Spring 2008

'Hussel,' 'Bussel' and 'Kussel,' Or, Using Google Books to Stalk the Elusive Alfred Russel Wallace

Charles H. Smith, *Western Kentucky University*

"Hussel," "Bussel" and "Kussel," Or, Using *Google Books* to Stalk the Elusive Alfred Russel Wallace

(Preprint of an article published in the Spring 2008 issue of *Kentucky Libraries*)

Charles H. Smith, Professor and Science Librarian, Western Kentucky University

From time to time the question arises as to whether the librarian, especially the reference librarian, should also be engaged in "librarian as scholar" activities--that is, in primary research activities. Among those who have written on this subject are Mark Winston, Cheryl LaGuardia, Monica Brooks, Stanley Chodorow, and Allen Kent, as a quick trip through *Google* or *Library Lit* will show. In general, the conclusion seems to be yes, for several main reasons: in conducting research themselves librarians: (1) become more familiar with the operational needs of their patrons (2) become, through the practice, more adept at search technique (3) encounter new tools and databases useful to aiding their patrons (4) add to the services provided by, and draw additional attention to, their library and university, and (5) find that those they serve develop higher levels of confidence in them. Here I should like to discuss one research adventure bearing especially on items 2 and 3 above, and concerning use of the relatively new service *Google Books*.

My original training was in the sciences and history of science, and even now as a librarian I spend a fair amount of time working in these directions. I have also gotten involved in creating and maintaining bibliography-centered websites that pertain to these subjects--especially, to the history of natural history. A key interest in this regard has been the naturalist and social critic Alfred Russel Wallace (1823-1913), best known (if he is remembered at all) as "the other man" in the history of the emergence of the natural selection concept in biology.

My Ph.D. was in geography, and more specifically biogeography. Biogeography is among the most interdisciplinary of all the natural sciences, being concerned with the study of "what plants and animals live where, and why": its relation to the modern biodiversity studies movement, accordingly, should be apparent. Biogeographers look at all the factors relevant to plant and animal distribution, including geological and evolutionary histories, current ecology, and anthropogenic influences. Wallace, in addition to being one of the co-discoverers of the principle of natural selection, is also regarded as the "father" of zoogeography, that division of biogeography which deals with the distribution patterns of animals.

As my interest in Wallace grew as a graduate student, I came to realize that his work extended to far beyond zoogeography and evolution--to aspects of geology and glaciology, astronomy, anthropology, sociology, land and economic reform, social criticism, etc., etc. He is also revered as the greatest field biologist in history, for his extraordinary collecting activities in South America and Indonesia circa 1848 to 1862. He

was a vocal spiritualist as well, and because of this and his constant criticisms of societal flaws and the powerful elite, many came to regard him as something of a crank (a Ralph Nader-like character, actually, beyond his many scientific contributions). By the end of his life he was perhaps the most famous scientist in the world, but his name declined rapidly after his death and it has only been recently that he has been revived to any extent.

Early on I realized that such a complicated person could only be fully appreciated if the entire body of his writings were taken into account. The sole existing bibliography of Wallace's publications was a 1916 compilation that listed around four hundred items; I soon discovered, however, that it was both inaccurate and very incomplete. I made it a personal object to start seeking out unreferenced and forgotten works of his, hoping to produce a much fuller bibliographic profile. I have so far managed to more than double that original figure of four hundred. Many of the rediscovered works are very minor pieces, but some are not, and have helped provide insight into the many remaining questions that exist regarding his philosophy and beliefs.

I have applied all manner of effort to the search, including going through a number of key serial publications page by page across periods of up to thirty years or more (Wallace's active publication career extended into eight decades), and checking the indexes of scores and scores of others (he is already known to have published in nearly three hundred *different* titles!). In the last ten years the widespread availability of electronic databases, including those focused on the nineteenth century, has significantly aided the cause. So too has *Google*: in 2006 a very early, unpublished, manuscript by Wallace I found through a website search for "correspondence OR letters" was deemed interesting enough to merit a special article in the prestigious scientific journal *Nature*.

