledge from Bad Data: An Experiment with Eighteenth Century French Texts

François Dominic Laramée

Université de Montréal, CA

fdl@francoisdominiclaramee.com

From a digital historian's point of view, Ancien Régime French texts suffer from obsolete grammar, unreliable spelling, and poor optical character recognition, which makes these texts ill-suited to digital analysis. This paper summarizes methodological experiments that have allowed the author to extract useful quantitative data from such unlikely source material. A discussion of the general characteristics of hand-keyed and OCR'ed historical corpora shows that they differ in scale of difficulty rather than in nature. Behavioural traits that make text mining certain eighteenth century corpora particularly challenging, such as error clustering, a relatively high cost of acquisition relative to salience, outlier hiding, and unpredictable patterns of error repetition, are then explained. The paper then outlines a method that circumvents these challenges. This method relies on heuristic formulation of research questions during an initial phase of open-ended data exploration; selective correction of spelling and OCR errors, through application of Levenshtein's algorithm, that focuses on a small set of keywords derived from the heuristic project design; and careful exploitation of the keywords and the corrected corpus, either as raw data for algorithms, as entry points from which to construct valuable data manually, or as focal points directing the scholar's attention to a small subset of texts to read. Each step of the method is illustrated by examples drawn from the author's research on the hand-keyed *Encyclopédie* and *Bibliothèque Bleue* and on collections of periodicals obtained through optical character recognition.

Keywords: text mining; data mining; textometrics; production of space and place; digital history; error correction

Du point de vue d'un historien numérique, les textes français d'Ancien Régime souffrent d'une grammaire obsolète, d'une orthographe irrégulière et d'une reconnaissance optique des caractères de faible qualité. Cet article résume les expériences méthodologiques qui ont permis à l'auteur d'extraire des mesures quantitatives utiles de ces improbables matières

premières. Une discussion des caractéristiques générales des corpus de textes historiques transcrits à la main et des corpus produits par reconnaissance optique révèle qu'ils diffèrent en degré de difficulté mais non en nature. Les comportements qui rendent certains de ces corpus particulièrement difficiles à traiter numériquement, dont la distribution non aléatoire des erreurs, un coût unitaire d'acquisition relativement élevé, la dissimulation des documents atypiques et l'imprévisibilité des erreurs répétées, sont ensuite expliqués. L'article trace ensuite les grandes lignes d'une méthode qui contourne ces problèmes. Cette méthode repose sur la sélection heuristique de questions de recherche pendant une phase d'exploration ouverte des données; la correction sélective des erreurs à l'aide de l'application de l'algorithme de Levenshtein à un petit nombre de mots-clés choisis pendant la phase d'exploration; et l'exploitation des mots-clés et du corpus corrigé soit en tant que données brutes, soit comme points d'entrée permettant l'extraction manuelle de données probantes, soit comme boussoles permettant d'orienter l'attention du chercheur vers un sous-ensemble de documents pertinents à lire. Des exemples tirés de la recherche de l'auteur, qui porte à la fois sur des corpus océsés de périodiques et sur les corpus reconstitués manuellement de l'*Encyclopédie* et de la *Bibliothèque bleue*, illustrent chacune des étapes.

Mots-clés: fouille de texte; fouille de données; textométrie; production de l'espace; histoire numérique; correction d'erreurs

What is a digital historian supposed to do with data that is barely tractable to digital methods? Most humanists, whether they use computational methods or not, have to contend with incomplete, inconsistent, error-ridden, or otherwise problematic data. When the amount of this messy data required to answer a research question is small enough, it may be possible to clean it up by hand or even to fill in the blanks and filter out the inconsistencies mentally as one reads through sources and computational results. However, this strategy grows less feasible as the volume of data increases, especially for an individual scholar with finite reserves of time. Historical sources compound the problem by introducing issues that are not found in more recent documents. For example, eighteenth century French books and periodicals are peppered with obsolete grammar and irregular spelling, which natural language processing software designed with modern digitized text in mind is ill-equipped to handle. Historical text is also prone to poor optical character recognition (OCR) results, and error correction techniques that perform well when applied to modern

text do not always translate well to sources that do not follow modern text's OCR error *production* patterns. Thus, in many cases, perhaps in most cases, cleaning an entire historical corpus prior to performing a digital analysis is unrealistic.

By analogy with Big Data, this paper calls a large corpus of text that defies cleanup efforts because of its size or its internal characteristics *Bad Data*. This paper contends that, when handled with proper care, this Bad Data can still yield good quantitative results. The theoretical-methodological framework required to do so involves an intimate knowledge of the corpus, a targeted approach to error correction, and a measure of humility about the historical questions that can be answered given the limits of the two. Throughout the paper, this theoretical framework will be presented step by step and illustrated by examples from the author's own research. The framework also illustrates the inescapable need, in making a Bad Data corpus tractable, for a symbiosis between digital methods and human judgement. Finally, the paper contends that its framework applies to many situations in which mining a Bad Data corpus is likely to be useful (albeit at the cost of some customization based on the idiosyncrasies of the case at hand), if only as a parable about the limits of quantification and about the value that one can derive despite these limits.

This article is divided into two major parts. First comes a characterization of Bad Data, how it differs from Big Data, and why historical text qualifies as Bad Data. In regard to the last, first and foremost of these reasons is the way in which defects in historical text, including actual errors and artefacts of language that may pose similar challenges to scholars, tend to be clustered rather than spread more or less uniformly. When these error clusters happen in parts of the source material that is highly salient to the research questions under study, they can skew the results to an unacceptable level. Other reasons include potentially high unit cost of acquisition of a large corpus of text; the ways in which text encoding methods may introduce crucial defects; the ambiguity caused by polysemy; and the partial incompatibility between historical text and current digital language processing tools. The second half of the article is devoted to outlining a method that can extract reliable information from a Bad Data historical text corpus despite these defects. This method is divided into three steps, and it may need to be iterated several times before it reaches a stable solution. The first step is a heuristic process of research question design that relies

on exploration of the corpus to figure out what it may be able to answer. The second step consists of a targeted error correction scheme built around a limited number of keywords that are likely to lead to an answer. In the third step, the scholar assesses how the keywords and the corpus can be mined for the answer, either as raw data themselves, as tools to guide the indirect extraction or construction of further data, or as a way to focus the scholar's close reading on a small number of particularly salient elements of the corpus.

