The *Maeſtro di Muſica*, or Quality Control in the Virtual Library[1]

Eleanor Selfridge-Field[2]

English Abstract

Many readers will have noticed bizarre mistakes in unverified scans of textual sources available with or without license on the internet. Those searching for sources describing music—whether in periodicals, books, program notes, lyrics, or the broad miscellany of documents pertinent to music—may suffer disproportionately from aberrant spellings in scanned resources. This enquiry formulates three levels of inaccuracy and presents representative errors. Among other findings, it can be shown that certain widely used textual repositories systematically fail to differentiate between running text and tables, illustrations, footnotes, and other hallmarks of the scholarly apparatus. The problems described have known solutions, but librarians and scholars must make a stronger case for remedies, and for verification of scanned texts.

French Abstract

De nombreux lecteurs auront remarqué des erreurs bizarres dans les numérisations non vérifiées de sources textuelles qui sont disponibles sur Internet, avec ou sans licence. Ceux qui recherchent des sources décrivant la musique – que ce soit dans des périodiques, des livres, des notes de programme, des paroles de chansons ou dans la large gamme de documents pertinents à la musique – risquent de souffrir de manière disproportionnée des orthographes aberrantes que l'on retrouve dans les ressources numérisées. Cette étude formule trois niveaux d'imprécision et présente des erreurs représentatives. Entre autres résultats, il est démontré que certains dépôts textuels largement utilisés échouent systématiquement à faire la distinction entre le texte courant et les tableaux, les illustrations, les notes de bas de page et les autres éléments de l'apparat critique. Il existe des solutions connues aux problèmes décrits, mais les bibliothécaires et les chercheurs doivent de leur côté insister pour que des correctifs soient apportés et pour que les textes numérisés soient vérifiés.

German Abstract

Vielen Lesern werden schon bizarre Fehler in unredigierten Scans von Textquellen aufgefallen sein. Diese kommen sowohl in kostenpflichtigen als auch in kostenfreien Datenbanken vor. Gerade bei der Suche in Musikquellen – sei es in Zeitschriften, Büchern, Programmheften, Textvorlagen oder in einer der vielen weiteren einschlägigen Dokumentenarten von Musikmaterialien – stolpert man überproportional häufig über Schreibfehler in gescannten Texten. Diese Untersuchung beschreibt drei Ebenen von Ungenauigkeiten und stellt typische Fehler vor. Neben anderen

[1] *Il maestro di musica* was a common antecedent phrase in *opera buffa* titles in the eighteenth century. There was a wide range of consequents, introduced by the word "or". Misrepresentation of social station or musical skill provided the foundation for the comic plot. Valuable advice has been contributed to this article by Ilias Chrissochoidis and Maureen Buja.

[2] Eleanor Selfridge-Field is consulting professor of music at Stanford University and research director of the Center for Computer Assisted Research in the Humanities (CCARH). She teaches music informatics to students in several disciplines, maintains a historical research agenda in Italian music, and has given several recent lectures on digital humanities topics. She is the author of six books, the editor of sixteen, and a contributor to many journals in musicology (most recently *Early Music, Journal of Interdisciplinary Musicology, Notes,* and *Musicae Scientiae*).

Ergebnissen wird aufgezeigt, dass bestimmte weitverbreitete Text-Repositorien grundsätzlich nicht die Möglichkeit bieten, zwischen Text und Tabellen, Illustrationen, Fußnoten oder weiteren wesentlichen Bestandteilen des wissenschaftlichen Arbeitens zu unterscheiden. Für die beschriebenen Probleme gibt es durchaus Lösungen, aber Bibliothekare und Wissenschaftler müssen größeres Augenmerk auf Abhilfe und auf die Korrektur gescannter Texte legen.

George Bernard Shaw famously summed up his frustrations with the irregularity of English pronunciation by explaining that the word “fish” might as well be written *gh-o-ti*—*gh* as in “enough”, *o* as in “women”, and *ti* as in “nation”. When noting the misconstructions of Google Books' optical character recognition, one could easily believe he is being confronted by a similarly perverse logic, but, unfortunately, there is no logic to scanning errors. They cannot be explained by orthography or phonology. Optical-recognition software seeks to categorize the shapes of letters, to interpret them by their physical location, and to refine them by contextual clues. Its results slowly improve but are rarely perfect. Google takes pride in the superior quantity of its scans but evidence of quality control or of efforts to incrementally improve performance seem to be indefinitely lacking. Here we document some common impediments to searches for documents discussing music.

Of the two common defenses of scanning errors, the major one is that “only a few” exist. Here one needs to understand the scale of the metric. Quoted rates of accuracy sound respectable on an academic scale of 1–100. The claims have slowly risen from 88% to 92%, 96%, and so forth. This is usually gauged against a simple text—an office computer script, a legal document, or a similarly regular writing. To the naked eye of a scanner, documents come in many levels of graphical complexity. Tables, illustrations, large blocks of white space, footnotes, and inconsistent type quality will all affect accuracy. A simple metric is illusory. Recognition Metrics, an OCR consultancy near Seattle focusing on recently created documents, explains an accuracy rate of 98% as representing a single page of 2,000 characters in which 40 will be incorrect.[3] Google Books, in contrast, attempts to render whatever is on its virtual shelf. A hypothetical error rate of 40/page can mean 40 words/page without a dictionary match. Google Books has greater difficulty with early (*c.* 1500–1825) publications than with modern ones. Many books were set in larger type than is customary today, but local variations existed. Early typography favored bigger margins and careful centering. Calculating error rates for early books is not possible without knowledge of page formats (quarto, octavo, etc.) as well as margin allowances, fonts, etc. In manual encoding, texts are verified by sight or through double entry and comparison.[4] When neither produces an acceptable result, language-specific search-and-replace routines can spot and fix most errors. From a lexical perspective, most scanning errors are so predictable that they can systematically be located, then filtered by language and typography.

The second defense of “a few errors” in scanning is that recognition software is ostensibly “trainable”. We examined this point in the context of musical notation in a controlled test of optical music recognition in *Computing in Musicology*.[5] One program consistently

[3] See http://www.primerecognition.com/cost_justification.htm.

[4] Double transcription and comparison is a process whereby 2 separate encoders separately prepare the same text. Divergences revealed through comparison lead to necessary corrections. Studies from the 1980s confirmed a near-perfect result from this method.