And so we come to *Google Books*. Whereas *Google Scholar* has become the preferred tool of the scientist and librarian, *Google Books* represents a godsend to the historian. It took the folks at Google a while to get the system operating in earnest, but by the time I decided to investigate its full potential--this past summer--some one and a half million items had been scanned, with a sizable percentage of these actually being available for full viewing.

I soon found, however, that *Google Books* has a number of very annoying deficiencies. First and foremost, of course, one needs to remember that only a very small sample of *all* works--those that are no longer under copyright restriction--are included in the collection. This means, as it turns out, that only about ten percent of the totality of published works in our library collections is targeted for inclusion. (And, don't forget, Google still has a long way to go before it treats all of the volumes comprising even that ten percent.) Yes, a fair number, in absolute terms, of the authors of more recent works have given their permission to include their productions, but in relative terms these permissions represent a drop in the bucket. And, though in theory the authors (or their descendants) of many obscure works from the earlier or middle parts of the twentieth century might give their permission to be included were they contacted, Google is not about to invest the kind of time it would take to locate them and obtain such permissions.

It is also the case that a very large percentage of the books and serials that have been scanned are not actually accessible, though they can be searched for particular words and terms from the main search access point. For these many volumes, a "hit" will register for a particular search term and the page on which it exists in the work will be noted, but on pulling up the individual record for the work one finds that no further information can be obtained. Another very large percentage of the entire collection is set up to allow an item-specific secondary search, but only three hits and pages are specified (along with a line or two of surrounding text). In either case this might be enough to lead to an interlibrary loan request, but in most instances the dearth of contextual information produces a disappointing outcome when a loan is actually requested.

Another annoyance: though an increasing number of runs of serial volumes are being added, there is no way to search for a particular term within the full run of holdings of a *particular* serial title. Thus, if one searches "Alfred Russel Wallace," comes up with a hit on the title *Journal of the Linnean Society*, and proceeds to the detailed screen for that item, it is only possible to search for further hits within that one volume. Other available volumes of the series are listed, but one must then investigate them *individually* for further inclusions (and when the series includes dozens or scores of volumes, this becomes a real pain).

There are other things regarding coverage and entry into the system with which one might take issue, but let's move on--and I also leave it to the reader to explore the several interesting *positive* features that Google has created to enhance the service (but that usually help relatively little with the actual process of historical research). This gives us an opportunity to look in some detail at a critical weakness of the service: its patchy ability to retrieve materials matching up with a particular search.

Wallace turns out to be a particularly good subject for testing the capacity of databases, and *Google Books* in particular, to return searched-for information. First, he is both a moderately famous (and thus commonly represented) person in history, and one who was associated with a very wide range of subjects. There is also the matter of the peculiar (though not altogether rare) spelling of his middle name, with only one 'l'. Further, he is referred to in a number of ways that are correct ("Alfred Russel Wallace," "Alfred R. Wallace," "A. R. Wallace," "Russel Wallace," "Alfred Wallace," "A. Russel Wallace," "A. R. W.")--as well as some that are not, using the more common but in his case incorrect spelling "Russell." Equipped with this knowledge, I carried out some preliminary tests on *Google Books*. I concluded that his name was probably going to come up several thousand times if I made a concerted effort, and decided to split up the work into sub-searches limited by ten-year time periods, beginning in 1840 and ending in 1920.

I then proceeded through about ten searches (including variations where Wallace's last name came first, as "Wallace, A. R.") for each time period. This took weeks, at an hour or two a day. Some individual searches yielded as many as a hundred hits, and though most

of these were easily dismissed, others had to be investigated more closely (including, as mentioned above, those that required looking through individual volumes of serial titles).

Well into the project I began to have some doubts about the quality of the database I was searching, as certain combinations of terms seemed to be producing fewer results than they should have. This fact, combined with my observation that many of the pages I pulled up were of poor, sometimes unreadable, quality, led me to wonder just how reliable the scanning and optical character recognition (OCR) systems were that Google was using: it is one thing, of course, to say you are "making available" a wide range of materials covering a great span of time, but quite another to make the whole thing truly searchable.

