Résumé

La littérature scientifique est affligée par des publications de dupliqués. La lutte contre la plagiarism est sur le point de prendre un tournant crucial avec l’arrivée d’un large éventail de logiciels de détection de plagiat. Le plus puissant d’entre eux est le laboratoire d’innovation de Virginie ‘Déjà vu’. Ce système de base de données de citations dupliquées conçu par Garner utilise l’eTBLAST, un moteur de recherche de similarité de texte.

Les journaux de nature ont publié des rapports ces dernières années où de nombreuses citations dupliquées ont été détectées et déposées dans la base de données Déjà vu et les rédacteurs ont commencé à retraiter des articles. La dualité de l’eTBLAST libre d’accès et de la base de données Déjà vu se comporte comme un ombudsman éthique et peut très bien être un frein contre les pratiques éthiques.

Mots clés

Publications de dupliqués, plagiat, eTBLAST, Déjà vu, Bioéthique

Introduction

Le célèbre physiologiste Walter Cannon une fois dit que la recherche est l’acquisition systématique de nouvelles connaissances “qui profondément satisfait l’esprit et sa curiosité de l’explorateur.” Dans un monde concurrentiel d’aujourd’hui, l’‘publish-or-perish’ culture a dilué cet esprit aventureux. La littérature scientifique est confrontée à la montée des publications de dupliqués. Il faut une méthode à cette folie que nous appelons recherche.

La menace des publications de dupliqués

Les publications de dupliqués ont été un défi pour les journaux médicaux pendant de nombreuses années et ils souvent sur le programme des réunions des rédacteurs. Dans une enquête, 3,247 chercheurs américains ont avoué un comportement non éthique dans la forme de répétition des publications (4,7%) et du plagiat (1,4%). Nous avons besoin de références croisées parfois mais la duplication mène à une inflation des résultats. Par exemple, dans un méta-analyse du BMJ, l’efficacité antiémétique d’Ondansetron est surévaluée de 23% en raison des duplications. 17% des essais randomisés et 28% des données patientes dans l’étude ont été significativement dupliquées.

Le comité international des rédacteurs des journaux médicaux (ICMJE) décrit la publication de dupliqués (ou répétitifs) comme :

“…publication of a paper that overlaps substantially with one already published in print or electronic media.”

A surgé avec la montée de nouveaux journaux dans leandalone et on perdtrait le niveaux de qualité de la littérature scientifique si des outils étaient développés pour détecter des duplications aussi tôt que possible dans le processus de production.

Il existe un nombre d’outils de détection de plagiat disponibles pour identifier les publications redondantes. L’un d’entre eux a été identifié par le comité de publication de l’ethique (COPE) est CrossCheck, un service de détection de plagiat. Il est offert par le service CrossRef.

Le CrossCheck communauté d’utilisateurs de plus de 50 éditeurs se compose de scientifiques, médicaux et techniques éditeurs et sociétés utilisant iThenticate technologie in
the editorial process. iThenticate is a wholly owned subsidiary of iParadigms, LLC, a web based service for collaborative, online educational support. The fight against duplicate citations is about to take a crucial turn. Recently, Dr Harold Garner’s innovation of flagging duplicate publication by identifying highly similar citations (based on their abstract) from MEDLINE is making waves. Garner’s lab uses a text similarity-based information retrieval and search engine named Déjà vu.

eTBLAST uses a text similarity-based engine to search literature collections, where MEDLINE, NASA, IOP (Institute of Physics), PMC (PubMed Central), Arxiv, Clinical trials and CRISP (Computer Retrieval of Information on Scientific Projects, a database of federally funded biomedical research) are available. Unlike other search engines it does not utilize Boolean operators but provides a simple interface to scan the whole of word-by-word text. eTBLAST performs at its optimum when it uses large number of words. In this regard PubMed Central’s growing list of full text articles makes it possible to identify the frequency of duplication of portions of submitted manuscripts. Copying of articles is easily identified and the duplicate details are available in the public domain via Déjà vu. This is freely accessible and users are encouraged to contact the researchers about the authenticity of these suspected duplicates. Garner has started contacting the editors and authors of the duplicates Déjà vu has identified, and is submitting the results for publication.

Garner’s team uses eTBLAST to build Déjà vu, a continually updated database that holds over one hundred thousand abstracts listed in Medline that seem highly similar. This watch dog has so far found dozens of near-100% clone papers. The researchers have put these numerous suspected duplicates in the public domain via Déjà vu. This is freely accessible and users are encouraged to contact the researchers about the authenticity of these suspected duplicates. Garner has started contacting the editors and authors of the duplicates Déjà vu has identified, and is submitting the results for publication.