[5] Eleanor Selfridge-Field, “Optical Recognition of Musical Notation: A Survey of Current Work,” *Computing in Musicology* 9 (1993–94), 109–145; same author, “How Practical is Optical Music Recognition as an Input Method?” *Computing in Musicology* 9 (1993–94), 159–166.

misplaced bar-lines. How does one quantify that kind of error? The object is present, but when many notes wander into the wrong measures, the cost of correction is high. In text as well, some objects deserve to be weighted more heavily than others. Among these initial letters of new sentences, paragraphs and words merit higher weights to reflect their role in segmentation. Google also omits certain special characters. To judge from the number of times CCARH has had to report copyright violations to Google Books' legal department, one deduces that recognition of the copyright sign © (being curiously absent in some reverse title-page scans of our own books of the 1990s) is beyond the capabilities of Google's recognition software to detect.

Error categories in book-text recognition

Errors can be grouped into three general categories according to their impact. These are the misreading (1) of single letters within words; (2) of groups of letters that may make single words unrecognizable; and (3) of errors so numerous that the text is unintelligible. Errors of the first kind can usually be eliminated (in principle) by systematic orthographic search-and-replace functions. Errors of the second kind are often arbitrary in nature. Since they cannot be anticipated, they elude systematic correction. Errors of the third kind may altogether obscure the language of the text. Once a sentence or two is completely off-track, it is unlikely that accuracy will improve. Table 1 illustrates the first two kinds of errors. Table 2 shows words containing these letters in early printed books.

Search term	No. of matches	Missing match: text
Mufik	776,000	Abdruck feiner ganzen Vortreffl ic**lg**keit, feines ächt menfchliche**tu** ächt künfilerifchen Charakters, Form und Inhalt aber finden **fiets** den wahrfien, anfprechendfien, befriedigend**fien** Ausdruck. Wir nennen Mozarts ***Mufik*** kla**ffifch**. (Heinrich Sattler, 1856)
Jefus	371,000	See text and « Jefu meine Freude » entry below for examples.
La même chofe	322,000	Mon **guieu** [Mon sieur ?], Piarrot [Pierrot], tu m**j** vient toujou dire *la même chofe*. PIERROT. Je te dis toujou *la même chofe*, parce c'eft toujou *la même chofe*, & **fi** ce n'étoit pas toujou *la même chofe*, je ne te dirois pas toujou *la même chofe*.[6]
Maeftro di Mufica	292,000	See main text.
Efpagna	119,000	«Tout le monde fait la fortune immenfe que Farinelli a faite en Efpagne» . "...fymphoniej dediée à M**gr** le Comte de Noailles, Grand d'Efpagne...."[7].
maeftro	88,400	See main text.

TABLE 1 Examples of **f>s** substitutions and miscellaneous errors in online search, in declining order of frequency in November 2014, are in bold type. Numbers and quotations come from Google Books unless otherwise specified. The emphasis is on spelling errors in the rendering to scanned texts where specialists would see correct renderings in now unfamiliar typographical formations.

[6] Quoted from "La Festin du Pierre," in ***The Works of Moliere in French and English*** (London: Watts, 1748), p. 274. Why "toujours" is consistently truncated [as "toujou"] is unclear.

[7] Quoted from ***Le Mercure de France*** (Mai 1768), p. 171.

Search term	No. of matches	Missing match: text
Mufic	29,500 [vs "music" in Google search: 15,100,000]	« Triju'gum (i. iff old recordi) The junfdiftion [jurisdiction] of three hundreds. TRILATERAL (ad<. from tbt Lat. tres tbrety and latus a fidt) Having three fides. Trilat'eralnels [s. from trilateral) **Tbt** quality of having three fides. Scott. TRILETTO (1- in ***mufic)* A fhort trill**. [Consecutive entries (run together) from John Ash: *The New and Complete Dictionary of the English Language* (1775), unpaginated.][8]
Meifter	23,800	See main text.
Univerfidad	14,800	« hizo á efte Colegio Mayor, ya la Univerfidad para las Cathedras, defpuesde agre- gar á cftas el Beneficio de Yecla; no folo por los tavo- res, que con tanta bizarria hizo áquantos individuos de éftos dos Iluftres Cuerpos á S. Erna- acudieron; fino porque de nueftra Univerfidad fue S ... ».[9]
«Jefu meine Freude»	5,960	*Jefu. meine Freude* _ ll. 306. 322. 323. Jefu komm. mein Trofl und Lachen — ll. 480, 552. Jefu. Kraft der blöden Herzen - ll. 514. 515. M-V. 11. Nr. 185. Jefu Kreuz. Leiden und Pein - l. 502. M-V. l. Ne. 156. Jefu Leiden. Pein und Tod - l. 122. - lll.[10]
La Mufique	1,297[*gallica.bnfr.fr*]	Summary: Found in bibliographies, periodicals (*Le Mercure galant*), commentaries on art (Vasari), military endeavors, an edition of Rousseau's letters with a response by Madame de Staël, et al.
"Ma maitreffe"	63 [*gallica.bnfr.fr*]	"Extrait 1: et Rude aux voleurs doux à l'amant 1 » J'aboyais & tailàis careilè » Ainfi j'ai fu diverfement u Servir mon maître **8c** ma maîtreffe ».' Sonnet de la belle Matineufe.[11]

TABLE 1 continued

Class-1 errors

The single most common misreading in *Google Books* is the substitution of the letter **f** for **s** [hereafter **s>f**].[12]

It is especially prevalent in works published up to about 1825 anywhere in Europe or North America. Because scanning is so dependent on letter-shape, a high degree of consistency can be found across cognates in Latin-alphabet languages.[13] This has a substan-

[8] This quotation runs together a sub-entry under TRIGYNOUS through TRILATERAL to the end of TRILETTO.

[9] *Oracion funebre panegyrica* (Seville: En la Imprenta de la Universidad, 1744).

[10] From contents listing for Carl von Winterfel[d], *Der evangelischen Kirchengesang und seine Verhältnis zur Kunst des Tonsatzes* (Lepzig: Breitkopf und Härtel, 1847).

[11] Quoted from La Borde [writing as « Onfroy »], *Essai sur la musique ancienne et moderne* (Paris : De l'Imprimerie de Ph.-D. Pierres) t. 4, 1780.

