SPECIAL REPORT

Taking on the cheats

The true extent of plagiarism is unknown, but rising cases of suspect submissions are forcing editors to take action. Jim Giles reports.

A. MACDONALD

The fight against plagiarism is about to take a decisive turn. Academic publishers have told *Nature* they hope that software designed to catch cheating students could soon be used to unmask academics who plagiarize other researchers' — or their own — work.

Big publishers such as Elsevier and Blackwell, which between them publish more than 2,500 journals, have been prompted to act by reports that plagiarism is becoming more common. "We're hearing about it more frequently from editors," says Bob Campbell, president of Blackwell Publishing in Oxford, UK.

Self-plagiarism, in which authors attempt to pass off already published material as new, is a particular problem. In an increasingly competitive environment where appointments, promotions and grant applications are strongly influenced by publication record, researchers are under intense pressure to publish, and a growing minority are seeking to bump up their CVs through dishonest means.

The extent of the problem is hard to assess. Defining plagiarism is not straightforward (see 'Where to draw the line?', below), and measuring the incidence of even the most clear-cut cases is difficult. Studies in certain fields have estimated that anything up to 20% of published papers contain some degree of self-plagiarism (see 'How common is plagiarism?', opposite). This may not be representative of basic research, but no rigorous, multidisciplinary study has ever been conducted.

And although most cases are never discovered, almost all of the editors and publishers contacted by *Nature* agreed that self-plagiarism is on the rise. "Editors are noticing many more cases," says Scott Dineen, director of editorial services at the Optical Society of America, which publishes ten journals. Last month, the increase prompted the society to issue an editorial statement on its commitment to expose plagiarism[1].

The advent of antiplagiarism software, such as that used by universities to check student essays, means that editors and publishers finally have a practical way to tackle the problem. Online services check essays against massive stores of documents generated from web trawls and purchases from media outlets. Supervisors can see which parts of the essays seem to be plagiarized and where the copied material comes from.

Adapting such technology for use with academic papers should be easy, say publishing experts, as the software could be bolted onto the online systems that publishers use to manage peer review. The system would work in the background and editors would only be aware of it when it flagged up a suspicious degree of overlap, which they could then check out in detail.

"We see it as an idea with great potential," says Campbell. "It will eventually be part of the editorial office."

ArXiv, the popular physics preprint server based at Cornell University in Ithaca, New

Where to draw the line?

Determining what constitutes plagiarism is tricky. Few scientific organizations have quantified the fraction of material that can be legitimately reused between papers, and few researchers have heard of those rules that do exist. So although plagiarism is relatively simple to define in a qualitative sense, it can be extremely difficult to rule on in practice.

Statements on plagiarism usually define the act as attempting to pass off someone else's work as your own. Duplicate publication, or self-plagiarism, occurs when an author reuses substantial parts of their own published work without providing the appropriate references. This can range from getting an identical paper published in multiple journals, to 'salami-slicing', where authors add small amounts of new data to a previous paper.

When large chunks of text have been cut-and-pasted, such definitions work well. But researchers routinely commit minor plagiarism without dishonest intent, such as reusing parts of an introduction from an earlier paper. To help editors resolve these cases, some journals set an upper limit for the amount of text that can reused, usually about 30%.

But what should an editor do when a paper looks similar to one already published, but does not contain chunks of text that have obviously been copied? This technique, dubbed 'intelligent plagiarism', is likely to evade detection tools that simply compare strings of text, although including sets of data in the comparison may help to flag up suspect cases.

Generally, the problem can only be resolved by editors studying the two papers, talking to the authors and making a personal decision on whether misconduct has occurred. "Plagiarism is always a human judgement," says Fintan Culwin, a plagiarism-software expert at London South Bank University.

Such cases can also be extremely time-consuming to investigate; one reason why data from the Committee on Publication Ethics, a UK-based group of biomedical journal editors, suggest that many allegations remain unresolved a year after they were first made. **J.G.**

CLIFF QUIVERS WARN OF COLLAPSE
Seaside cliffs may shake before they fall, giving hours of notice.
www.nature.com/news

York, is almost ready to deploy plagiarism detection software. Paul Ginsparg, the Cornell physicist who runs arXiv, acted after 22 plagiarized papers were discovered on the archive[2].

Physical exercise

Daria Sorokina, a computer science PhD student at Cornell, has tuned an established algorithm to look for any two documents that share at least six of the same words in a row. The system is already finding "plenty of awkward things", says Ginsparg. One test-run revealed a PhD thesis that shares large chunks of material with a paper posted to the archive three years earlier. So far, Ginsparg says, the tests have revealed a few thousand pairs of articles by different authors that have "excessive overlap".

Ginsparg plans to post all of the pairs on the arXiv website — without accusing the authors of wrongdoing — and ask the researchers involved to respond. He hopes that the results will help to refine the algorithm, which could then be used on new submissions to generate a warning if papers seem to overlap.

The world's largest scientific publisher — Amsterdam-based Elsevier, which publishes about a quarter of a million papers each year — has also decided to act. Last month, it initiated a year-long assessment of the various technology options.