Thus, the approach described in this article relies on constant back and forth between data curation, judgement calls, and digital methods, far more than it does on unadulterated technical wizardry. This approach has, however, served the current author well.

Big Data, Bad Data, and the Perils of Historical Corpora

Historians are accustomed to working with Small Data. The typical historical argument relies upon a limited collection of highly salient documents, painstakingly exhumed from the archive at considerable expense in time and toil. These documents are then critically interpreted, sometimes against the grain, to filter out the biases of their creators or to unearth nuggets of information that the documents' creators never directly intended to transmit to posterity. In other words: each unit of Small Data comes at a high price, but it yields correspondingly high value. Big Data, inasmuch as it can be defined, is the opposite in all aspects. Big Data all but accumulates on its own: once a pipeline has been set up to harvest tweets or Web search queries, the incremental effort required to obtain millions of them is negligible. However, since we collect Big Data to reveal trends and patterns that escape the human eye, it is only meaningful in very large amounts. Finally, the promise of Big Data is that quantity trumps quality (Mayer-Schönberger and Cukier 2013, 16–33). If one has easy access to millions of units of content, says the theory, there is no need for critical assessment of each individual unit because the errors will spread more or less uniformly, and a useful signal will still percolate from underneath the noise.

For digital historians, it is tempting to approach large textual corpora as if they were Big Data. And indeed, it is certainly possible to assemble corpora that are large

enough to qualify. Broadly speaking, such corpora can be divided into two categories. The first category includes the small number of sources that have been hand-keyed into digital form through the efforts of scholars and volunteers, either as plain text files or as sets of TEI-encoded, metadata-enhanced documents. Chief among them, for the historian of Ancien Régime France, is the treasure trove provided by the University of Chicago's invaluable ARTFL project, including the *Encyclopédie* (Morrissey and Roe 2017), which was painstakingly reconstructed from microfilm in the late 1990s, and the *Bibliothèque Bleue* (ARTFL 2016), ARTFL's collection of 284 works of popular literature published between the 16th and 19th centuries. As scholarly editions, these digital archives reproduce the original source materials, with all of their idiosyncrasies, as faithfully as possible. Far more common, of course, is the second category, which includes the sources to which scholars only have access thanks to optical character recognition. Gallica (Bibliothèque nationale de France 2018), the French national library's online archive, provides an enormous collection of such documents, including complete or nearly complete runs of several eighteenth century periodicals such as the news-oriented *Gazette*, the literary *Mercur de France*, and the western world's first scholarly publication, the *Journal des Savants*. In both cases, it is relatively easy for a scholar to mine these resources to assemble data sets containing tens of millions of word tokens; hardly comparable to the billions of tokens in Google's word vector training set, perhaps, but plenty to make a Big Data approach seem appealing.

However, treating such data sets as Big Data would be hazardous because historical text tends to violate the rules that define Big Data. OCR errors and other artefacts of language are definitely *not* distributed at random. Mining textual corpora that have been created by others means abiding by the decisions of others, including the authors and editors of the original source material in the past, which may or may not be appropriate for one's needs. Critical interpretation and close reading are always necessary because words have multiple meanings and their usage changes over time. Perhaps worst of all, extracting numerical features from a large volume of text may suggest the existence of patterns that are mere artifacts of the ways in

which the text has been encoded – something that the transformation into numbers has made invisible. And of course, if the sources we want to examine have not yet been digitized and only exist in print or microfilm, acquiring data in bulk can be extremely time consuming and, especially when the source material has been poorly preserved, devilishly tricky. For these reasons, historical text corpora should not be considered Big Data, but rather a form of Bad Data that combines some of the most troublesome features of Small Data and Big Data, *even when they have been hand-keyed to perfection*. The next two sections will explain why.

Why Historical Text Violates the Random Distribution of Errors

As mentioned earlier, one of the key assumptions of Big Data theory is that defects are spread more or less uniformly, which makes them irrelevant when the amount of data is large enough. All historical text corpora violate this rule because they contain *clusters of defects*, some but not all of them predictable. The very nature of language is the source of most of these clusters; print technology and editorial decisions create others. And while common sense dictates that hand-keyed corpora are preferable in the abstract, they are no more immune to the clustering effect than those assembled through OCR.

The clustering effect emerges as a consequence of the three types of defects identified by Michael Piotrowski (2012) in his discussion of the pitfalls of historical text processing: changes in spelling and word meanings over time, irregular spelling in the case of sources that predate the standardization of orthography, and uncertainty due to errors in transcription or optical recognition. The first two types of defects are not errors *per se* but rather historical phenomena that may or may not be significant to a scholar's work. For linguists, these defects may be salient pieces of data; for historians interested in measuring the number of references to a place whose name is spelled in multiple ways, they are functionally identical to OCR errors unless the scholar knows the list of possible spellings ahead of time and plans accordingly. In any case, as we will now see, none of these defects are randomly distributed. Intimate knowledge of the corpus is necessary to figure out what defects are present, how they can influence the research process, and how to implement the appropriate corrective measures.

In the case of eighteenth century French texts, a particularly cumbersome cluster of defects due to spelling changes over time arises from the relatively recent (historically speaking) replacement of an *o* with an *a* in such ubiquitous French words as *avoit/avait* (had), *étoit/était* (was), and even *françois/français* (French). This seemingly innocuous change can wreak havoc on digital text analysis because most natural language processing tools have been designed with contemporary grammar and spelling (and *only* contemporary grammar and spelling) in mind. The popular TreeTagger part-of-speech parser (Schmid 1997), for example, does not possess an Early Modern French grammar, and its contemporary French grammar regularly misidentifies the word types *avoit* and *étoit* as nouns instead of the archaic past-tense spellings of the two most common verbs in the French language that they are. This mistake repeats itself thousands of times in any large corpus, with potentially dire consequences for the unwary. When using TreeTagger along with the TXM textometric software package (Heiden, Magué and Pincemin 2010), scholars studying vocabulary divided by parts of speech must compensate for these tagging errors by hand.