[12] Between the fifteenth and nineteenth centuries the letter **s** existed in at least three forms (often simplified to two in modern discussions). Early English and French exemplars often extend to both upward and downward. The intermediate version (Old Dutch, Renaissance Italian) extended upward but not "below the line" of most characters. The **s** in seventeenth- and eighteenth-century English and North American colonial typography was more notable for its lack of a cross-piece than its upward extender. (Both of these forms are classified as belonging to the "long s" class.) The round **s** was determined less by locale than by function within a word. It was used especially for initial and terminal positions, while the other form with used in most interior positions, sometimes modified to parse the word itself. For example, on Felix Mendelssohn's tombstone the surname (rendering long **s** here as **f**) reads Mendel**sf**ohn. This tells us that the successive s's belonged to different syllables, while in a word such as "recess" or "progress" both **s**'s would be long and thus resemble "reseff" and "progreff".

[13] Polish is an outlier (because of its large number of diacriticals). German *Fraktur* is problematical both because of overlapping ascenders (**b, d, f, h,** et al.) and descenders (**g, p,** and **y**) and because of decorative tendrils distracting the "eye" away from a letter's essential shape. Specialized software enables optical recognition

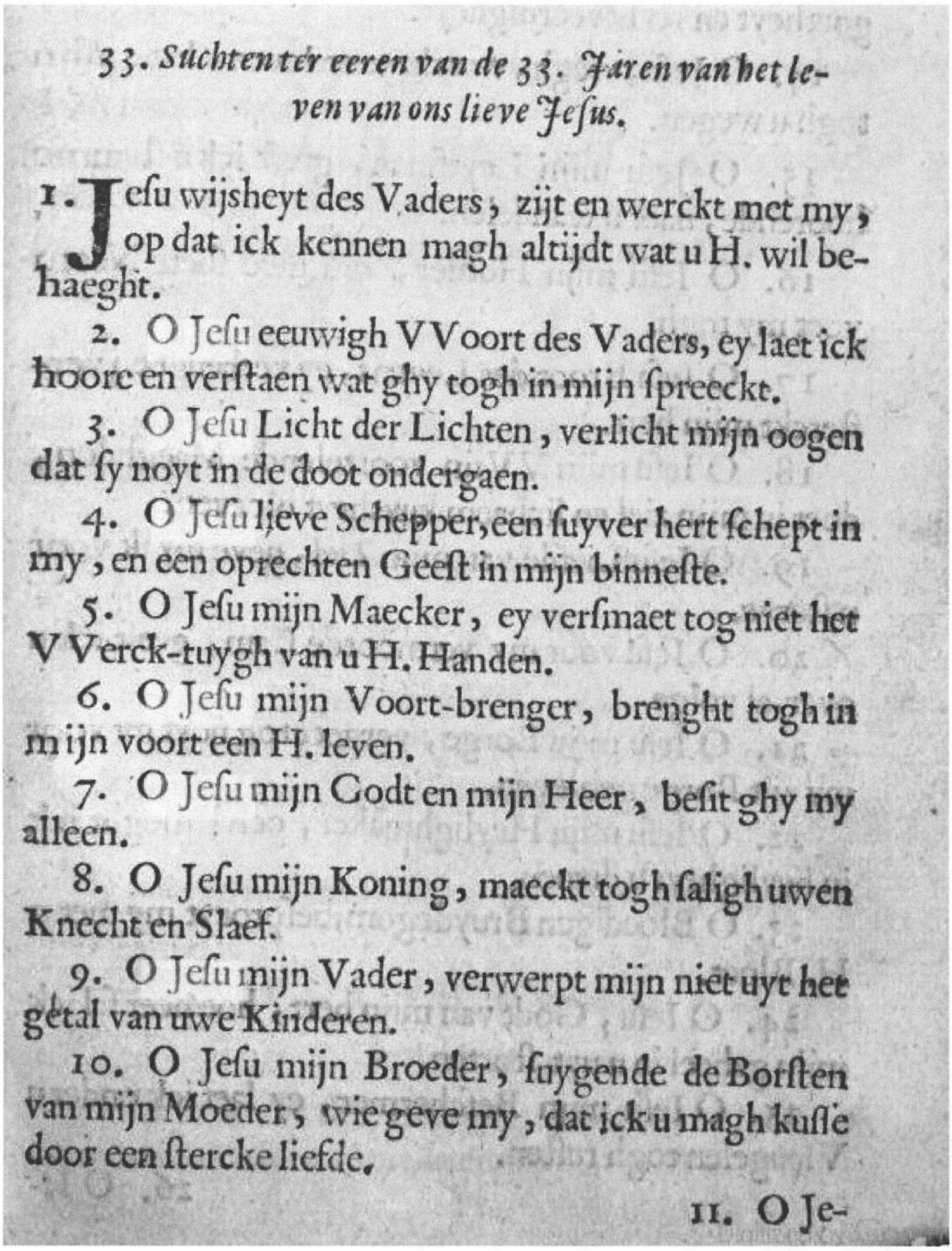
Year	Language (place of publication)	Focus of image	Image	Transliteration	Source
1668	Latin	ae ligature	Græcas	Graecas	*Historiae Alexandri*
	(Vienna)	Lower-case **s**	Helleſponto	Hellesponto	*Magni...*
		ct ligature	cuncta	cuncta	
		Lower-case **f**, lower-case **s** (x3)	frigidiſſimus	frigidissimus	
1606	Italian	Lower-case **s** (x3)	ſteſſa	stessa	*Preparatione*
	(Venice)	Lower-case **f**	fedeliſs.	fedeliss. [=fedelissimo]	*dell'anima alla divina gratia*
1614	Italian	Lower-case **s**	Caſtello	Castello	*Lettera annua dal*
	(Rome)	Upper-case **F**	Febbraro	Febbraro	*Giappone del 1614*
		Upper-case **s**	Signore	Signore	
1708	English	Lower-case **s**	laſt	last	*The British Apollo, or,*
	(London)				*Curious amusements*
		Upper-case **F**, lower-case **s**	Froſt,	Frost	*for the ingenious....*
1762	French	**ff** ligature	difficulté	difficulté	*Journal ecclésiastique*
	(Paris)	**ss** ligature	expreſſior	expression	*ou Bibliothèque raisonnée, vii/3*
1601	French	Lower-case **f**	fait	fait	*Actes de la*
	(Evreux)	**st** ligature	Requeſte	Requeste	*conférence tenue entre le sieur Evesque d'Evreux...*
1740	Spanish	**st** ligature	Mageſtad	Magestad	*Coleccion de los*
	(Madrid)	**ss** [in successive syllables]	aſsiſten	assisten	*tratados de paz...* Part II
		ss [in one syllable]	Miſſas	Missas	
1600	English	**fl** and **ct** ligatures	afflict	afflict	*Titus Andronicus*
	(London)	**ss** ligature	diſſembled	dissembled	*partly by William*
		fl and **sh** ligatures	flouriſht	flourisht	*Shakespeare: The*
		st ligature	ſubſtance	substance	*First Quarto*

