Nucleic Acids Research Advance Access published September 8, 2008

Nucleic Acids Research, 2008, 1–4
doi:10.1093/nar/gkn546

Déjà vu: a database of highly similar citations in the scientific literature

Mounir Errami[1,*], Zhaohui Sun[1], Tara C. Long[2], Angela C. George[2] and Harold R. Garner[1,2]

[1]Division of Translational Research and [2]McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9185, USA

Received August 7, 2008; Accepted August 9, 2008

ABSTRACT

In the scientific research community, plagiarism and covert multiple publications of the same data are considered unacceptable because they undermine the public confidence in the scientific integrity. Yet, little has been done to help authors and editors to identify highly similar citations, which sometimes may represent cases of unethical duplication. For this reason, we have made available Déjà vu, a publicly available database of highly similar Medline citations identified by the text similarity search engine eTBLAST. Following manual verification, highly similar citation pairs are classified into various categories ranging from duplicates with different authors to sanctioned duplicates. Déjà vu records also contain user-provided commentary and supporting information to substantiate each document's categorization. Déjà vu and eTBLAST are available to authors, editors, reviewers, ethicists and sociologists to study, intercept, annotate and deter questionable publication practices. These tools are part of a sustained effort to enhance the quality of Medline as 'the' biomedical corpus. The Déjà vu database is freely accessible at http://spore.swmed.edu/dejavu. The tool eTBLAST is also freely available at http://etblast.org.

INTRODUCTION

Authorship of scientific papers is one of the most valuable currencies for scientists and engineers, and is an asset not only for climbing the corporate or academic ladder (1), but also most importantly to secure funding for academic laboratories. The fierce competition in most scientific disciplines and the increasing necessity to publish may lead authors to engage in questionable behavior such as publishing a single piece of work more than once, or emulating the style, or copying the content of another person's work. Duplicate publication may be useful to provide wider access to the scientific community or to report important updates to surveys or clinical trials, but publications that simply reproduce a previous work with virtually identical results and conclusions often lack the novelty to justify additional publication. The latter types of duplicate publication are considered unethical because they undermine the public confidence in scientific integrity. Others have previously described additional duplicate publication behaviors referred to as 'salami slicing' (dissecting a scientific work into multiple least publishable units) and 'meat extenders' (building on a previous publication with new data that would not be publishable alone) (2–4). Most previous studies of duplicate publication have been limited to a particular scientific field where duplication was painstakingly identified manually, underscoring the need for an automated method to detect putative duplications (5–16).

We have established a method to identify highly similar citations in Medline, the comprehensive literature database of life sciences and biomedical information, using the text similarity search engine eTBLAST (17,18). We were able to statistically calibrate eTBLAST to identify citations that have unusually high similarity, which were then saved in Déjà vu pending manual inspection (19,20).

CONTENT AND METHODS

Identification of highly similar citations

Technical details describing the detection of highly similar citations and its application to the entire Medline database have been reported previously (19,20). Briefly, the method which has contributed the preponderance of entries in Déjà vu involves 'eTBLASTing' each Medline citation against its most related article (a feature available from Medline). Upon comparison, citation pairs are so highly similar that predetermined similarity thresholds exceeded are flagged as a highly similar pair and stored in Déjà vu awaiting manual verification by human curators.

*To whom correspondence should be addressed. Tel: +1 214 648 5992; Fax: +1 214 648 1445; Email: mounir.errami@utsouthwestern.edu

Table 1. Déjà vu content by category and category definitions

Duplication type	Count	Description
DISTINCT	1379	There are a number of reasons for different citations to have a high similarity, including citations that describe related, but very distinct publications. A pair of citations identified by computer similarity, which after inspection is, for example, clearly a continuation of a study which has evolved, and the text represents new information that is categorized as a distinct and unique work
DUPLICATE	2443	A pair of citations that was identical or nearly identical. The citations report on a study with the same or very similar results and conclusions.
ERRATUM	188	Only a fraction of the MEDLINE records that are apparently corrections to previous entries are marked as errata. If a title/abstract pair is either labeled as errata or if it is clear that a correction has been made (author list, spelling, small changes to abstract or title wording, etc.), then the errata classification is used.
SANCTIONED	1619	There are a number of reasons for different citations to have a high level of similarity, some of which play a special, very important, and very legitimate role in the reporting of science. Examples include periodic reviews, periodic guidelines, specialized databases and specialized federal register citations. Citation pairs of this type, identified through computer text similarity have been manually classified to the category sanctioned.
NO ABSTRACT	16	In some cases highly similar titles are flagged as potential duplicates, but the non-identity MEDLINE record does not contain an abstract, we designate that pair as a 'NO ABSTRACT' to indicate that its status cannot be determined.
UNVERIFIED	69115	Deja vu is a database of duplicate publications, as identified using a number of different techniques, with the principle one being text similarity comparisons. Those putative duplicates identified by any of these techniques, prior to human verification and assignment to another category, are initially loaded into these categories, and since our software also inspects the author lists, they are loaded into unverified categories that have either overlapping authors (SA) or not (DA).
TOTAL	74 760	