Spelling variance also tends to cluster in highly salient parts of historical text, such as named entities (people, places, etc.) For instance, the word *Louisiane* is spelled three different ways in the ARTFL *Encyclopédie*, and a keyword search for *Encyclopédie* content that mentions Louisiana and that only takes the “correct” variant into consideration would miss nearly a quarter of the relevant entries, including the main article about Louisiana itself. (Called *Louysiane* with a Y, this article contains all three occurrences of the *Louysiane* word type found in the entire seventeen-volume encyclopedia, and no trace of either the “correct” spelling or of any other.) Only through an iterative process of trial and error can language artefacts such as these be uncovered, and it is all but impossible to guarantee that none will escape the scholar’s attention.

Even transcription errors may cluster, especially in OCR data. In the *Gazette*, for example, article headers contain highly valuable information about the cities from where the news originates and the dates on which they were sent to the editor. However, *Gazette* headers are italicized and therefore misread by OCR at a much

higher rate than the surrounding text. The author has observed that the ubiquitous *Versailles*, for example, is misread in headers in more than a dozen different ways, some of which are completely unrecognizable as words at all. For a scholar interested in news dissemination patterns, this type of error cluster is extremely damaging.

As an aside, transcription defects are by no means limited to OCR. The ARTFL *Encyclopédie* was hand-keyed by professionals, and yet more than 650,000 corrections had to be applied to the database between 1998 and 2013 (Morrissey 2016), a process that was undoubtedly made more difficult by the fact that, to twenty-first century eyes, the difference between a transcription error and a correct transcription of a word that was incorrectly or fancifully spelled in the eighteenth century is far from obvious. OCR data derived from eighteenth century periodicals is itself of much lower and much less predictable quality than what a scholar accustomed to working with twentieth century sources would expect. Eighteenth century printers often packed text tightly (paper wasn't cheap) and had to contend with irregular type and with ink that seeped through one sheet to the next. Many old documents accumulated stains, rips and mildew in musty attics for several decades before they even entered the archive. Some sources only survive on microfilm, as slightly misaligned or warped pictures that cause no trouble to the human readers for whom they were produced but bedevil OCR software. As a result, while the OCR success rates reported by Gallica can reach 95% or more for many issues of the *Gazette*, they fall below 50% for some annual compendia of the *Journal des Savants*, whose most problematic passages are almost indistinguishable from strings of characters generated at random.

In summary, because of the clustering effect, the difference between hand-keyed corpora and those obtained through OCR seems to be one of degree rather than of nature. The fact that the examples given for the first two types of defects have been drawn from the most recent release of the hand-keyed *Encyclopédie*, which may very well be the highest-quality digital source available for Early Modern French studies, is sobering indeed. The lesson: every historical corpus must be treated as potential Bad Data until proven otherwise.

Further Characteristics of Bad Data

Beyond the clustering effect, which applies everywhere, other characteristics of historical text may sometimes violate the rules that define Big Data.

First, acquiring a unit of textual data can be relatively expensive compared to its salience. Recovering the places and dates of origin from the *Gazette's* italicized article headlines had to be done by hand. Each of the 1,184 *Gazette* articles that discuss the colonial Atlantic world between 1740 and 1761 also had to be cut from the raw OCR files and pasted into its own .txt file by hand, one at a time; the process could not be automated in any way because the endpoints of an article are just as likely to be misread by OCR as anything else. In both cases, it was obvious that the data would reveal interesting patterns only when acquired in bulk, but the acquisition process required effort at retail.

Second, the sheer volume of text and lack of a regular structure in large corpora make it more difficult to pinpoint outliers. This is dangerous because some outliers are highly salient while others are mere artefacts of the way in which the sources were encoded and must therefore be discarded. **Figure 1** shows an intriguing pattern that emerges from correspondence analysis (Benzécri c1992; Cibois 2007) of the 14,547 *Encyclopédie* articles that discuss geography: the articles extracted from Volume XIII seem to have very little in common with the others. At first glance, nothing seems to explain this phenomenon. A deep dive into the word frequency statistics calculated by volume, however, reveals an odd discrepancy. The letter P, written as a single-character word type, appears no fewer than 9,613 times in Volume XIII and no more than a few dozen times in any of the others. Further inquiry reveals that nearly all of these unexpected occurrences belong to a single article, about the Italian city of Reggia. It turns out that the ARTFL *Encyclopédie* has encoded this article and Volume XIII's appendix in a single file. The appendix in question contains a table listing the prime factors of every number between 1 and 100,000, a table in which prime numbers are marked with a P. Deleting the offending table from the dataset makes the abnormal correspondence analysis result disappear; a less conspicuous culprit, however, could easily have gone unnoticed, with potentially deleterious consequences.

Third, because words are polysemic, word type counts can never be taken at face value. In the corpus of eighteenth century periodicals, for example, the word type *Halifax* refers to a port city in Nova Scotia and to a British lord (usually spelled *Hallifax*). When trying to figure out how often a reader is reminded of the existence

