TO CATCH A

Tools and Tips to Combat Digital Content Plagiarism

The saying goes, “Imitation is the sincerest form of flattery.” Yet in the world of web content, imitation often goes beyond emulation of style or even subject matter. With cut-and-paste ease, imitation becomes plagiarism.

Jonathan Bailey, a journalist and writer, had been posting his original poetry and essays to literature websites when he got a wake-up call in his email inbox about five years ago. “A reader of mine tipped me off that my site was being plagiarized,” Bailey said. “It was a jaw-dropping moment for me.”

Since then, Bailey said he’s found over 500 other instances of his work being duplicated across the internet. “I’m fine with copying—all I asked for was attribution,” he said.

Print plagiarism used to be considered an occupational hazard for scholars and writers, like Bailey. With the advent of the internet, however, plagiarizing someone else’s original work requires less heavy lifting than ever. With a few keystrokes, content thieves can copy sentences and mirror entire sites, claiming false credit, taking ad revenue out of content creators’ pockets, or snagging search engine hits away from their rightful owners.

Luckily, the digital environment that makes plagiarizing content easier can make spotting poached pieces easier too. Digital tools are available to help authors discover when others are replicating their work without permission, and other developments can help teachers and managers make sure an author’s words are his or her own. For savvy users who want to protect their creations from falling into the wrong hands in the first place, there are some smart strategies to make it harder for human thieves and scheming bots to steal the credit for someone else’s original ideas. It’s about keeping others honest—and making sure you do the same.

A TOOL TO MATCH THE CRIME

First, the bad news: Digital tools can’t stop plagiarism before it happens. Original ideas get out—that’s the whole point of publication and distribution. And unless all proprietary content is locked away from public view, plagiarists will find a way to get their hands on it.

Until recently, a hunch was all teachers, editors, or other suspicious readers had to go on in order to catch the thief. They’d pour over the plagiarized work and try to match key phrases, paragraphs, and even pages to try to find what triggered the warning bells. The process could take weeks or months and involved thumbing through paper pages in endless archives.

The good news is that certain tools can tap the power of automated search and the depth of digital content databases to make the same standard process faster and more thorough. The basic idea underlying these tools isn’t all that different: The checker uploads the document in question, each word pattern is checked against the full text of original sources numbering in the dozens up to the billions, and then the system generates a report highlighting suspicious passages.

However, as Tom Holt, president and CEO of the search technology company Surf Wax, noted, “Each person is going to find that their definition of plagiarism is not identical [to anyone else’s]. The comparison engine has to accommodate some degree of defining what constitutes plagiarism, depending on the underlyng purpose of the organization using the tool.” Teachers might want to allow for the greatest amount of creative freedom. Lawyers might be looking for the smallest amount of leeway to protect the company’s reputation. And webmasters might need to check for duplicates of a site’s code. The following tools might share the underlying source-search report functions, but they all allow for different interpretations of what constitutes content theft.

THE ACADEMICS OF PLAGIARISM

Every student knows that turning in plagiarized work is worse than turning in no work at all—if you get caught. But getting caught isn’t certain. About 80% of college students
admitted to plagiarizing content at least once, according to The Center for Academic Integrity.

"The problem of plagiarism seems to take root in the academic field," said Alena Siameshka, head of the marketing department for SearchInform Technologies Inc., which released the plagiarism search tool PlagiatInform in June 2007. "Unlike a few decades back, today it’s next to impossible for teachers to read all the literature and publications on a particular discipline. The internet has granted access to rare sources, libraries, and journals, enabling students to get information from the sources their instructors might not be familiar with."

As a result, academic institutions from grade schools on up have been looking for ways to stem the tide by arming professors and teachers with digital tools. PlagiatInform allows professors to cross-check student papers against the university’s submission database and returns an estimate of what percent of the work has been lifted from another source. If the estimate is low, the paper is divided into paragraphs for comparison; if the estimate is high, the search highlights any suspicious text and links to possible sources line by line.

PlagiatInform’s design takes into account the massive source material from which cheaters draw. As papers are turned in, they’re added to the master database of student works, while bots automatically mine the internet and add relevant academic papers and sites. And as the database expands, the PlagiatInform software allows for storage and search capacity to increase accordingly without any manual upgrades or overhauls. And as more universities get on board with PlagiatInform, Siameshka said that those schools would be able to network with each other and search each other’s paper databases.