Up to date statistics and definitions are available at http://spore.swmed.edu/dejavu/help and http://spore.swmed.edu/dejavu/statistics/.

Manual classification of highly similar citations

Déjà vu was designed and developed to allow for collaborative work among the multiple curators. It was also necessary to define a broad, flexible and extensible classification scheme to accommodate a wide range of highly similar documents dealing with all areas of biomedical research, reflecting different publication behaviors, styles and agreements. Upon manual verification, highly similar citation pairs were classified in one or more of the categories listed and defined in Table 1. In particular, we sought to distinguish between appropriate and inappropriate duplication, a process which is admittedly subjective. A pair of duplicates with different authors may indicate potential plagiarism, while two publications with shared authors may indicate multiple publication of the same study. Updates to clinical trials or survey type research are instances where complete duplication is not necessarily inappropriate. Similarly, studies with different outcomes using similar phraseology may bring valuable new information. Errata, which may or may not be tagged as such in Medline, are most similar to the initial record, often involving only a typographical correction. All of these determinations are difficult or impossible to accomplish computationally, and thus are best made by human curators.

Déjà vu in numbers

All data collected have been consolidated into a web-accessible database, available at http://spore.swmed.edu/dejavu. As of 22 July 2008, Déjà vu contains a total 74 760 records of which 5645 have been manually inspected (Table 1). Déjà vu has received over 40 000 visits since 1 January 2008 and currently receives an average of about 2000 visits per month.

QUERIES AND INTERFACE

The Déjà vu interface was 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). Déjà vu was designed to allow real-time collaborative annotation by multiple curators who need not be programmers to add comments and updates or create new records.

On the Déjà vu website users can: (i) browse Déjà vu entries with no specific search method (Each entry links to the scientific citation along with full text when freely available.); (ii) perform generic searches within the Déjà vu content by authors, address, title word, abstract word, year and comment word; (iii) perform detailed searches by specifying search criteria specific to PMID, journal names, title words, abstract, address and year; (iv) filter and view Déjà vu results in a particular category or identified by particular authors (same or different), language, availability of full text, discovery method, etc.; (v) send comments or reports to contest a record or submit a potential duplication to be reviewed by human curators; and (vi) access statistics using different filters including category, language, country, journals, etc.

For each duplicate record, a viewing window presents citations side-by-side with similarities or differences highlighted (Figure 1), providing a user-friendly interface to search, browse and facilitate rapid and rigorous interpretation of the results. Déjà vu data are also available for data mining in two formats: comma-separated values and a MySQL script to recreate the MySQL database.

CONCLUSION AND FUTURE DIRECTIONS

The Déjà vu database is the first of its kind to publically present cases of highly similar citations in Medline.

