Edward Tenner

The Rise of the Plagiosphere

How new tools to detect plagiarism could induce writer's block.

The 1960s gave us, among other mind-altering ideas, a revolutionary new metaphor for our physical and chemical surroundings: the biosphere. But an even more momentous change is coming. Emerging technologies are causing a shift in our mental ecology, one that will turn our culture into the plagiosphere, a closing frontier of ideas.

The Apollo missions' photographs of Earth as a blue sphere helped win millions of people to the environmentalist view of the planet as a fragile and interdependent whole. The Russian geoscientist Vladimir Vernadsky had coined the word *biosphere* as early as 1926, and the Yale University biologist G. Evelyn Hutchinson had expanded on the theme of Earth as a system maintaining its own equilibrium. But as the German environmental scholar Wolfgang Sachs observed, our imaging systems also helped create a vision of the planet's surface as an object of rationalized control and management—a corporate and unromantic conclusion to humanity's voyages of discovery.

What NASA did to our conception of the planet, Web-

based technologies are beginning to do to our understanding of our written thoughts. We look at our ideas with less wonder and with a greater sense that others have already noted what we're seeing for the first time. The plagiosphere is arising from three movements: Web indexing, text matching, and paraphrase detection.

The first of these movements began with the invention of programs called Web crawlers, or spiders. Since the mid-1990s, they have been perusing the now billions of pages of Web content; indexing every significant word found; and making it possible for Web users to retrieve without charge and in fractions of seconds, pages with desired words and phrases.

The spiders' reach makes searching more efficient than most of technology's wildest prophets imagined, but it can yield unwanted knowledge. The clever phrase a writer coins usually turns out to have been used for years worldwide— used in good faith because until recently the only way to investigate priority was in a few books of quotations. And in our accelerated age, even true uniqueness has been limited to 15 minutes. Bons mots that once could have enjoyed a half-life of a season can decay overnight into clichés.

Still, the major search engines have their limits. Alone, they can check a phrase, perhaps a sentence, but not an extended document. And at least in their free versions, they generally do not produce results from proprietary databases like LexisNexis, Factiva, ProQuest, and other paid-subscription sites or from free databases that dynamically generate pages only when a user submits a query. They also don't include most documents circulating as electronic manuscripts with no permanent Web address.

Enter text-comparison software. A small handful of entrepreneurs have developed programs that search the

open Web and proprietary databases, as well as e-books, for suspicious matches. One of the most popular of these is Turnitin; inspired by journalism scandals such as the *New York Times*'s Jayson Blair case, its creators offer a version aimed at newspaper editors. Teachers can submit student papers electronically for comparison with these databases, including the retained texts of previously submitted papers. Those passages that bear resemblance to each other are noted with color highlighting in a double-pane view.

Two years ago I heard a speech by a New Jersey electronic librarian who had become an antiplagiarism specialist and consultant. He observed that comparison programs were so thorough that they often flagged chance similarities between student papers and other documents. Consider, then, that Turnitin's spiders are adding 40 million pages from the public Web, plus 40,000 student papers, each day. Meanwhile Google plans to scan millions of library books—including many still under copyright—for its Print database. The number of coincidental parallelisms between the various things that people write is bound to rise steadily.

A third technology will add yet more capacity to find similarities in writing. Artificial-intelligence researchers at MIT and other universities are developing techniques for identifying nonverbatim similarity between documents to make possible the detection of nonverbatim plagiarism. While the investigators may have in mind only cases of brazen paraphrase, a program of this kind can multiply the number of parallel passages severalfold.

Some universities are encouraging students to precheck their papers and drafts against the emerging plagiosphere. Perhaps publications will soon routinely screen submissions. The problem here is that while such rigorous and robust policing will no doubt reduce cheating, it may also give

writers a sense of futility. The concept of the biosphere exposed our environmental fragility; the emergence of the plagiosphere perhaps represents our textual impasse. Copernicus may have deprived us of our centrality in the cosmos and Darwin of our uniqueness in the biosphere, but at least they left us the illusion of the originality of our words. Soon that, too, will be gone.