TABLE 2 Original appearance of the letters f and s plus selected ligatures in books printed before 1800. Exact details varied by a letter's position in a word, by font, and by publisher. A transliteration and brief indications of year, place of publication, and title are given.

tial impact on searches that involve the word "music" or its equivalents. The root is common to both Germanic and Romance languages. The readings "mufic", "mufique", "mufica", and "Mufik" seem to be ubiquitous. Google Books is not the only offender, simply the biggest. Google Translate cannot digest more than one or two instances of non-lexical results of its own scanning without launching into an endless loop.[14] The caveats for those searching for Psalm settings, hymns, and liturgical music can be summed up simply with

of Greek, Hebrew, and Cyrillic, which have finite numbers of characters but wide variation in their rendering. In Asian scripts Hiragana and Katakana syllables are manageable because of their finite number, but pictographs as found in Kanji and Mandarin pose big challenges. Languages based on cursive script (Arabic, Persian) present a range of different choices related to variability in letter formation and in use of interpretive marks.

[14] If one clicks the "translate" prompt shown with a citation that is obviously garbled, the "translate" software churns away until someone turns it off.



33. *Suchten ter eeren van de 33. Jaren van het leven van ons lieve Jeſus.*

1. Jeſu wijsheyt des Vaders, zijt en werckt met my; op dat ick kennen magh altijdt wat u H. wil behaeght.

2. O Jeſu eeuwigh VVoort des Vaders, ey laet ick hoore en verſtaen wat ghy togh in mijn ſpreeckt.

3. O Jeſu Licht der Lichten, verlicht mijn oogen dat ſy noyt in de doot ondergaen.

4. O Jeſu lieve Schepper, een ſuyver hert ſchept in my, en een oprechten Geeſt in mijn binneſte.

5. O Jeſu mijn Maecker, ey verſmaet tog niet het VVerck-tuygh van u H. Handen.

6. O Jeſu mijn Voort-brenger, brenght togh in mijn voort een H. leven.

7. O Jeſu mijn Godt en mijn Heer, beſit ghy my alleen.

8. O Jeſu mijn Koning, maeckt togh ſaligh uwen Knecht en Slaef.

9. O Jeſu mijn Vader, verwerpt mijn niet uyt het getal van uwe Kinderen.

10. O Jeſu mijn Broeder, ſuygende de Borſten van mijn Moeder, wie geve my, dat ick u magh kuſſe door een ſtercke liefde.

11. O Je-

ILLUSTRATION 1 The 1676 image comes from *Heyligh Tydverdryf: Bestaende in diverse Godtvruchtige Oeffeningen, door Jaer, Maent, Week en Dagh: Aen alle Godtminnende zielen op-geoffert; Om door deselve verdienstelijck te warden besteedt,* Vol. 1. (Antwerp: Pieter van Overlant, 1676), p. 111. The highlighted words in Nos. 2, 3, 4, 6, 7, 8 identify those instances recognized by Google Books as "Iefu". On multiple trials, No. 8 was sometimes highlighted, sometimes not.

the warning to be on the lookout for such non-words as "Ble**ff**ed", "Je**f**u", "Chri**ft**", "Ho**f**anna", "I**f**rael" and other common terms with a native **s**.

Class-2 errors: Unpredictable character misreadings

When two or more adjacent characters are misread in a single word, there is often a typographical ligature involved. Letter sequences that used to be joined into one physical character included **fi**, **ffi**, **li**, **lli** in English, **ae** [æ]in British renderings of words derived from ancient languages and the **ß** [originally **sz**] of formal German.[15] (See Table 2) In parallel with ligatures, diacritical marks usually appear in a single composite character (à, è, ê, et al.).[16] Google Search seems to lack any sense of how to spot misinterpreted characters even when, with a language filter, many illegal character combinations could readily be found.

This kind of error becomes especially problematic when it occurs in close proximity to an **s>f** conversion, since entire words and phrases may become irredeemably unrecognizable. In writings on music the **s>f** permutation has the greatest negative impact, it seems, on books in German. **Sc-**, **sch-**, and **-sf** are frequently replaced by such alexical constructions as **fc-** and **fch-**, or the viable but often misintended **–ft**. One example referring to C. M. von Weber's *Der Freischütz* yielded the snippet "... [Frei**fch**üß] was [war?] die deut**fch**e ***Mufik*** für die Bühne werden könnte. wenn **fie** ... in meiner zur Feier von Schillers hunder**tft**em Todestag er**fch***ienenen* ***Feftfchrift***.").[17] An **s** can also be misread as a **p**, **d**, or **t**. (The number of permutations is seemingly endless.) Consider this reported title: "Gott**f**ched: Gedanken vom Ur**t**prung und Alter der ***Mufik***; in de**ff**en kriti**fcd**er Gei**h**te**h**te [Geschichte?] der Dichtkun**fi** [-kunst] der Deut**f**chen. Leipzig. 1757."

Google's perennial exclusion of punctuation marks exacerbates the proper segmentation of words, phrases, and sentences.[18] However, punctuation marks are used liberally for unrecognized characters.[19] Punctuation specific to particular European languages—the Spanish inverted question mark (¿) or French quotes («...»), for example—may, together with currency signs and mathematical symbols, be sprinkled liberally (and inappropriately) throughout scanned texts.[20] Characters, numerals, and letters with similar shapes may be confused, as in misreadings of lower-case **l**, the numeral **1**, and the exclamation point **!**. Nonsensical anagrams for "The" include any middle letter that is as high as the "T": "T**b**e", "T**d**e", and so forth. At the start of titles and sentences **J** and **I** are regularly confused (they were represented by the same letter in early typesetting). In the listing of ten titles beginning with the word "Jesu" in Illustration 1, six are highlighted as matches,

[15] The Romanization of ***Fraktur*** in the nineteenth century lacked an appropriate ligature. In this instance recent books can produce more errors.

[16] Those who use Adobe® fonts will appreciate their support for joined characters continues unblemished, while word processors offer no support for ligatures. Bembo® is a particular favorite of those trying to imitate early typography and could be a useful base font for training recognition software intended for use with early books, although the objective is to recognize ligatures in any font.

[17] Mistaken punctuation replicates that in the screen view. The citation comes from ***Westermanns Monatscheft*** (1908), no page number shown.

[18] Punctuation marks can interfere with the indexing of n-grams—character strings of progressively larger lengths—which facilitate the profiling of word-usage statistics along a time-line.