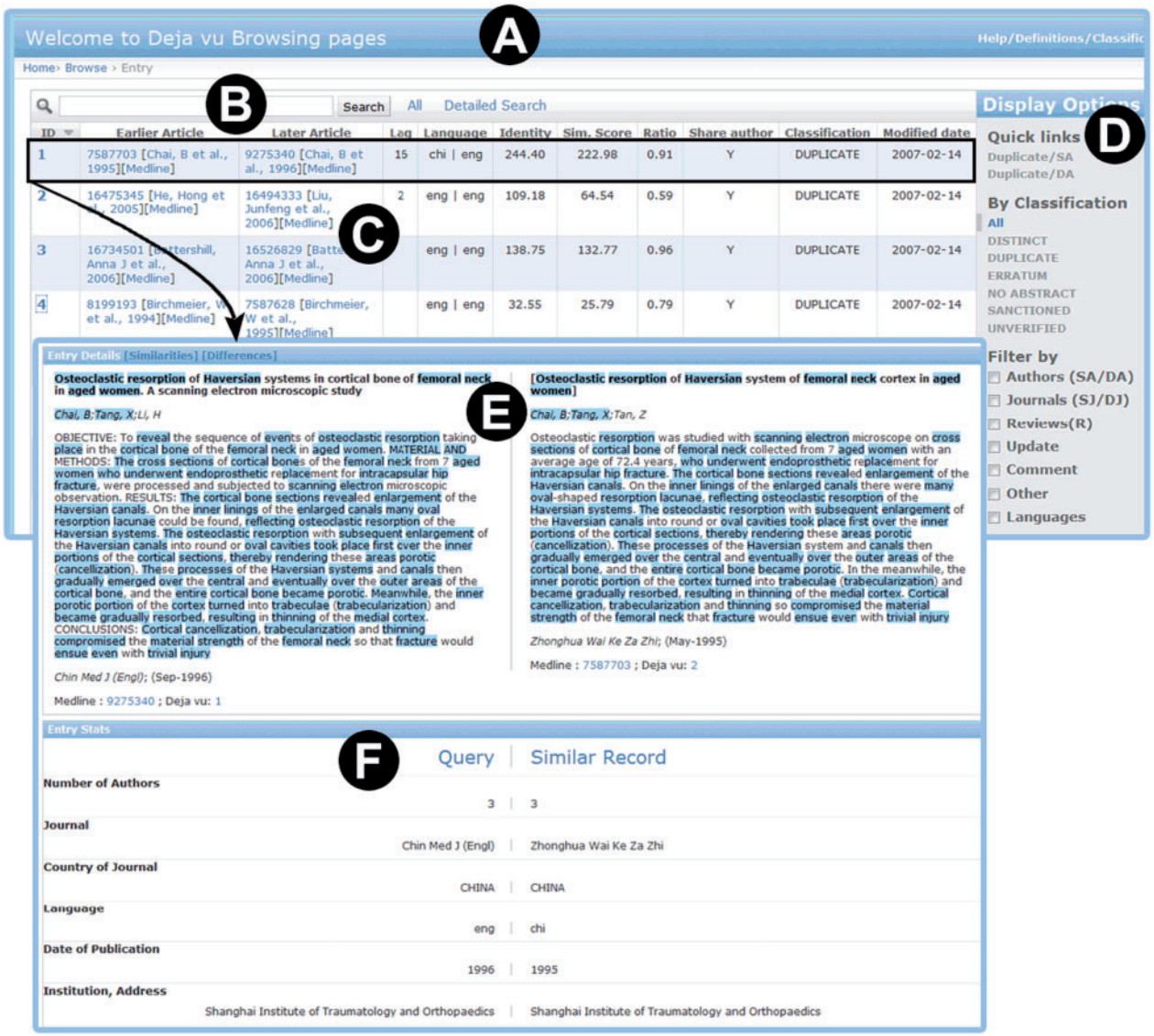

Figure 1. The Déjà vu citation presentation output. (**A**) Browsing interface for database content. (**B**) Query box to search duplicate records by author names, title, abstract, year of publication and comment words. (**C**) List of records in Déjà vu including PMIDs, author names, publication date and links to Medline citations and free full text when available. (**D**) Category filters to browse records in a particular category. (**E**) Side-by-side view of a duplicate record highlighting overlapping keywords in blue. (**F**) Miscellaneous information for each article involved.

In addition to presenting the list of highly similar citations, a goal of Déjà vu is to help scientists study in depth the behaviors of authors and the characteristics underlying multiple publications and related ethics issues surrounding the process of scientific publication. A friendly interface provides users with various browsing options along with a graphical representation of the overlapping information between citations. Ultimately, Déjà vu may act as a deterrent to the unethical practice of duplication.

Further work, currently in progress, that will substantially improve Déjà vu includes: (i) a streamlined process to update Déjà vu on a daily basis. (ii) a more collaborative approach for recruitment and qualification of topical experts as volunteer curators for specific publication areas. (iii) New methods to better address the question most often asked by authors introduced to Déjà vu, '*Am I in it, or has my work been duplicated*?' Authors can now check if their work has been duplicated by submitting their abstracts one by one directly to eTBLAST, which then flags highly similar citations for the authors to pursue. Utilities are being developed to allow authors to scan their entire bibliography at once (retrieved using Medline Entrez keyword queries) to obtain a list of highly similar citations for each citation entered. Authors will also be able to automatically submit suspicious highly similar citations found by this process directly to Déjà vu curators. (iv) Currently, duplications found in Déjà vu were obtained from Medline citations. Other literature databases will be added as they are scanned by eTBLAST, including the Institute of Physics, NASA and NIH CRISP.