Another option is Turnitin, iParadigms’ anti-plagiarism tool and digital assessment suite, used by over 7,500 schools so far. It operates either as a standalone application or as an integrated part of the school’s content management and communications systems. When a paper is submitted by a student, it’s checked against a cache of more than 12 billion internet pages
collected and stored by Turnitin’s specially-designed spider, a database of 40 million student papers and licensed third-party content like newspapers, magazines, and books from educational information providers including Thomson Gale.

With that much content to cross-reference, similar sentence patterns are bound to turn up even on wholly original student work. To avoid career-ruining accusations, Turnitin offers Originality Reports and shows what percent of the flagged copy is suspicious, much like PlagiatInform’s percent-based results. “The report in no way states that a submission is or is not plagiarized,” pointed out Melissa Lipscomb, COO of iParadigms. “Rather, we create an unambiguous, objective report that can be used by our client to make the final decision on whether a paper has been cut and pasted.”

Still, some institutions feel that these tools undermine the trust it shows in its students, and several of the most prestigious universities in the country—including Harvard, Yale, and Princeton—have no institution-wide tool to catch plagiarists. If a professor has a hunch, he or she just has to follow it up in the “old-fashioned” way: through Google or another search engine.

**THE BUSINESSES OF PLAGIARISM**

Employee plagiarism can tarnish the integrity of any company, whether it’s a front-page story in The New York Times or an ad slogan lifted by a lazy copywriter. On the other end, companies whose protected content is being ripped off lose out on profit from that content.

“The financial liability is tremendous to the infringing business. Regarding businesses whose content is stolen, the impact is obvious: It de-values their intellectual property,” said Lipscomb.

The corporate counterpart to iParadigms’ Turnitin plagiarism-detection tool is iThenticate. The web-based application scans an extensive page cache and content database and, if any pattern matches are detected, the content is flagged and shown alongside the source.

---

**Dealing with Digital Plagiarism**

There’s no detective unit devoted to uncovering crimes of plagiarism. Even if you’ve got the right tools to catch copying when it happens, here are some tips from digital content experts to make your work less vulnerable to misuse:

**Post warning signs.** Let readers know you’re paying attention to how your content is being used and re-used. “It’s still better to prevent your content from being stolen in the first place,” said Copyscape’s Gideon Greenspan. “A simple yet effective measure is to warn potential content thieves that they will be discovered. A range of warning banners is available for free at Copyscape.com. These banners are already included in over 20,000 different websites.” If a banner isn’t your style, a simple notice that the content is protected posted in a prominent place can get the message across.

For schools and enterprises, the knowledge that deeply sourced anti-plagiarism tools are in place to verify content’s authenticity can discourage authors from cutting corners. “At the end of the day, what’s important is the understanding that there’s a chance that documents you have been using are in the source file,” said Holt. “As an employee or student, you’d be more inclined to work carefully.”

**Copyright it.** Affixing copyrights all over the internet bogs down the flow of information. But, for digital authors, the ability to protect original work when necessary is important.

In the United States, most digital creators are assumed to have certain copyright protections over their creations. But, said Bailey, registering with U.S. copyright officials can give authors a wider range of control over their work, including the right to sue for punitive damages if the work is being egregiously misused.

Another option is to establish a Creative Commons license for the content. The more flexible model helps authors strike a suitable balance between reader freedom and user protection. FeedBurner’s RSS tool allows authors to publicize and attach Creative Commons protections to content on its way out to subscribers.

**See what’s out there.** Even if you’re using a tool like CopyGuard or Copyscape to check in on your content, it doesn’t hurt to venture beyond their limitations and make sure plagiarists aren’t slipping through the net.

“Listen to your readers, and do searches yourself,” suggested Bailey. “Even if you don’t find anything, at least you know. Unfortunately, the odds of you finding nothing these days are pretty slim.”
In 2005, LexisNexis partnered with iThenticate to create CopyGuard. Using the same basic match-and-report system of iThenticate, CopyGuard is available to LexisNexis subscribers. It broadens search comprehensiveness by combining over 6 billion LexisNexis digital documents with iThenticate's web page archive.

