Preprint: *Fontes Artis Musicae* 16/2 (2015), xx-yy.

The *Maeftro di Mufica*, or Gremlins in the Virtual Library

Alfred Einstein 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 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 searching 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.[1] *Google Books*, in contrast, attempts to render whatever is on its virtual shelf. A hypothetical error rate or 40/page can mean 40 *words/*page without a dictionary match. *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. 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 margins allowances, fonts, etc. In manual encoding texts are verified by sight or through double entry and comparison. 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

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

controlled test of optical music recognition in *Computing in Musicology*.[2] One program consistently 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 had had to report copyright violations to *Google Books*' legal department, one deduces that recognition of the sign **©** (curiously absent in some reverse title-page scans) is beyond the capabilities of Google's engineers 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.

Class-1 errors

The single most common misreading in *Google Books* is the substitution of the letter **f** for **s** [hereafter **s**>**f**]. 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.[3] This has a substantial impact on searches that involve the word "music" or its equivalents. The root is common to both Germanic and Romance languages. The readings "mu**f**ic", "mu**f**ique", "mu**f**ica", and "Mu**f**ik" 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 an endless loop.[4] The caveats for those searching for Psalm settings, hymns, and liturgical music can be

[2] Vol. 9, 1993-94, ISBN 0-936943-08-4.

[3] 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 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.

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

summed up simply with the warning to be on the lookout for such non-words as “Ble**ff**ed”, “Je**f**u”, “Chri**f**t”, “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.[5] In parallel with ligatures, diacritical marks usually appear in a single composite character (à, è, ê, et al.).[6] 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, in 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 hundert**ft**em Todestag er**fch***ienenen Fe**ftf**chr**ift**.*”).[7] An **s** can also be misread as a **p**, a **d**, or a **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**f**c**d**er Ge**i**h**t**e**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.[8] However, punctuation marks are used liberally for unrecognized characters.[9] 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.[10] Characters, numerals, and letters with

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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). One listing of ten titles beginning with the word "Jesu" highlights seven with "J" and misses three more with "J". 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 modern reader but to preserve for the scholar. Ultimately the purpose of recognition should be clarified. Serving many audiences simultaneous 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** Q**g**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 "*Mendel**f**ohn* - Bartholdy wurde Mu**f**ikdirector und übernahm die Leitung der Oper. Das Haus wurde renovirt und verziert und mit einer wenig**f**tens an**f**tändigen Außeu**f**eite ge**f**chmückt. **J**n kurzer Zeit ent**fi**and unter die**f**er Leitung ein Theater".[11]

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 from the year 1800 of Seneca. 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.

[11] From a 1940 study said to be by "Robert Blum and K Herlozsohn". The second name is not traceable nor, consequently, is the source.

> 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 contributes to the derangement of the text.[12]

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.[13] Any number of JSTOR listings, even for recent articles, have **s>f** substitutions combined with other bizarre misspellings, but for numerous reason the overall rate is much lower. However, once in while JSTOR completely misfires, as in this example:

> ..., por Fr. Franci**f**co Xi**-** menez, hijo del Conuento de S.Domingode Mexico, Natural de la Villa de Luna del Reynode Aragon. A , **bie R&.** P. Mae**f**tro Fr. Hermando **Bana,Ppior** Prouincalde 14 Pro**t**incia de **S, iidio** de Mexic**,de** l Orden de 1ie F redicadoer,e yCatbedratic hubiladode Tbeologia eI Il l **niMe,fdad**....[14]

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.

[12] Another snippet from the same work contains the phrases "*^'a lqU;don°PPO"ihlr'^'a lqU;don°PPO"ihlr'*" and " *7Hbehs fac¡ee, 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.)

[13] A useful account of JSTOR's formative years is provided in Chapter 4 of Roger C. Schoenfeld's *JSTOR: A History* (Princeton, 2012). It divulges many details of the quandaries encountered in 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. An originally scientific model required accommodation for humanities journals.

[14] 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 (Autumn, 1992), pp. 621-634. This match responded to a search for the English term "arias".

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 (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 Books*: it contains very few errors of the kinds discussed here. It has relatively good success in avoiding the pitfalls of archaic French.[15]

Archive (http://www.archive.org) is much older in origins 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.[16] 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 **f**ixth Chord of Ba**S**s-Viols, and the tenth of large Theo**o**bos, con**fif**t of 50 Threads, or Guts : There are Some of them 100 Foot long, twisted and polish'd with…. ";

b. "lerrawit ob**f**erves, that of late they have invente**,** C changing the Chords, to render their Sound mor without altering the Tone. **f**ixth Chord o**t** Ba**f**s-Viols, and…."[17]

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.[18] The second

[15] E.g., by correctly rendering the **s** in "plutost" (rather than presenting pluto**f**t) before the word became "plutôt".

[16] Within *Archive*'s multiplicity of formats instances of "cla**ff**ical" music together with such words as "pre**f**erve", "**f**uch", and "in**f**titution" are ubiquitous in *.txt files but do not necessarily occur in corresponding passages in more finished formats.

[17] 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.

[18] The Wisconsin case in particular merits comparison with the Google paraphrase. See http://digicoll.library.wisc.edu/cgi-bin/HistSciTech/HistSciTech-

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 "**f**ixth" 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.[19] The sample adds in its summary that "the sample clearly shows 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.[20]

The carefully curated *Deutsches Text Archiv* (http://www.deutschestextarchiv.de/), in which only two matches for "Mu**f**ik" 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 plain, *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.[21]

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*,[22] users may prefer to explore other search engines. The 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 "mae**f**tro" and "mae**s**tro". Table 2 offers a short list of the operators (e.g. AND, OR, NOT) supported by some common search engines to support nuanced and

idx?type=turn&id=HistSciTech.CycloSupple02&entity=HistSciTech.CycloSupple02.p0895&q1=fixth&q2=Chord.

[19] 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."

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

[21] 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.

[22] Unix is particularly dependent on the verticule (|), which in *Google Books* results seems to be a random marker for unintelligible characters.

delimited searches. A comprehensive introduction to the subject of operator usage in search engines is available in a 2011 PowerPoint talk by Paul Barron.[23]

Data repositories that emerged in the decades before Google as well as 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 "mae**f**tro" 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."[24] 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 in a former time.

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 on 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), but they provide a 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*.[25] 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 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.[26]

[23] "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.

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

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

[26] 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 of page images. (N.B. Recent efforts to access that data according to Badger's advice did not yield searchable results.)

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.[27] 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.[28]

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.[29] Their aim was to produce a usage timeline for designated terms.[30] 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 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 tools 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 catalogue, 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).

[27] See also Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle of Higher Education*, April 31, 2009 (https://chronicle.com/article/Googles-Book-Search-A/48245/).

[28] 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".

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

[30] 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

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 *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.