ACKNOWLEDGEMENTS

The authors thank David Trusty for computer administrative support, Dr John Loadsman as a substantial contributing curator, Dr Wayne Fisher for useful comments and discussions and Linda Gunn for administrative assistance. They also wish to thank numerous Déjà vu users who have reported inaccuracies or have alerted them to questionable publications.

FUNDING

P.O'B. Montgomery Distinguished Chair (to H.G.); the Hudson Foundation (to H.G.); National Institute of Health/National Library of Medicine grant (R01 LM009758-01 to H.R.G.). Funding for open access charge: P.O'B. Montgomery Distinguished Chair.

REFERENCES

1. Budinger,T.F. and Budinger,M.D. (2006) *Ethics of Emerging Technologies, Scientific Facts and Moral Challenges*. John Wiley and Sons, NJ
2. Broad,W.J. (1981) The publishing game: getting more for less. *Science*, **211**, 1137–1139.
3. Huth,E.J. (1986) Irresponsible authorship and wasteful publication. *Ann. Intern. Med.*, **104**, 257–259.
4. von Elm,E., Poglia,G., Walder,B. and Tramer,M.R. (2004) Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *J. Am. Med. Assoc.*, **291**, 974–980.
5. Schein,M. and Paladugu,R. (2001) Redundant surgical publications: tip of the iceberg? *Surgery*, **129**, 655–661.
6. Rosenthal,E.L., Masdon,J.L., Buckman,C. and Hawn,M. (2003) Duplicate publications in the otolaryngology literature. *Laryngoscope*, **113**, 772–774.
7. Roig,M. (2005) Re-using text from one's own previously published papers: an exploratory study of potential self-plagiarism. *Psychol. Rep.*, **97**, 43–49.
8. Mojon-Azzi,S.M., Jiang,X., Wagner,U. and Mojon,D.S. (2004) Redundant publications in scientific ophthalmologic journals: the tip of the iceberg? *Ophthalmology*, **111**, 863–866.
9. Kostoff,R.N., Johnson,D., Rio,J.A.D., Bloomfield,L.A., Shlesinger,M.F., Malpohl,G. and Cortes,H.D. (2006) Duplicate publication and 'paper inflation' in the Fractals literature. *Sci. Eng. Ethics*, **12**, 543–554.
10. Gotzsche,P.C. (1989) Multiple publication of reports of drug trials. *Eur. J. Clin. Pharmacol.*, **36**, 429–432.
11. Durani,P. (2006) Duplicate publications: redundancy in plastic surgery literature. *J. Plast. Reconstr. Aesthet. Surg.*, **59**, 975–977.
12. Chennagiri,R.J.R., Critchley,P. and Giele,H. (2004) Duplicate publication in the Journal of Hand Surgery. *J. Hand Surg.*, **29**, 625–628.
13. Bloemenkamp,D.G., Walvoort,H.C., Hart,W. and Overbeke,A.J. (1999) [Duplicate publication of articles in the Dutch Journal of Medicine in 1996]. *Ned. Tijdschr. Geneeskd.*, **143**, 2150–2153.
14. Blancett,S.S., Flanagin,A. and Young,R.K. (1995) Duplicate publication in the nursing literature. *Image J. Nurs. Sch.*, **27**, 51–56.
15. Barnard,H. and Overbeke,A.J. (1993) [Duplicate publication of original manuscripts in and from the Nederlands Tijdschrift voor Geneeskunde]. *Ned. Tijdschr. Geneeskd.*, **137**, 593–597.
16. Bailey,B.J. (2002) Duplicate publication in the field of otolaryngology-head and neck surgery. *Otolaryngol. Head Neck Surg.*, **126**, 211–216.
17. Lewis,J., Ossowski,S., Hicks,J., Errami,M. and Garner,H.R. (2006) Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, **22**, 2298–2304.
18. Errami,M., Wren,J.D., Hicks,J.M. and Garner,H.R. (2007) eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res.*, **35**, W12–W15.
19. Errami,M. and Garner,H. (2008) A tale of two citations. *Nature*, **451**, 397–399.
20. Errami,M., Hicks,J.M., Fisher,W., Trusty,D., Wren,J.D., Long,T.C. and Garner,H.R. (2008) Deja vu–a study of duplicate citations in Medline. *Bioinformatics*, **24**, 243–249.