Here I began to use all that experience I've had over the years searching for mis-shelved books. Mainly, what letters are most likely to be (electronically) mistaken for one another, and thus foil a well-intended search? What I found, starting from that point, is both interesting and revealing, if not entirely diagnostic of all the problems involved.

First, it must be admitted that scanning and OCR application are still an evolving science. It would not surprise me to find that Google's scanning technique falls rather short of the mark in terms of state of the art standards, as the human element in processing over a million volumes must be very considerable, and taken into account. At the same time, it *would* surprise me to find that the OCR systems they use are anything less than topnotch. But even assuming that, it was to be expected that much of the source material was just not of a clarity that could yield unambiguous results. And it doesn't.

Take, for example, a search on the name "Russel" alone. Okay, so you think you're going to be smart, and because there are few others of any fame during this period (I used 1780-1920, the wider spread to deal with journals that had started up before his time) with this name, that this should be sufficient to round up most references to him. I just now repeated this search, and it serves up 1770 hits. Fine. But, as it turns out, parallel searches employing "Hussel," "Bussel," "Kussel," and "Eussel" return, respectively, 634, 754, 678, and 654 hits, and the vast majority of these also deal with our Wallace. If one does a search of the form "Wallace AND Russel" over the same time period, things improve: 1710 hits for this combination, and with "Hussel," "Bussel," "Kussel," and "Eussel," respectively, 192, 174, 47, and 145 hits. But the search "A. R. Wallace" retrieves 801 hits, with the other ("A. H.," "A. B.," "A. K." and "A. E.") combinations scoring 138, 457, 240, and 337, respectively. Nor do inverted name searches using "Wallace, A. R." improve things much: it garners 708 hits, with the others totaling 83, 519, 134, and 151 (though for a couple of these, many of the hits are for other Wallaces).

By contrast, a search for "Alfred Russell Wallace"--incorporating the misspelled version of his middle name--yields 658 hits, while the substitutions of H, B, K, and E yield only 6, 6, 15, and 24. My conclusion from all this is that the OCR software is sophisticated enough to recognize slight deviations from an easily recognized word such as "Russell," but is less able to deal with variations on an uncommon word such as "Russel." It also

appears to do better with certain kinds of phrases than others. Despite the bad results noted above on the reversed phrase "Wallace, A. R.," a search on "Wallace, Alfred R." yields 457 hits, whereas the parallel searches with H, B, K, and E produce only 2, 9, 14, and 19 (and many of which are not actually our Wallace).

Oh, and did I mention that there are many other such substitution confusions that further complicate matters? For example, the search "Alfred Russcl Wallace" turns up 66 hits. A search on "Wallace" will produce 22100 hits, whereas parallel searches on "Wallaee," "Wallacc," and "Wallaec" return 678, 630, and 52 hits, respectively.

So, the moral is to think before you jump in. Whereas for many applications such fineries will not matter (in helping an undergraduate write a term paper, for example), for those doing professional level research the implications of haste may be a loss of as much as fifty percent or more of the potentially accessible information.

Am I trying to come down hard on *Google Books*? Actually, no. But you should be aware that strange proper names, whether they be people, places, or other things, may require special attention. Meanwhile, there's lots of potential for creative approaches to your charge as well. One particularly interesting one that I developed was to look for published letters by Wallace by using his mailing addresses as search terms--these often appear along with the body of the letter.

Overall I spent a total of perhaps fifty or sixty hours on the project, and so far have come up with about twenty-five "new" (actually, long-forgotten) writings *by* Wallace, dozens of "new" articles and reviews *about* him worthy of recording bibliographically, and at least that many finds again contributing productive facts and leads for further investigation; that is, some one hundred or more useful items in all. This ratio of time input to rewards received is very satisfactory, to say the least--but it also demonstrates that tracking down the elusive Wallace is not a chore to be taken on lightly!