[19] When, in 2011, Google introduced the cypher "+" to identify its Google+ social network, accommodations seem to have been made in its advanced search to obviate confusion.

[20] Scanning software does not in general admit to its defeat, although some ***Google Books*** texts are full of "?"s, that may or may not indicate the software was admitting confusion.

while four are ignored. These kinds of errors can rarely be anticipated. In Latin and Romance languages lower-case **v** was rendered as **u**, upper-case **U** as **V**. Here recognition of early printed texts carries an implicit obligation to modernize for the reader but to preserve for the scholar. Ultimately the purpose of recognition should be clarified. Serving many audiences simultaneously is not destined to produce results that will satisfy all of them.

Some letter-changes are too idiosyncratic to classify. One extract from Gio. Battista Martini's *Storia della musica* (1757) refers to "Joa**cbirn Qg**antz" [= Joachim Quantz] just before citing "Pier France/20 *Tofi* [= Opin. de' *Cantori*." The actual author would be Pierfrancesco Tosi, the work in question his *Opinioni de' cantori antichi e moderni* (1723). The surname Mendelssohn is particularly prone to distortion, as in "*Mendelfohn* - Bartholdy wurde Mufikdirector und übernahm die Leitung der Oper. Das Haus wurde renovirt und verziert und mit einer wenigftens anftändigen Außeufeite gefchmückt. **J**n kurzer Zeit entfiand unter diefer Leitung ein Theater".[21]

Class-3 errors: Gobbledygook

Gobbledygook can start out innocuously with a **b** substitution for **h** in "the", a **J** at the start of any sentence starting with an **I**, or the overuse of **?** and other punctuation marks for any unrecognized character. Small problems are compounded by the absence of spaces. Consider a citation from what proves to be the preface to an edition of Seneca from the year 1800. Google's snippet says this:

> "r K ^ L r ^ I' I 0. ViäsÄL ... ^nal. II, zi. 10H.) noto. ' Huo mre^ »lii via'eriin. Vi6e O^ttio^. Zel. ^112. „ «um *1800*. nu. 36. r>. 36a. NN8 6 t lloetilliinis, <nü c^nnni ex ni8ce epilta- lis.

The quotation comes (ostensibly) from Ruhkopf's "Praefatio" to the *Opera Omnia* edition of Seneca's works published by Weidmannische Buchhandlung, Leipzig. The passage is supposed to match Note 6 of the preface found in Vol. II, p. xii, which (in contrast to the snippet) reads:

> Cf. Wernsdorf I. 1. p. 12. Addi nunc potest Iunioribus aliis, a *W.* ibi allatis Iunio poeta, cuius epigramma elegans nuper primus protulit Ennius Quirinus *Visconti* in libro docto: Lettera su due monimenti, ne' quali è memoria d'Antonia Augusta p. 20. et vindicavit M. Pompeio iuniori, iam ex Anthologia (Brunk. Anal. II, p. 105.) noto. Quo iure, alii viderint. Vide Götting. gel. Anz. anni 1800., nu. 36. p. 360.

The passages do not exactly coincide, but some common material is faintly identifiable. What is clear is that the absence of a Google lexicon for bibliographical abbreviations con-

[21] From a 1940 study said to be by "Robert Blum and K Herlozsohn". The second name is not traceable nor, consequently, is the source. On further research, it may be that the source is *the Allgemeines Theater-Lexikon oder Encyklopadie alles Wissenswerthen fur Buhnenkunstler, Dilettanten und Theaterfreunde unter Mitwirkung der sachkundigsten Schriftsteller Deutschlands*, edited by R. Blum, K. Herlosssohn, H. Marggraff, etc., a multi-volume work published in Altenburg and Leipzig, Germany, between 1839–1946. (The second editor, Herlosssohn, suffers from the same transcription problems illustrated in this paper because of the repetition of the letter 's' in his name: in the original, the first two 's' were actually ß, which is replaced in modern German by 'ss'. Since this would have been written in Fraktur, it appears to modern eyes like a 'long s' followed by a 'z', or even just an elaborate 'z', hence the transcription error.)

tributes to the derangement of the text.[22] The more fundamental problem seems to be that Google's scanning algorithms cannot differentiate running text from footnotes, and so, in this case, has simply run to text to footnote without cognizance of the independent verbal contexts.

Systematic errors in other large digital collections

JSTOR is generally above the fray in scanning errors, but it is not free of a few persistent defects.[23] Any number of JSTOR listings, even for recent articles, have **s>f** substitutions combined with other bizarre misspellings, but for numerous reasons the overall rate is much lower. Once in a while JSTOR completely misfires, as in this example:

> ..., por Fr. Francifco Xi- menez, hijo del Conuento de S.Domingode Mexico, Natural de la Villa de Luna del Reynode Aragon. A , **bie R&.** P. Maeftro Fr. Hermando **Bana,Ppior** Prouincalde 14 Protincia de **S, iidio** de Mexic,**de** 1 Orden de 1ie F redicadoer,e yCatbedratic hubiladode Tbeologia eI Il 1 **niMe,fdad**....[24]