As many businesses and publications launch websites or digital publications, they face the possibility that their copyrighted content will wind up on someone else's site. "Any website with good content or marketing copy is likely to be copied—the problem is that it's just so easy for someone to copy the text from your site and make a few minor modifications to suit their purposes," said Gideon Greenspan, co-founder of Indigo Stream Technologies Ltd., which launched the anti-plagiarism web tool Copycape in 2004.

Copycape is a search engine that scans the content of a specific URL and runs a pattern-matching algorithm to find potential plagiarized copies across the rest of the visible web. Suspicious sites are returned with the questionable content highlighted.

Anti-plagiarism tools designed for the enterprise sector such as iThenticate and Copycape are designed to protect the company both from potential liability and potential loss. They rely heavily on access to a rich source pool, both from across the web and from the protected digital databases of companies like LexisNexis. "The real issue is what you have access to," said Tom Holt. The more access, the less likely corporate plagiarists will get away with their theft.

TAMING THE WILD WILD WEB

After Jonathan Bailey was beleaguered by digital plagiarists, he fought back by starting a blog about online content theft. Besides blogging about the subject, Bailey still finds himself dealing with those bold enough to plagiarize PlagiarismToday.com.

According to Bailey, a threat most online authors don't see coming is automated RSS scrapers, which use syndication channels to dump content minus attribution on a spam blog, or “splog.” Using keyword-rich content, the splogs set up a contextual ad scheme and divert search engine traffic to pick up the ad hits from the original author's site. And these days, Bailey says he's seen RSS scrapers mining directly from Google or Technorati feeds the author set up himself.

"In the world of RSS feeds, it seems to me that most people don't think of re-using content as plagiarism," said Holt. The rush to set sites up for syndication may inadvertently put the content at risk of being ripped off by some bad bots.
That’s not to say that RSS is a harmful tool. The key, said Bailey, is to keep an eye on who’s subscribing to what. To that end, Rick Klau, vice president of publisher services at FeedBurner, which creates RSS-management tools, said that FeedBurner developed its Uncommon Uses tool to help content creators analyze traffic patterns, identify suspicious use, and see where their feed winds up.

“Uncommon Uses helps to identify re-syndication of your feed beyond standard consumption points, including contact with your feed by non-subscribers,” said Klau. “I think the combination of blog plugins, services like Uncommon Uses, and traffic analysis can go a long way to minimizing the potential harm.”

There’s no widely used tool to deal with the explosion of text-free intellectual property like video, graphics, and pictures yet, because pixel matching is still hard to implement on a large scale. However, a unique file name or tag can make duplicate copies more findable. Bailey recommended overlaying images with a subtle watermark and keeping an eye on server logs to see who’s linking to images and videos directly.

The watermark principle is already being employed as protection on some self-publishing platforms, usually as a plugin or optional tool. For instance, blog host WordPress allows users to tag their content with an invisible digital watermark. The service, called Digital Fingerprinting, then monitors major information feeds and search engines to see if the watermark turns up anywhere other than your site.

“A lot of bloggers that know there’s a problem don’t know how this can impact them whether there’s anything they can do about it,” said Bailey. By plugging unique phrases or tags into a standard web search engine like Google every week or so, or setting up a search alert on those key phrases, internet authors can keep control over their content.

If plagiarism is discovered, Bailey said, bloggers don’t need the legal resources of a corporation to regain that control. He recommended sending a cease-and-desist letter to the plagiarist or notifying advertising companies such as Google or Yahoo!, whose services are supporting the sploggers—the plagiarist is probably in violation of the company’s usage policy. And, if all else fails, a DMCA violation notice usually gets people’s attention.

Although content digitization has made it easier to point, click, and grab, tools are emerging—making it easier for authors, teachers, managers, and any other online publishers to nab the grabbers—to help balance out the scales of just content use. 

JESSICA DYE (JESSICA.DYE@GMAIL.COM) IS AN ILLINOIS-BASED FREELANCE WRITER. COMMENTS? EMAIL LETTERS TO THE EDITOR TO ECLETTERS@INFOTODAY.COM.