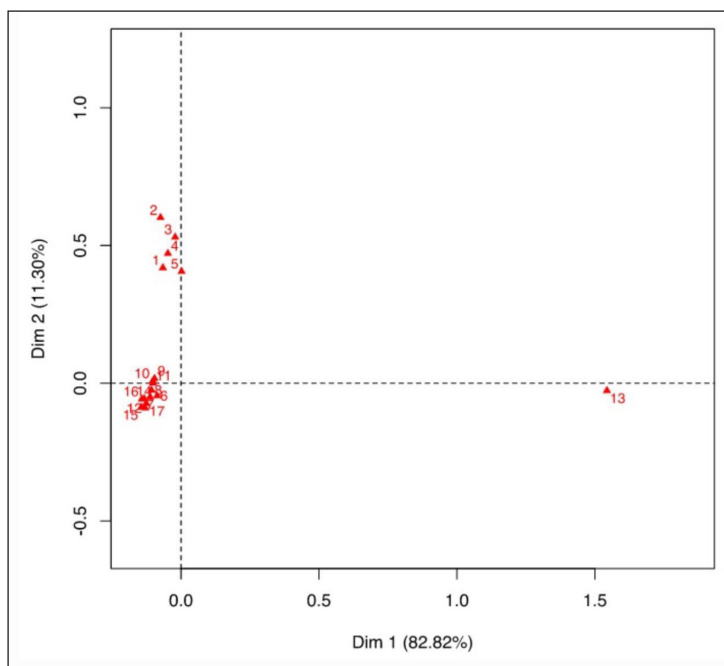


Figure 1: Correspondence analysis of a corpus of 14,547 *Encyclopédie* articles about geography. A data error produces a seemingly impressive result.

of the city, should mentions of the lord be counted? And if the text also talks about the warship *HMS Halifax*, does that count? How do we know whether the ship was named after the colonial port, the lord, or some other town in Britain? A similar problem occurs with the word type *colonie* (colony), which is sometimes used in the periodicals to refer to Atlantic world colonies but far more often to talk about ancient Greco-Roman cities – or, in a few cases, about Cardinal Coloni of the Roman Catholic Church. It is surprisingly easy to second guess one’s judgement calls in matters like these.

Finally, from a purely technical standpoint, the application of OCR to historical text tends to produce defects that do not follow the patterns found in OCR data obtained from contemporary text. Among the latter are relatively high numbers of words split into two parts by a phantom period or blank space, and the letter *m* recognized as a sequence made up of an *i* and an *n* (Lopresti 2009). Experience has shown that applying algorithms designed to fix errors in modern OCR data to historical corpora yields mediocre gains. For example, an attempt to repair words

that OCR had split in two only resulted in 322 repairs over 22 years' worth of *Gazette* issues, an improvement of less than 0.05% in overall corpus accuracy.

From Bad Data to Good Science, Step 1: Heuristic Project Design

Turning our attention back to the original question: what *is* a digital historian supposed to do with this Bad Data? The key insight lies in the fact that not all Bad Data is created equal. Some corpora, because of their internal structures or because of their contents, can be made to answer specific historical questions – and nothing else. Others may be more flexible. To probe the boundaries of what can be achieved with a given corpus, we need to follow a process of heuristic project design, which means letting the data guide the selection of a research question, through visual and computational exploration of the corpus, rather than choosing the question *a priori*.

This is useful for two reasons. First, exploration may summon from the corpus realities that are more relevant than any prior hypothesis. In the words of French scholar Damon Mayaffre: “The hypothesis-deduction method is dangerous both because of the risk of projecting artifactual realities onto the text and because it can obliterate real facts too numerous to be handled by human memory” (Mayaffre 2002, 158, my translation). For example, mining the *Encyclopédie* for articles that mention the four major parts of the world as conceived in the eighteenth century (America, Europe, Asia, and Africa) suggests a link between America and botanical vocabulary, something that would not be obvious to a casual reader. Eventually, part of the current author's study settled on asking why America was so closely associated with plants in the *Encyclopédie*, and whether a similar pattern would hold true for other types of natural resources. The second reason to favour heuristic project design is that, when clear patterns emerge from exploratory analysis, they focus attention towards potentially crucial parts of the data. At worst, this focus helps the scholar show that the emerging pattern is merely the consequence of a data error, as was the case with the geographic articles in Volume XIII of the *Encyclopédie* mentioned earlier. Sometimes, however, the patterns can reveal a small set of keywords and concepts worthy of study, thus showing which errors need to be corrected before useful interpretations can be derived and which do not. In other words, a heuristic

approach to the data essentially neutralizes much of the noise in a corpus by removing its non-crucial parts from consideration. For example, when mining the *Gazette* for traces of the French colonial Atlantic, a much higher number of mentions of British and Portuguese colonies emerge at the same time. This requires explanation: why would a French periodical be so much more concerned with non-French America than with the kingdom's own colonies? This unexpected finding led to a study of the ways in which the colonial Atlantic was covered in the *Gazette* and in the other periodicals of the time, which eventually yielded a thesis chapter and an upcoming peer-reviewed article (Laramée forthcoming) about the peculiar lack of enthusiasm for migration to the colonies shown by the French under the Old Regime. Only 23 keywords, most of them place names, were required to perform this research. Many other avenues of inquiry, and the keywords associated with them, have been set aside for another time.

From Bad Data to Good Science, Step 2: Focused Error Correction

Designing a research question that can be answered by mining a large corpus for a small number of keywords naturally orients error correction efforts towards making sure that the presence of these keywords is measured accurately. This means finding and fixing, through some sort of fuzzy search algorithm, as many misspelled and badly recognized keyword tokens as possible, so that every article, page or paragraph relevant to the research question can be extracted from the corpus. (At this time, it may not be obvious whether this extracted content will itself be suitable for digital analysis or whether the scholar will have to examine it through close reading, but the extraction process is the same in either case.)

Some Web-based resources, like ARTFL, may include their own fuzzy search engines. When dealing with raw text files downloaded from an archive like Gallica, however, the scholar must apply their own solution. Levenshtein's algorithm (Levenshtein 1966), which defines the distance between two strings as the number of characters that must be deleted, inserted or swapped to transform one into the other, provides an easily customizable model. **Table 1** illustrates the types of calculations that the algorithm performs.