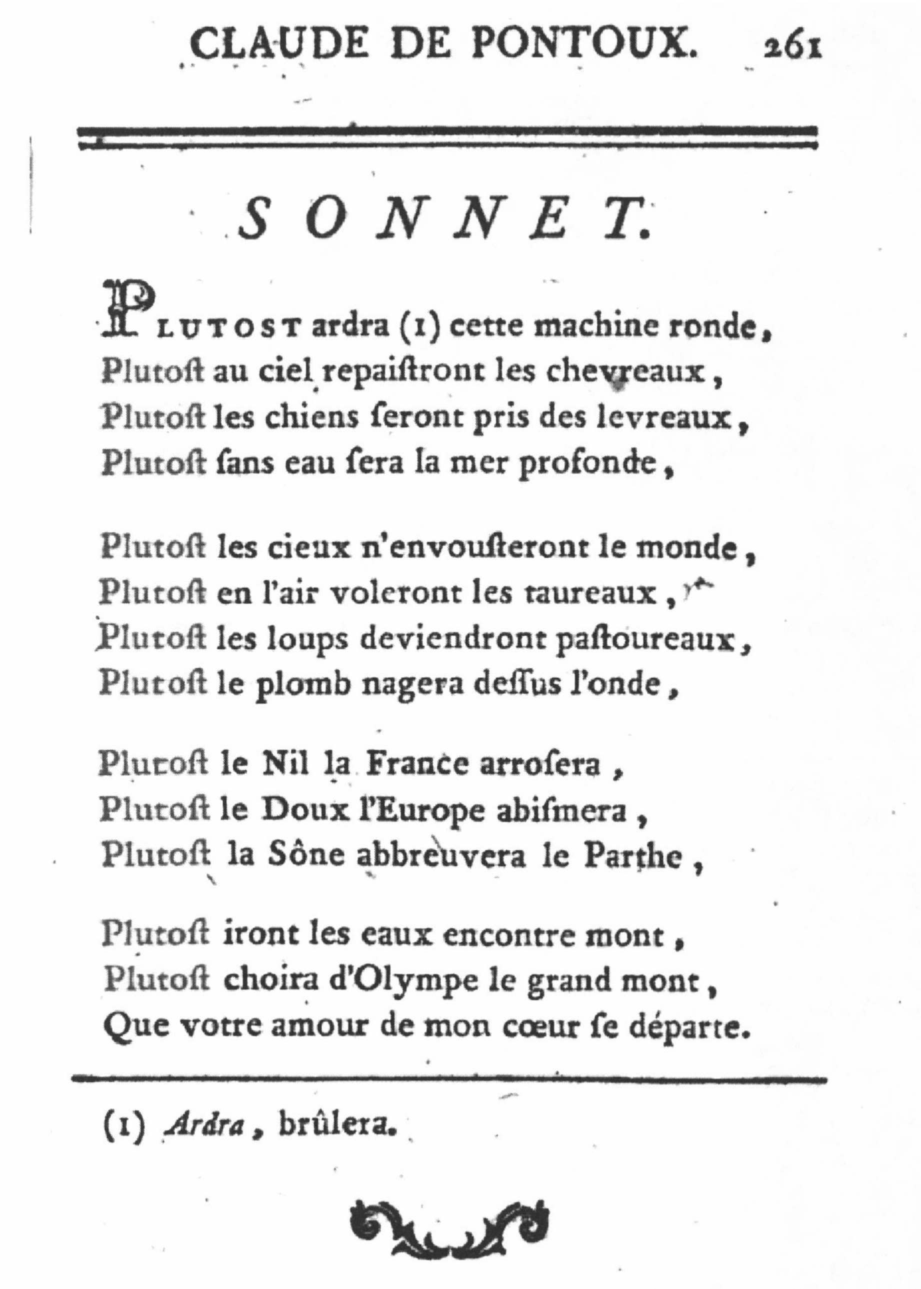
The quotation comes from a facsimile of a 1615 title-page (*De la Natura raleza, e Virtudes de las plantas,* i.e., a book on botany). The title-page was shown as an illustration in a modern article that was labeled a "match" in a search for the word "arias". The original title-page text of the work carried an elaborate dedication to Francisco Ximenes and to "N.ro [Nuestro] R. P. Maestro Fr. Hernando Bazan, Prior Provincal de la Prouincoia de Sa[n]ctiago de Mexico, de la Orden de los Predicadores, y Cathedratico Iubilado de Theologia en la Vniuersidad Real" [Our Rev. Father Hernando Bazan, provincial prior of Santiago of Mexico, from the order of preachers and professors of theology in the Royal University]. Facsimiles of title-pages from early prints within modern publications present a consistent trap comparable with that of abbreviation. Spurts of garbled text occur in any number of JSTOR republications of recent articles from journals such as *Early Music,* which also reproduces title-pages similar to this one.

Gallica (http://gallica.bnf.fr), which offers a cross-medium search engine spanning early and recent prints, manuscripts, images, and sound files, is not directly comparable with others. The extreme care it gives to difficult projects, such as its exquisite (and easily found) scans of illuminated manuscripts of Machaut's poetry and Cavalli's operas, for example, demonstrate a high regard for both quality and retrievability. Because it includes a large number of early printed books, Gallica offers an interesting antidote to Google

[22] Another snippet from the same work contains the phrases "*^'a lqU;don°PPO"ihlr'^'a lqU;don°PPO"ihlr'*" and " *7Hbehs facjee, OmПe^o ^11Г,Г^Г que eft*". These were not retrievable in a literal Google search, presumably because of the exclusion of non-alphabetic marks in search input. (An alternative scan of the same work is available on request from the National Library of the Czech Republic via Europe's Books2ebooks with the listing found at http://search.books2ebooks.eu/Record/nkcr_stt20110031756.)

[23] A useful account of JSTOR's formative years is provided in Chapter 4 of Roger C. Schoenfeld's *JSTOR: A History* (Princeton, NJ: Princeton University Press, 2003). It divulges many details of the quandaries encountered in JSTOR's development. Scanning errors make up a small part of the picture when the contributions of intermediate technologies, storage media, graphical detail, and vendor particularities are factored into the picture. Preferences also vary by discipline. The original scientific model required accommodation for humanities journals.

[24] This example comes from Rafael Chabrán and Simon Varey, " 'An Epistle to Arias Montano': An English Translation of a Poem by Francisco Hernández," *Huntington Library Quarterly,* 55/4 (1992), pp. 621–634. This match responded to a search for the English term "arias".



CLAUDE DE PONTOUX. 261

SONNET.

PLUTOST ardra (1) cette machine ronde,
Plutoſt au ciel repaiſtront les chevreaux,
Plutoſt les chiens ſeront pris des levreaux,
Plutoſt ſans eau ſera la mer profonde,

Plutoſt les cieux n'envouſteront le monde,
Plutoſt en l'air voleront les taureaux,
Plutoſt les loups deviendront paſtoureaux,
Plutoſt le plomb nagera deſſus l'onde,

Plutoſt le Nil la France arroſera,
Plutoſt le Doux l'Europe abiſmera,
Plutoſt la Sône abbreuvera le Parthe,

Plutoſt iront les eaux encontre mont,
Plutoſt choira d'Olympe le grand mont,
Que votre amour de mon cœur ſe départe.

(1) *Ardra*, brûlera.

ILLUSTRATION 2 The 1778–79 image comes from *Annales poétiques, ou Almanach des muses, depuis l'origine de la poesie française.* Vol. 7. (Paris: De la lain, 1778–79) p. 261.

Books: it contains very few errors of the kinds discussed here. It has relatively good success in avoiding the pitfalls of archaic French.[25]

Archive (http://www.archive.org) is much older in origin and still more heterogeneous in the range of materials it provides. Its lapses are far fewer than those of Google Books, but some of the categories into which the errors fall are the same.[26] One persistent glitch shared by Archive and JSTOR is an inability to suppress hyphens used in line segmentation when searching for single words. A search for an author named Gastone Vio in JSTOR encounters numerous "matches" for "vio-" in contexts in which the following word is "loncello". Case sensitivity would clearly go some distance in fixing the problem.