Tools are now becoming available for individual editors and peer reviewers. One has been developed by Christian Collberg, a computer scientist at the University of Arizona, Tucson. He was asked to review a paper for a conference in 2003, and recalls carrying out a Google search to research it. "I found an earlier published version that had just been reformatted," he says. Reviewing for the same conference the following year, Collberg found another submission that had copied substantial parts of the author's own earlier work. "This really pissed me off," he says. "I spent time reviewing those papers."

Copycats

In response, Collberg began work on what has become the Self-Plagiarism Detection Tool (SPlaT). Whereas publisher-run plagiarism detection services are likely to take years to set up, Collberg's software is available, free-to-use, and targeted at editors and peer reviewers.

The software grabs papers from authors' websites and compares them to each other and to other manuscripts added manually, such as papers under review. Collberg declined to reveal details of what happened when he let SPlaT loose on the websites of 50 computer-science departments, but summary results released last month show that the software turned up more than one pair of conference publications with more than 50% common text and no reference to each other[3].

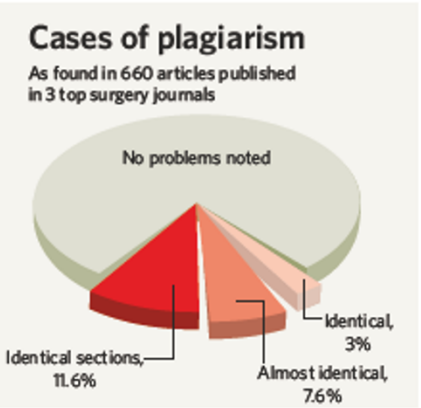
Student antiplagiarism services are another option. These offer the possibility of detecting plagiarism itself, not just duplicate publication. The firm behind the iThenticate software, for example, says its product is licensed by 5,000 institutions and that it can check documents against a database of more than seven billion pages for cases of suspicious overlap. Because this store has been generated in part from web trawls, it contains papers from authors' websites and some open-access journals. If a university subscribes to the service to tackle student fraud, researchers in those institutions can also use it when refereeing papers.

When *Nature* gave the software a test-drive, it did a good job of identifying areas where the same text had been legitimately used in different places, such as on authors' websites and in properly referenced quotes. But when a known plagiarism was submitted to iThenticate, it failed to turn up the original, even though it had been published in a high-impact journal.

The reason, says John Barrie, president of iParadigms, the California firm that developed the software, is that most online literature is locked behind subscription barriers and so cannot be added to the iThenticate database.

An overall solution to plagiarism, says Campbell, is likely to come when publishers collaborate on industry-wide detection systems. Preliminary discussions about this have already taken place at meetings of CrossRef. This company, based in Lynnfield, Massachusetts, collaborates with publishers to develop systems that allow researchers to search across journals from many different companies.

Such a system could, in theory, catch almost all instances of direct plagiarism, although it will take several years to set up, even if negotiations go smoothly. In the meantime, say editors, plagiarized papers will continue to creep into the literature as a minority of researchers dishonestly beef up their reference lists. After all, says one researcher who admitted to *Nature* that he occasionally neglects to mention papers that overlap with new publications, there is a perception that "the Dean can't read, but he can count". ■



How common is plagiarism?

A literature search on the word 'plagiarism' reveals numerous editorials bemoaning the problem, but little in the way of hard facts. The data are patchy, often anecdotal and rarely applicable to more than one field. To make matters worse, it's impossible to know how many cases evade detection.

One approach to gauging the frequency of plagiarism and other forms of misconduct is to search for notices of retraction. The results of one such trawl, which used biomedical literature in the PubMed database, were published in January and put the incidence of "recognizable fraudulent material" at less than 0.02% of all papers[4].

But surveys that compare individual papers report higher figures. Most of these focus on duplicate publication of clinical papers. Last year, for example, a trawl of 1,234 articles on anaesthesia and analgesia found that 5% were duplicates that did not reference the appropriate original[5].

A 2001 study of surgical journals put the figure even higher: nearly a quarter of articles published that year had some form of redundancy, and 11% were suspected to be dual publications (see chart). "Redundant publications must be recognized as a real threat to the quality and intellectual impact of surgical publishing," the authors were moved to conclude[6]. Duplicates have also been shown to cause meta-analyses to overestimate the efficacy of drugs[7].

The extent of the problem for most basic research is probably somewhere between these two extremes. Rigorous studies have not been performed on basic-research papers, so the problem may be going undetected. But commercial pressures may also encourage the duplication of papers that report positive findings about a new drug. In basic research, where the link between data and profits is less direct, the problem may be less common. **J.G.**

1. *Opt. Lett.* **30**, 813 (2005).
2. *Nature* **426**, 7 (2003).
3. Collberg, C. & Kobourov, S. *Commun. ACM* **48**, 88–94 (2005).
4. Claxton, L. D. *Mutat. Res.* **589**, 17–30 (2005).
5. von Elm, E., Poglia, G., Walder, B. & Tramèr, M. R. *J. Am. Med. Assoc.* **291**, 974–980 (2004).
6. Schein, M. & Paladugu, R. *Surgery* **129**, 655–661 (2001).
7. Tramèr, M. R., Reynolds, D. J. M., Moore, R. A. & McQuay, H. J. *Br. Med. J.* **315**, 635–640 (1997).