Given a list of keywords, it is a relatively simple matter to write a Python script that leverages a pre-packaged Levenshtein distance module (such as the *editdistance* module available on the standard Python repositories) to scan a corpus for strings that closely match the keywords. This, however, generates a large number of false positives that can only be identified through visual inspection of the results. For example, the application of Levenshtein's algorithm to the raw *Gazette* OCR data obtained from Gallica yielded no fewer than 2,956 different string types, representing a total of 16,297 tokens, that were separated from one of the 23 keywords in the colonial Atlantic study discussed earlier by a Levenshtein distance of three or less. Of those, only 148 turned out to be actual keyword variants, including 35 at distance zero (the keywords themselves and some frequent alternate spellings that had been identified during the exploratory phase) and 103 at distances between one and three. This means that a visual inspection of the 2,956 candidate string types eliminated 95% of them. This was easier than one might think: the vast majority of these discarded candidates were either French words themselves and therefore unlikely to represent misread keywords, or else nothing more than random OCR detritus from which nothing could be recovered. The word *musique* (music), which stands at a Levenshtein distance of 3 from *Amérique* (America), is an example of the first case; a string made up of the letter *a* repeated six times, which stands at a Levenshtein distance of 3 from *Canada*, is an example of the second.

Table 1: An illustration of Levenshtein distances.

Word Type #1	Word Type #2	Levenshtein Distance	Operations required
Amérique	Amérique	0	None
Amérique	Amrique	1	Deletion of <i>é</i>
Amérique	Cmévrique	2	Replacement of <i>A</i> by <i>C</i> Insertion of <i>v</i>
Amérique	Musique	3	Deletion of <i>A</i> Replacement of <i>é</i> by <i>u</i> Replacement of <i>r</i> by <i>s</i>

Once the 148 keyword variants were identified, each of their tokens was examined in context in the source material to determine whether it actually represented a keyword and not some alternate meaning of the same word. This second-pass inspection was designed to eliminate potential sources of confusion in the data, such as the ones involving *Halifax* and *colonie* mentioned earlier in this article. Strictly speaking, this step is not required if there is no risk of confusion, but this is a condition that is hard to guarantee ahead of time. (The existence of Cardinal Coloni, for example, came as a complete surprise to this article's author.)

Table 2 summarizes the results of this two-part process as applied to the *Gazette*. Overall, Levenshtein's algorithm was able to recover 1,867 tokens of the 'canonical' tokens themselves, plus 532 damaged or misspelled keyword tokens of 103 different types, which resulted in an increase of 29.5% in the total number of tokens compared to a simple perfect-match search. In total, 2,399 keyword tokens (damaged or not) were found in 1,184 different articles.

Note that, while most of the variants of keyword tokens recovered by this method are the results of one-, two- and three-character OCR errors, the method is equally adept at finding unexpected but correct keyword spellings. For example, the algorithm found 143 instances of the unaccented word type *Bresil* (Brazil), against only 5 occurrences of the accented form *Brésil* used in twenty-first century French. This latest example is particularly telling of the method's value: while the application of Levenshtein's algorithm to the raw OCR data uncovered relevant articles about every part of the colonial Atlantic, Brazil's importance in the corpus would have been vastly underestimated without it.

While the process described in this section of the article was designed to handle a few dozen keywords and several thousand documents, it is relatively straightforward to adapt it to different contexts. If the list of keywords is very long, for example, extracting candidate word types that are only separated from a keyword by a Levenshtein distance of 2 or less might provide an acceptable compromise, since the number of such candidates is approximately ten times smaller than for a distance of 3 and experience has shown that only a handful of candidates at distance 3 turn out to be useful. (In

Table 2: Results of the application of Levenshtein's algorithm to the *Gazette* corpus.

Canonical type	Canonical tokens	Recovered types	Recovered tokens	% of recovered tokens	Sample recovered types
Amérique/ d'Amérique/ l'Amérique	485	36	128	20.9%	l'amerique (59), d'amcrique, ramérique
Acadie	3	1	15	83.3%	l'acadie
Antilles	1	0	0	0%	n/a
Boston	56	2	8	12.5%	bofton, b^fton
Brésil	5	10	159	97.0%	bresil (143), bretil, brcfil
Canada/ Canadiens	139	3	3	2.1%	en.canada, canada*
Cayenne	8	0	0	0%	n/a
Colonie(s)	411	15	16	3.7%	5lonie, coioniej
(Saint) Domingue	88	11	13	12.9%	jjomingue, saintdomingue
Guadeloupe	48	4	4	7.7%	guadecoupe, quàdeloupe
Halifax	7	3	43	86.0%	hallifax (14), d'hallifax (19)
Jamaïque	343	5	5	1.4%	jamai-, jamàlque
Louisbourg	9	3	51	85.0%	louifbourg (48), louisbôurg
Louisiane	10	1	1	9.1%	louiiifane
Martinique	160	4	4	2.4%	martinique**
Montréal	13	1	2	13.3%	montreal
Philadelphie	69	1	1	1.4%	philadelphie.
Québec	12	3	79	86.8%	quebec (77)
TOTAL	1867	103	532	22.2%	Net gain 28.5%

the case of the *Gazette*, approximately 50% of the recovered tokens were at distance 1, 45% at distance 2, and only 5% at distance 3.) It may also be a good idea to run the algorithm more than once with short distances instead of once with a longer distance, augmenting the list of keywords with frequent alternative spellings after each iteration.

This is how *l'amérique* and *d'amérique* were included in the keyword list, which allowed a second pass to find a handful of misread tokens separated from *l'amérique* by a distance of 2 but from the original *amérique* by a distance of 4. Finally, if two keywords are very similar, such as *Russie* (Russia) and *Prusse* (Prussia), some candidate word types may lie within a short Levenshtein distance of both. A good rule of thumb, in this case, is to assign the candidate to the keyword to which it is closest; if there is a tie, the decision must be made through a judgement call, possibly one token at a time, after visual inspection of the token in context.

From Bad Data to Good Science, Step 3: Resolution

Now that the corpus has assisted in the process of heuristic research question design and that targeted error correction has solidified our understanding of the presence of a certain number of crucial keywords in the corpus, it is time to return to the source material. Broadly speaking, the project has reached one of four states, depending on the prevalence of the remaining defects and on their relevance to the research question.