Evaluating incidental errors found in Google Search that match writings on third-party websites rather than in Google Books is not straightforward. However, a strong resemblance to lapses in Google Books will be noted. A random search for letter transpositions turned up these two versions of the same passage from Ephraim Chambers' *Cyclopædia, or, An Universal Dictionary of Arts and Sciences* (1728):

a. "The fixth Chord of BaSs-Viols, and the tenth of large Theoobos, confift of 50 Threads, or Guts : There are Some of them 100 Foot long, twisted and polish'd with....";
b. "lerrawit obferves, that of late they have invente, C changing the Chords, to render their Sound mor without altering the Tone. fixth Chord ot Bafs-Viols, and...."[27]

In these cases the content is unambiguous, and it is available to the user. Whether the user will be enticed by such misinterpretations to view it is open to question.[28] The second quotation comes not from the original four-volume work (1728) but from a 1753 supplement found in a separate PDF at the same Wisconsin web location. The Wisconsin digital search engine provides said page in response to the (local) Boolean search "fixth" and "Chord".

Remedies

The sad part about the survival of so many ragged passages is that tools to remedy most of their defects are available. ABBYY FineReader offers what it calls "Historic OCR" for now unfamiliar kinds of typography. It has an alluring "before and after" example at its "Frakturschrift" page: http://www.frakturschrift.com/en:start.[29] The example adds in its

[25] E.g., by correctly rendering the **s** in "plutost" (rather than presenting plutoft) before the word became "plutôt", cf. Illustration 2.

[26] Within *Archive*'s multiplicity of formats instances of "claffical" music together with such words as "preferve", "fuch", and "inftitution" are ubiquitous in *.txt files but do not necessarily occur in corresponding passages in more finished formats.

[27] In modern English: "The sixth string of bass viols, and the tenth of large theorbos, consist of 50 threads or guts: There are some of them 100 feet long..." and so forth. The first quotation comes from *Chambers' Cyclopaedia* as found at the ARTFL server at the University of Chicago—http://artflsrv01.uchicago.edu/cgi-bin/philologic/getobject.pl?c.0:2364. The second quotation, at the University of Wisconsin, Madison, comes from http://digicoll.library.wisc.edu/collections/HistSciTech/Cyclopaedia.

[28] The Wisconsin case in particular merits comparison with the Google paraphrase. See http://digicoll.library.wisc.edu/cgi-bin/HistSciTech/HistSciTech-idx?type=turn&id=HistSciTech.CycloSupple02&entity=HistSciTech.CycloSupple02.p0895&q1=fixth&q2=Chord.

[29] Those interested in technical information will find it at http://www.frakturschrift.com/_media/en:white_paper_gothic-fraktur_ocr_e.pdf. Digital librarians will be pleased to note this addendum: "...improvements achieved in processing documents mean that today's OCR software can also be applied to image collections and historical documents that are already scanned."

summary that "tuned and optimized recognition technologies have to be used when processing historic documents printed in old fonts." At the same time ABBYY Historic OCR offers a discussion of "challenges" that were studied in the European Libraries IMPACT [IMProve ACcess to historical Text] project.[30]

The carefully curated Deutsches Text Archiv (http://www.deutschestextarchiv.de/), in which only two matches for "Mufik" could be found, has a built-in safeguard against nonsense. It shows the original text and the modern script side-by-side, which allows the user to easily identify any lapses. On a more general note, *The Signal*, an online blog of the Library of Congress's digital preservation program, offers a rigorous, detailed account of optical recognition and its efficiencies—when done consistently and well.[31]

In ordinary text-search on a single server, it would normally be possible to employ operators and delimiters (the "regular expressions" of the Unix grep tool) that would compensate for most spelling idiosyncrasies in Google Books. Because most characters used in grep queries are off limits in Google Search,[32] users may prefer to explore other search engines. The grep expression "[ch]at" would find all instances of "cat" or "hat" (the square brackets identify an either/or set). Likewise a search for "mae[fs]tro" would find all instances of both "maeftro" and "maestro". Table 3 offers a short list of the operators (e.g., AND, OR, NOT) supported by some common search engines to help narrow down the results. A comprehensive introduction to the subject of operator usage in search engines is available in a 2011 PowerPoint presentation by Paul Barron.[33]

	Google: Advanced Search; Developer	**Bing Query Language; MS Fast Query Language**	**DuckDuckGo (private web search)**	**Structured Query Language (SQL database search)**	**Yandex Advanced Search**	**Yahoo Advanced Search**
Logical (Boolean) operators (AND, OR, NOT)	Yes	Yes (alt OR = "|")	Yes, plus include/exclude commands	Yes	Yes	Partial: AND, OR [not = "-" followed by term to be excluded
String operators (BETWEEN, IN, NOT IN)	Limited (emphasis on titles)	No?	A few (e.g. CONTAINS)	Yes, plus additional ones	Equivalent	No?
Proximity (NEAR)	Shows context but without controls	Yes	No?	Equivalent	Yes	No?

[30] See http://www.frakturschrift.com/en:projects:impact.

[31] See http://blogs.loc.gov/digitalpreservation/2014/08/making-scanned-content-accessible-using-full-text-search-and-ocr/). This account discusses indexing, language-tuning, procedures to preserve metadata when corrections are made to recognized text and much else.

[32] Unix is particularly dependent on the verticule (|), which in *Google Books* results seems to be a random marker for unintelligible characters. Uses of this character in various programming contexts are discussed in the "Vertical bar" article in Wikipedia (http://en.wikipedia.org/wiki/Vertical_bar, accessed on March 18, 2015).

[33] "Advanced Web Searching for VEMAns," http://vaasl.org/pdfs/Conference_Handouts/2011/Barron%203.pdf. Barron is director of library and archives at the George C. Marshall Foundation.

Grammatical operators (for punctuation marks)	No	Selective	Yes	Yes	Yes	No
Search by date, date range	Yes	Yes	Yes	De factor	Yes	For email
Search by filetype	Yes	Yes	Indirectly	De facto		Yes
Search in URL	Yes	Yes	Indirectly	Not relevant	Yes	Yes
Wild card in search string	Yes but "removes some results"	Yes (weak results)	Yes	Yes	Yes	Yes (weak results)
Language filter	Yes	Yes	By changing user's "region"	Yes	Yes	Yes

TABLE 3 Permitted operators in selected text-search environments.