Key features

The key feature of eTBLAST is a ‘Biomedical Acronym Resolver’ and more importantly a ‘Pair Comparison.’ In the latter, two different paragraphs from different sources are evaluated for similarity as part of biomedical text comparison. It was this unique feature which enabled it to catch >87.5% text similarity between a Nepalese article and an American article. The editors of the Nepalese journal investigated and penalized the author after eTBLAST’s email.

The Déjà vu interface is designed using python (http://python.org) and the Django web framework (http://djangoproject.com). Data are stored in a backend MySQL Database (http://mysql.com) within the Garnerian innovation lab. The data entries are retrieved using PubMed ID, first article and the last article of similarity, the publication lag between these two, languages of both these articles and their ‘Similarity Ratio’ calculated by dividing the “Duplicate Score” by the “Identity score.” The database also indicates whether or not the duplications have shared authors.

The creators of this tool, which functions as a sort of ethical ombudsman, observed that duplications were predominantly in journals with low impact factor and that these articles were rarely cited. Escaping detection may be more likely because of low visibility of the journal. A further increased tendency was noticed in which reviews based on a previous publication duplicated matter from the first publication. This was easily picked up by the simple interface of Déjà vu.

In an effort to further enhance the sensitivity, Garner’s team has used ‘statistically improbable phrases’ (SIPs) for assistance in identifying duplicate content. The new innovation yields a much better precision of 78.9% in comparison to 50.3% for eTBLAST.

Recent events

Nature reported that an immunologist’s review article is to be retracted from an Iranian journal following allegations of duplicate publication. eTBLAST found that many paragraphs were lifted from Farsi-language forums and blogs in Iran. Déjà vu claimed about 85% similarity to five different papers by other writers. The author of one of the original paper quotes "The article is a veritable patchwork of other people’s work, word for word, grammatical error for grammatical error." In defence, the tainted Immunologist blamed it on her student and did not respond to e-mail queries from Nature.

Déjà vu identified that French gerontologist’s entire paper had been plagiarised in Korean Journal of Biological Sciences. It was reported to the editors of Experimental Gerontology who tried to investigate, but without any success. A problem confronting those working to identify plagiarism is that many journal editors seem reluctant to pursue the cases of alleged plagiarism. A previous study from Garner’s lab searched MEDLINE abstract from the previous 12 years with eTBLAST and found over one hundred thousand duplicate citations with the same authors. The false positive rate was only 1% in this study and the duplicate entries have been deposited in the Déjà vu repository.
Around three quarters of publications in MEDLINE come from USA, UK, Canada, Italy, Germany, France, China and Japan. Most contributions in the Déjà vu database are from China and Japan and the trend predicts more duplicated citations from non-English-speaking countries. Various reasons have been suggested to explain this and chief among them is complexity of translation, cultural norms and different ethical training. A word of caution

An editorial in Clinical Chemistry, examined the false-positive rates in the Déjà vu database. It checked the suspected duplicates in 3 journals, The New England Journal of Medicine (NEJM), Clinical Chemistry, and The Lancet, since 1975 and found misclassification of the reports. The authors find the reasons to be: articles published in different languages, two-part articles, follow up from same cohort, elaborated abstracts from a conference, Medline/publisher error and guidelines adopted and published by cooperating journals. To take a simple example, PubMed shows articles with same title and authors when you search “Toward more uniform conflict disclosures: the updated ICMJE Conflict of Interest Reporting Form.” These may be sensed as duplicate citation by an electronic database but the fact is that these are the guidelines which are adopted and published by several cooperating journals in the larger interest of the public. Hence it is advisable to do a manual verification first before claiming scientific misconduct.

Conclusions

The combination of the freely available eTBLAST tool and Déjà vu database can be a deterrent to unethical practices and is a positive step forward in making such detection easier for authors, editors and reviewers. Journals should make use of this freely available software to ensure high ethical standards.

References

3. Tramèr MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ. 1997; 315 : 635
5. White C. Plagiarism detection service to be launched in June. BMJ 2008; 336 : 797
6. www.ithenticate.com

PEER REVIEW
Not commissioned. Externally peer reviewed.

CONFLICTS OF INTEREST
The author declares that he has no competing interests.