In the best-case scenario, token counts for the corrected keywords themselves (or statistics that can be directly computed from them) are the signal that needs to be measured to answer the research question. Now that the keywords have been found and fixed, whatever defects remain in the corpus are irrelevant. For example, while searching for the message that print media transmitted to the lower classes of French society about the opportunities awaiting them in Old Regime colonies, the cheap and widely distributed books of the *Bibliothèque Bleue* are eloquent by their silence. The word type *Canada* only appears five times in this entire corpus, four of them in the names of plants found in a 1707 gardening manual. Louisiana, Acadia, Martinique and Guadeloupe are entirely absent. Saint Domingue, Europe's richest colony in the eighteenth century, is mentioned once. The word *Amérique* appears 17 times in total, including 5 times in a 1782 geography manual whose author merely describes America as a part of the world unknown to the Ancients and where strangers often get sick. For the humble readers of these books, the French Atlantic is all but invisible. Inasmuch as they can find a message in this silence, it is that the New World is none of their concern.

In the next best case, the keywords are not the answer but they allow us to extract a subset of the corpus that can be treated as a reasonable approximation of Big Data from which to derive the answer. To qualify as ersatz Big Data, the errors that remain in this subcorpus must be either relatively rare, randomly distributed, or irrelevant to the research question. In the *Encyclopédie*, for example, a set of several dozen keywords representing the four major parts of the known world (America, Africa, Asia and Europe) and some of the better known Atlantic World colonies of the eighteenth century allows us to extract a collection of 6,053 articles that refer to one or more of these parts of the world. The text of these *Encyclopédie* articles is of high quality and can safely be submitted to any number of textometric methods, provided that the scholar is aware of the spelling issues mentioned earlier in this paper. Linguistic specificity analysis (**Figure 2**), for example, shows that articles about America feature an abnormally high number of verbs in the present tense, such as *sont* and *est* (to be), *font* (to make), *trouve* (to find) and *servent* (to serve), compared to the rest of the world. Asia and Africa, on the other hand, show high linguistic specificity for verbs conjugated in the past tense (**Figure 3**), whereas America shows high *negative* specificity for the same words. Correspondence analysis and topic modeling of the same subcorpus also show that, while commerce and natural resources emerge as salient topics everywhere, botany has an abnormally high presence in America's subcorpus, with words like *feuilles* (leaves), *arbre* (tree) and *fruit* being vastly overrepresented when compared with other continents. Thus, in the *Encyclopédie*, America is mostly portrayed as a young land of current events, rich in living wealth represented by the plants that must be catalogued and studied before they can be exploited in colonial plantations. (A more complete discussion of the *Encyclopédie's* geography can be found in Laramée [2017].)

All may not be lost even if the keywords can neither answer the research question themselves nor point towards clean data that does. The keywords may be able to focus the scholar's attention towards a subset of the corpus which, through visual inspection, may reveal hidden information that can solve the research question indirectly or help to reorient the project in a more suitable direction. As mentioned

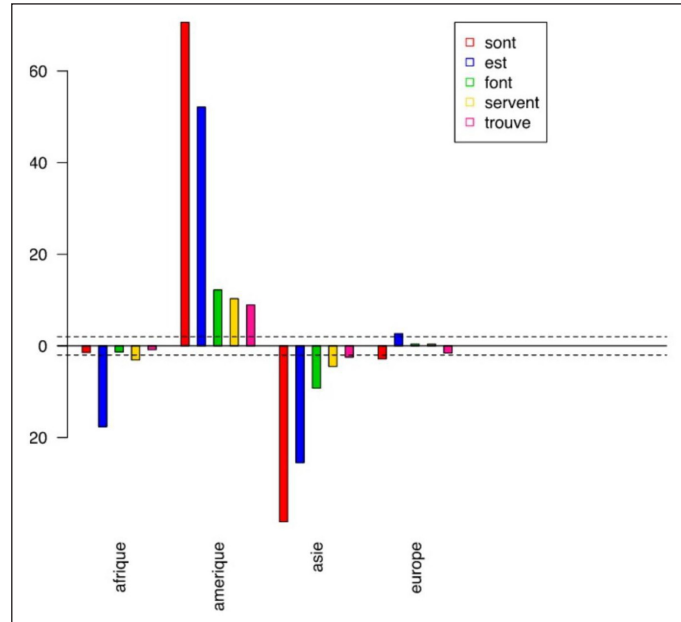


Figure 2: Lexical specificity of present-tense verbs in the *Encyclopédie*. Articles about America show high positives.

earlier, mining twenty-two years' worth of *Gazette* issues for articles discussing Atlantic World colonies yielded a relatively small set of 1,184 articles. The OCR data in these articles is too noisy for most purposes, and the idiosyncratic nature of *Gazette* articles, which read like diaries of unrelated events rather than carefully constructed narratives, makes techniques like topic modeling irrelevant anyway. However, visual inspection of the articles shows that far more of them seem to originate from abroad than from France itself. This information is invisible in the OCR data because of an error cluster, since articles' cities of origin appear in headlines, which the *Gazette* printed in italics, which OCR has difficulty understanding. It is, however, a relatively simple matter to retrieve this information visually, from the PDF versions of the periodicals, and to include it in a metadata file that also contains, for each article, computationally extracted values like year of publication and sentinels indicating the presence or absence of each keyword. This metadata file, created through a mixture of hand-crafting and calculation, contains reliable data that can be studied while the unreliable raw text is set aside. A k-means classification of the 1,184 articles into five