Data repositories that emerged in the decades before Google and newer archives that consist entirely of material entered by hand have the advantage that their holdings contain exactly what their users entered—and verified. No instance of "maeftro" or other misspellings cited here will be found in most curated collections, nor in Wikipedia. Some repositories do, by intention, provide exact transcriptions that capture the wondering spellings of earlier centuries. Notational errors in music manuscripts are faithfully recorded in all the RISM databases, for example. A text equivalent would be the *Early English Books Online* database (http://quod.lib.umich.edu/e/eebo?key=title;page=browse;value=ar). Among its 25,000+ titles, the 1600 print of Shakespeare's *Much Ado about Nothing* reads: "Much adoe about nothing. As it hath been sundrie times publikely acted by the right honourable, the Lord Chamberlaine his seruants. Written by William Shakespeare."[34] Scholars can turn to such sources to appraise the state of usage at a particular time without cringing when they see the word "seruant" because what the modern eye sees as deviations as the proper forms of printed language at the time of publication.

While Google Books is a great boon to many scholarly endeavors and indisputably saves many trips to a physical library, its rough texts impose a degree of inconvenience when accuracy and precision are required. The Advanced Search form for Google Books enables search by ISBN, publisher, and year of the print (all possible assets for the eventual resale of scanned out-of-print titles[35]), but they do not provide an adequate means of overcoming the errors described here. Dan Cohen's "Is Google Good for History?" (2010) is one of the most comprehensive and diplomatic evaluations of the strengths and weakness of Google Books.[36] As the executive director of the Digital Public Library, Cohen offers extensive praise, but he perceptively questions Google's possible privatization of aspects of its celebrated open-access model. Cohen defends the company on the ground

[34] http://name.umdl.umich.edu/A11991.0001.001.

[35] The confidential perception now exists among librarians who were among the first to allow Google access to their collections that Google's own enthusiasm for the project has waned as its "market potential" has remained elusive.

[36] See http://www.dancohen.org/2010/01/07/is-google-good-for-history/comment-page-1/.

that their aim was to work quickly. To do the job well, he supposes, might have taken a century instead of a decade. He objects, though, to the lack of availability of research data and bulk downloads.[37]

An earlier appraisal (2009) by Geoff Nunberg ("Google Books: The Metadata Mess") noted other kinds of errors, the most bizarre—a proliferation of books published in "1899" by living authors—having been fixed.[38] Nunberg lamented the hopelessness of genre classification for literature, noting that *Jane Eyre* surfaces under the rubrics of autobiography, governesses, love stories, architecture, antiques, and collectibles. In music this is a more complicated issue.[39]

Yoav Goldberg (Bar Ilan University) and Jon Orwant (a manager of Google Books) presented a case of their n-gram approach to "a very large corpus of English Books" in a 2013 paper entitled "A Dataset of large syntactic n-grams over Time..." based on a linguistic analysis of 345 billion words.[40] Their aim was to produce a usage timeline for designated terms.[41] The rise and fall of word usage is a perennial matter of interest to lexicographers but not one that is widely shared by most humanities scholars. "Big data" studies such as this one intermingle gleanings from texts the scans of which lie across a spectrum of accuracy rates. Humanities scholars generally want a result free of butchered words. The fact is, though, that Goole Books' own objectives would be better served by a higher degree of accuracy.[42]

The current state of fidelity of scanned early books to their physical originals suggests that we need the kinds of tools for search than we find mainly in curated repositories. In fact textual scholarship may be more efficiently served *qualitatively* by repositories that have existed since the days of mainframe computers. The Oxford Text Archive [http://ota.ox.ac.uk], established roughly 40 years ago, supports text search in 25 languages (ancient and modern) and includes the earliest encoded texts of Shakespeare, Milton, and the Bible plus numerous other writings studied by scholars. Project Gutenberg's book cata-

[37] In response to Cohen's post, Brandon Badger of Google Books pointed out that [Google's] epubs contain the optically recognized data that linguists would like to use, whereas PDFs contain only page images. (N.B. Recent efforts to access that data according to Badger's advice did not yield searchable results.)

[38] Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle of Higher Education*, 31 April 2009 (https://chronicle.com/article/Googles-Book-Search-A/48245/); rev. as "Google Books: The Metadata Mess," Presentation at the *Google Books Settlement Conference*, University of California, Berkeley, 28 August 2009, (http://people.ischool.berkeley.edu/~nunberg/GBook/GoogBookMetadataSh.pdf). The theme is newly expanded in Diana Kichuk, "Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books," *Libraries and the Academy* 15/1 (2015), pp. 59–91 (DOI: 0.1353/pla.2015.0005).

[39] Genre in music is a more vexing problem and one less susceptible to semantic remedies, given that in the popular/country/folk sphere *Billboard Magazine*, which is the arbiter of popular categories, has been accused of manipulating its classifications to stimulate sales of lagging "genres". For Google Books' approach to music see the pertinent section of their sitemap: http://books.google.com/sitemap/Sitemap/Music.html.

[40] *Second Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Atlanta, Georgia, USA (2013)*, pp. 241–247.

[41] Time-lines are also in course of implementation in JSTOR's bibliometric *Data for Research* project, on which see http://about.jstor.org/service/data-for-research. Since music cannot be isolated as a discrete subject area in JSTOR, these are currently of limited value. Further documentation can be found at http://about.jstor.org/sites/default/files/misc/Search_Documentation.pdf

[42] The mission statement of Google Books (accessed on March 18, 2015 at http://books.google.com/intl/en/googlebooks/library/) asserts that the aims are to "make it easier for people to find relevant books ... [and] "to create a comprehensive, searchable, virtual card catalog of all books in all languages".

logue, in process of development since 1971 (http://www.gutenberg.org/catalog/), consists entirely of materials (again in numerous languages) curated by volunteers. The project design is an early harbinger of "crowd-sourcing." Even when all licensed-access database holdings are added to these repositories, the ***quantity*** war has clearly been won by Google Books. Google offers simplicity of search and universal access (the latter degraded at times by disregard for copyright restrictions).

Who will win the quality war? We may want to consider whether parts of today's "digitized" world will, in a distant future, be seen to belong to a primordial past. Google's huge investment in Google Books seems to be undermined by its indifference to improvement. All projects founded on scanning face the risk of achieving a value inversely proportional to their error rates. The need to insist on intellectual rigor in our growing digital libraries looms large on the humanities horizon.