Déjà vu: The duplicate citation database as an ethical ombudsman

Satendra Singh

Department of Physiology
University College of Medical Sciences
Delhi, India

REVIEW

Please cite this paper as: Singh, S. Déjà vu: The duplicate citation database as an ethical ombudsman. AMJ 2010, 3, 12, 840-842
Doi http//dx.doi.org/10.4066/AMJ.2010.516

Corresponding Author:
Satendra Singh
Assistant Professor of Physiology
University College of Medical Sciences
Dilshad Garden, Delhi-110095, India
dr.satendra@gmail.com

Abstract

Scientific literature is plagued by duplicate publications. The fight against plagiarism is about to take a crucial turn with the advent of a plethora of plagiarism detection software programmes. The one which is making biggest waves is the Virginia Innovation laboratory's 'Déjà vu'. This duplicate citation database devised by Garner utilizes eTBLAST, a text similarity-based search engine. *Nature* has published reports in the last few years where many duplicate citations have been detected, deposited in Déjà vu databases and editors have started retracting articles. The dual combination of freely available eTBLAST tool and Déjà vu database act as an ethical ombudsman and can very well be a deterrent against unethical practices.

Key Words
Duplicate publications, Plagiarism, eTBLAST, Déjà vu, Bioethics

Introduction

The famous physiologist Walter Cannon once said that research is systematic acquisition of new knowledge "which deeply satisfies both the explorer's adventurous spirit and his persistent curiosity." In today's competitive world the 'publish-or-perish' culture has diluted this adventurous spirit. Scientific literature is challenged with the rise of duplicate publications. There has to be a method in this madness which we call research.

Menace of duplicate publication

Duplicate publication has been a challenge for medical journals for numerous years and it often tops the agenda of Editors' meetings.[1] In a survey 3,247 American researchers admitted unethical behaviour in the form of recurring publication (4.7%) and plagiarism (1.4%).[2]

We indeed need cross references at times but duplication leads to over inflation of results. For example In a BMJ meta-analysis, the antiemetic efficacy of Ondansetron is over-estimated by 23% because of duplications.[3] The researchers also found 17% of randomised trials and 28% of the patient data in the study to be significantly duplicated.

The International Committee of Medical Journal Editors (ICMJE) describes duplicate (or redundant) publication as:

"…publication of a paper that overlaps substantially with one already published in print or electronic media."[4]

A surge in the rise of new journals is correlated with lax standards and it would help preserve the quality of scientific literature if tools were developed to detect duplication as early as possible in the production process.

There are a number of software programs available to identify redundant publications. One of note identified by the Committee on Publication Ethics (COPE) is CrossCheck, a plagiarism detection service. It is offered by the independent publishers' membership association CrossRef.[5]

The CrossCheck user community of over 50 publishers consists of global scientific, medical and technical publishers and societies using iThenticate technology in the editorial process. iThenticate is a wholly owned subsidiary of iParadigms, LLC, a web based service for collaborative, online educational support.[6]

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 eTBLAST.[7]

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 a freely accessible database named Déjà vu.[8]

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 watchdog has so far found dozens of near-100% clone papers.[9] 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.[10]

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.[11] 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.[12] 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. [13]

Recent events

Nature reported that an immunologist's review article is to be retracted from an Iranian journal following allegations of duplicate publication.[14] 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." [14] 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* .[15]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.[16] 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.[9]

A word of caution

An editorial in *Clinical Chemistry*, examined the false-positive rates in the Déjà vu database.[17] 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.[18, 19] Journals should make use of this freely available software to ensure high ethical standards.

References

1. DeMaria AN. Duplicate publication: insights into the essence of a medical journal. J Am Coll Cardiol, 2003; 41:516-517.
2. Martinson, B. C., Anderson, M. S. & de Vries, R. Scientists behaving badly. Nature. 2005; 435: 737–738.
3. Tramèr MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ. 1997; *315 : 635*
4. http://www.icmje.org/publishing_4overlap.html (accessed October 19, 2010)
5. White C. Plagiarism detection service to be launched in June. BMJ *2008; 336 : 797*
6. www.ithenticate.com
7. http://etest.vbi.vt.edu/etblast3/ (accessed October 19, 2010)
8. http://dejavu.vbi.vt.edu/dejavu/ (accessed October 19, 2010)
9. Errami M, Garner H. A tale of two citations. *Nature. 2008;* 451: 397-399
10. Dixit H. Consultant Editor's Note regarding publication of a duplicate article: "Correlation between serum-ascites albumin concentration gradient and endoscopic parameters of portal hypertension (PMID: 16449830; Oct-Dec 2005)". Kathmandu Univ Med J (KUMJ). 2008;6(23):301.
11. Errami M, Sun Z, Long TC, George AC, Garner HR. Deja vu: a database of highly similar citations in the scientific literature. Nucleic Acids Res. 2009;37:D921–D924
12. Errami M, Hicks JM, Fisher W,et al. Deja vu–a study of duplicate citations in Medline. Bioinformatics. 2008;24:243–249
13. Errami M, Sun Z, George AC, et al. Identifying duplicate content using statistically improbable phrases. Bioinformatics. 2010;26(11):1453-7.
14. Butler D. Iranian paper sparks sense of déjà vu. Nature. 2008; 455: 1019.
15. Butler D. Entire-paper plagiarism caught by software. Nature. 2008; 455: 715.
16. Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res. 2007;35: W12–W15.
17. Rifai N, Bossuyt PM, Bruns DE. Identifying duplicate publications: primum non nocere. Clin Chem. 2008;54(5):777-8
18. Giles J. Special report: taking on the cheats. Nature 2005;435:258-259.
19. Marris E. Should journals police scientific fraud? Nature 2006;439:520-521.

PEER REVIEW

Not commissioned. Externally peer reviewed.

CONFLICTS OF INTEREST

The author declares that he has no competing interests.