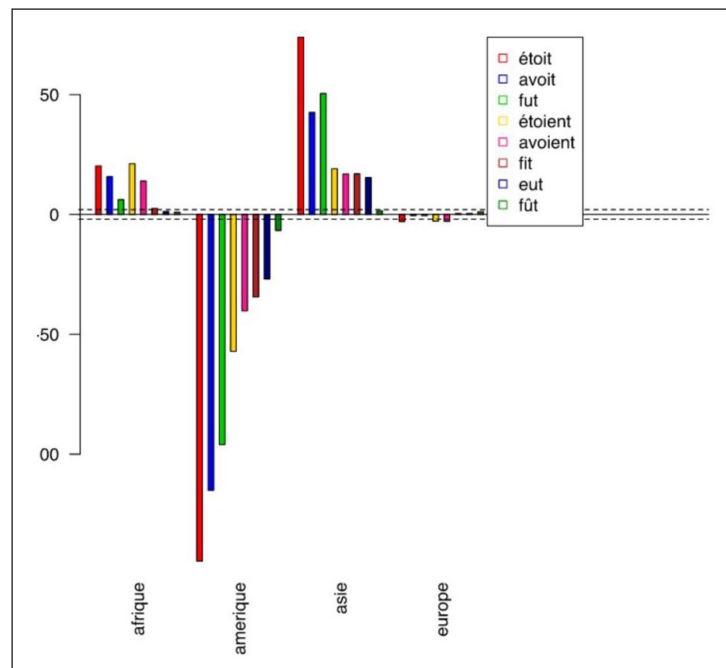


Figure 3: Lexical specificity of past-tense verbs in the *Encyclopédie*. Asia shows high positives; America, high negatives.

classes, based on the contents of this metadata file, confirms the suspected pattern. Four of the five classes, including the ones characterized by the presence of the keyword *Amérique* and by the presence of the keyword *colonie*, are overwhelmingly composed of articles emanating from abroad; many of them appear to be translations of material copied from English periodicals, translated so faithfully that they use first-person constructs such as *notre* (ours) and *nous* (us) when discussing British fleets and colonies. In only one class out of five do the names of French colonies appear more often than those of their foreign rivals (**Figure 4**), and even in this class, sources of French origin are in the minority compared to news briefs sent to the editor from London (**Figure 5**). Thus, a dirty corpus has indirectly revealed that the *Gazette* presents the Atlantic world to the French reading public as an essentially foreign phenomenon.

If all else fails, the keyword instances retrieved from the noisy data by Levenshtein's algorithm at least show the scholar a better picture of which parts

of the corpus to read. A modest achievement, perhaps, but in a large, noisy corpus where the keywords are relatively rare, such preprocessing of the sources can both increase coverage and save a considerable amount of research time. This was as far as digital processing could go for the *Mercure de France* and the *Journal des Savants*, with their very poor OCR and lack of metadata patterns similar to the one identified in the *Gazette* headlines. It still proved invaluable to the study covered in Laramée (forthcoming). Alternatively, the digitally-inclined scholar may want to reiterate the entire three-step process, redefining the research question or reframing the corpus until they become sufficiently compatible.

Conclusion

This article has shown that raw historical text, as a category, cannot be treated as Big Data. It also outlined a method that can neutralize the defects of this source material through an iterative process of heuristic project design, targeted error correction, and careful assessment of what can be computed from the two. The method relies not so

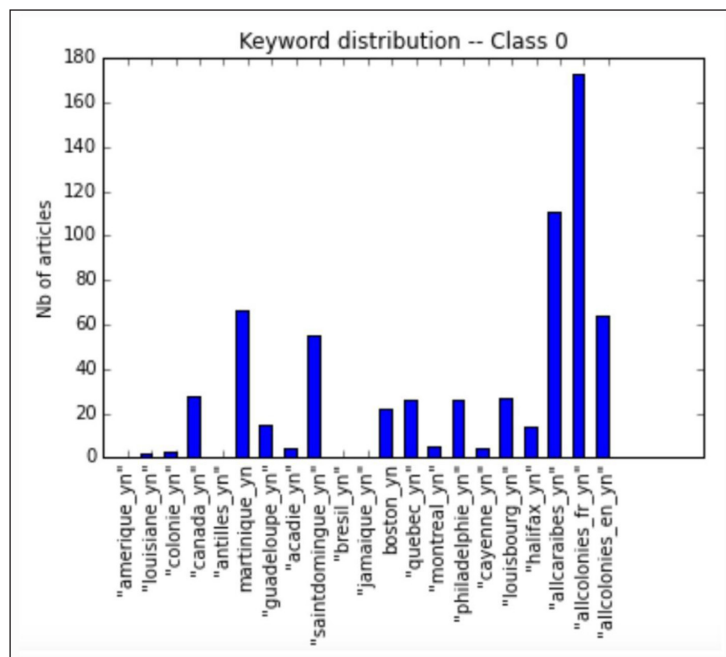


Figure 4: Keyword distribution in Class 0. This is the only class in which French colonies are mentioned more often than other colonies.

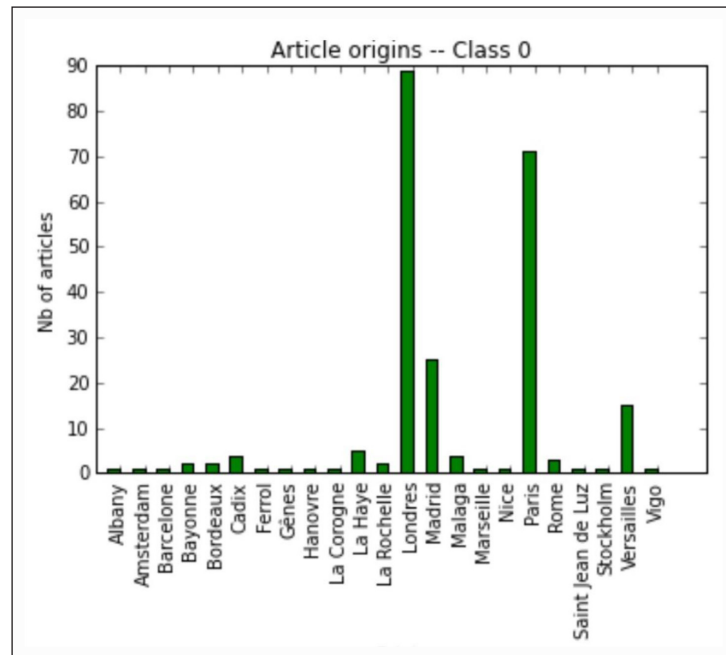


Figure 5: Origins of articles in Class 0. Even in this class, a majority of articles emanate from abroad.

much on technical legerdemain as on carefully crafted research questions that allow small subsets of the data to be isolated, repaired, or reconstructed by hand, while the rest is gracefully discarded. Thus transformed (at least locally) into a reasonable approximation of Big Data, some Bad Data corpora become, in the current author's experience, suitable for cartography, machine learning, textometrics, word vector analysis, and the like.

This method is somewhat labour intensive and relies on judgement calls at every step, which suggests a trade-off between the size of the corpus that can be mined, the internal structure of that corpus, the types of error clusters found within it, the size of the list of keywords that can be fixed, the frequency at which these keywords occur in the corpus, and the data quality in the raw text. A very large corpus made up of low-quality OCR data, for example, may not be compatible with a research question that can only be answered by looking at every instance of hundreds of ubiquitous and polysemic keywords. Within these parameters, the method has been applied to

several corpora and unrelated research projects. There is no reason to believe that it cannot be adapted across languages and time periods, as long as the language has clearly defined written word boundaries and uses an alphabet so that a Levenshtein distance between words can be computed.

Yet the prudent scholar will retain a healthy skepticism regarding results derived from Bad Data. Cross-validation of multiple experiments on a Bad Data corpus, involving different digital methods and visual confirmation of the results, is required to protect the scholar against software bugs and data accidents. Perhaps more importantly, the lower the quality of the original data, the stronger and more consistent across methods the results must be before they can be used to support, and only to support, humanistic interpretation.

A final word on reproducibility. It is easy to publish the general parameters employed in a given study, such as the list of original keywords, a table of additional keyword types and tokens identified using Levenshtein's algorithm, the number of keyword tokens that have been discarded from consideration as a result of judgement calls about polysemy, etc. However, the method outlined in this paper only provides a (very) partial cleanup of the source data. Further, what counts as a correction for a scholar's purpose, such as merging all of the spellings of *Louisiane* in the *Encyclopédie* into a single word type, may count as introducing even more noise for someone else's research. Thus, distributing the corrected data files to the community would be of limited value, except perhaps for those attempting an exact duplication of the original results. And of course, the judgement calls required at every step call into question the level of duplication that can be achieved anyway. Perhaps this should serve as a warning. In the current author's experience, digital history projects involving text reach the limits of what can be achieved through algorithmic approaches distressingly fast. In other words, the human era is far from over.

Acknowledgements

The author's research was supported by a doctoral scholarship provided by the *Fonds de recherche du Québec – Société et culture*. The author wants to thank his thesis advisor Susan Dalton, the 2017 CSDH/SCHN conference participants, and the anonymous reviewers for their insightful comments.

Competing Interests

The author has no competing interests to declare.

References


- ARTFL.** 2016. "Bibliothèque Bleue." Accessed December 21, 2018. <https://artfl-project.uchicago.edu/bibliotheque-bleue>.
- Benzécri, Jean-Paul.** 1992. *Correspondence Analysis Handbook*. New York: Marcel Dekker. DOI: <https://doi.org/10.1201/9780585363035>
- Bibliothèque nationale de France.** 2018. "Gallica." Accessed December 21, 2018. <https://gallica.bnf.fr/>.
- Cibois, Phillipe.** 2007. *Les méthodes d'analyse d'enquêtes*. Part of collection. *Que sais-je?* Paris: Presses universitaires de France.
- Heiden, Serge, Jean-Philippe Magué, and Bénédicte Pincemin.** 2010. "TXM: Une plate-forme logicielle open-source pour la textométrie – conception et développement." In *Proceedings of the JADT 2010 Conference*. Rome: Edizioni Universitarie di Lettere Economia Diritto.
- Laramée, François Dominic.** 2017. "La production de l'espace dans l'Encyclopédie: Portraits d'une géographie imaginée." *Document Numérique* 20(2–3): 159–77.
- Laramée, François Dominic.** Forthcoming. "Migration and the French Colonial Atlantic as imagined by the periodical press, 1740–61."
- Levenshtein, Vladimir I.** 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Cybernetics and Control Theory* 10(8): 707–10.
- Lopresti, Daniel.** 2009. "Optical Character Recognition Errors and Their Effects on Natural Language Processing." *International Journal on Document Analysis and Recognition (IJ DAR)* 12(3): 141–51. DOI: <https://doi.org/10.1007/s10032-009-0094-8>
- Mayaffre, Damon.** 2002. "L'Herméneutique Numérique." *L'Astrolabe. Recherche Littéraire et Informatique*, 1–11. special issue.
- Mayer-Schönberger, Vikto, and Kenneth Cukier.** 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. First paperback edition. London: John Murray.

- Morrissey, Robert.** 2016. "Introduction to the ARTFL Encyclopédie." University of Chicago: ARTFL Encyclopédie Project. Accessed December 21, 2018. <http://encyclopedie.uchicago.edu/node/16>.
- Morrissey, Robert, and Glenn Roe.** (eds.) 2017. "Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers." Eds. Denis, Diderot, and Jean Le Rond d'Alembert. *University of Chicago: ARTFL Encyclopédie Project*. (Autumn Edition). Accessed December 21, 2018. <http://encyclopedie.uchicago.edu/>.
- Piotrowski, Michael.** 2012. *Natural Language Processing for Historical Texts*. Lexington, KY: Morgan & Claypool. Available online. Accessed December 21, 2018. <http://www.morganclaypool.com/doi/abs/10.2200/S00436ED1V01Y201207HLT017>.
- Schmid, Helmut.** 1997. "Probabilistic part-of-speech tagging using decision Trees." In *New Methods in Language Processing*, edited by Daniel B. Jones, and Harold Somers, 154–64. London: Routledge.

How to cite this article: Laramée, François Dominic. 2019. "How to Extract Good Knowledge from Bad Data: An Experiment with Eighteenth Century French Texts." *Digital Studies/Le champ numérique* 9(1): 2, pp. 1–24. DOI: <https://doi.org/10.16995/dscn.299>

Submitted: 02 February 2018 **Accepted:** 20 December 2018 **Published:** 30 January 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Digital Studies/Le champ numérique* is a peer-reviewed open access journal published by Open Library of Humanities. **OPEN ACCESS** 