Usage Patterns of Open Genomic Data

Jingfeng Xia and Ying Liu

This paper uses Genome Expression Omnibus (GEO), a data repository in biomedical sciences, to examine the usage patterns of open data repositories. It attempts to identify the degree of recognition of data reuse value and understand how e-science has impacted a large-scale scholarship. By analyzing a list of 1,211 publications that cite GEO data to support their independent studies, it discovers that free data can support a wealth of high-quality investigations, that the rate of open data use keeps growing over the years, and that scholars in different countries show different rates of complying with data-sharing policies.

As an integral part of the e-science endeavour, making scientific data freely available for everyone to share has been a popular subject in research. Much has been published on underlining the significance of scientific collaborations in support of scholarly communication with the possibility of a broad data sharing.[1] In the literature of open data studies, a continuous shift of research topics can be observed that reflects each major stage of open scientific data development.[2] This development has experienced various efforts within a short history of less than two decades for constructing necessary infrastructure that either encourages or mandates researchers to deposit raw data in particularly designed online data repositories and fine-tunes the repository systems to facilitate data acquisition, curation, visualization, and access.

Thanks to the maturation of many data repository systems, the implementation of varying data-sharing policies, and the increasing familiarity of researchers with the concept and practice of free data circulation, a considerable body of datasets, particularly scientific data, has recently become available to the entire scholarly community. An increasing number of studies reusing these data have been published in peer-reviewed journals. Now is the time to evaluate the use of open scientific data, to identify the degree of recognition of data reuse value, and to understand how the e-science movement has impacted a large-scale scholarship. Currently, such studies are scarce in the professional literature.

This present study uses biomedical studies as a sample area of study to show the usage patterns of open data repositories. It identifies and examines journal articles that either cited genomic data in Genome Expression Omnibus (GEO), an open data repository in biomedical sciences, as evidence to support and complement their independent studies, or used GEO

Jingfeng Xia is Assistant Professor in the School of Library and Information Science at Indiana University; e-mail: xiaji@iupui.edu. Ying Liu is a Doctoral Student in the School of Economics and Management at Beijing University of Posts and Telecommunications; e-mail: fantasysoar@gmail.com.

data as the basis of statistical or analytical analyses or tools. The examination focuses on two characteristics of the publications, authorship and publication quality, to understand who have used and published the freely available data, which is represented by authors' institutional affiliations and geographic distributions, and where these articles are published, which is reflected by the journals' reputations (impact factors) and open access status. Such information will help stakeholders such as administrators, advocates, and repository managers to adjust data management policies so as to acquire more and better data and to enhance data usability. Scientists will also find this analysis to be a useful comparative resource for guiding their own data publishing.

Literature Review

Genomic Data Repositories

As early as in the 1960s, gene expression started attracting the attention of scientists.[3] Since then, numerous genomic investigations have been undertaken to generate adequate data, and new technologies have allowed simultaneous measurements across individual experiments. Although in the early days "entire strings of DNA and amino acid sequence were published in peer-reviewed journals,"[4] an increasing demand on making the data freely available online has been in existence, and a public data repository has been called for by scholars to make the data accessible for comparative analysis and for interoperability with other data resources.[5] It has been realized that "improved access to large electronic data sets, reliable and consistent annotation and effective tools for 'data mining' are critical."[6]

Prior to the 2000s, molecular biologists had witnessed the establishment of various free online genomic and proteomic databases where DNA and amino acid sequences and protein structure data could be deposited and shared.[7] In the new millennium, more efforts have been made to plan and implement open data repositories for various purposes in several major industrial countries such as the United Kingdom, the United States, Germany, and Japan. Attention has been paid to how to design an appropriate database structure to accommodate diverse genetic data formats; how to maintain a smooth data flow; how to simplify the process of data submission; how to standardize query functions; how to facilitate seamless data downloads, and how to optimize data visualization.[8] Table 1 is a list of open data repository examples that are in effect today.

The scholarly community, represented by funding agencies, journals, and professional associations, quickly responded to the data repository efforts and set policies to mandate grantees and authors to deposit their raw data in a public data repository.[9] Some data repositories made corresponding changes to provide authors the flexibility of depositing data before formal publishing of their articles and publicized the data afterward. The mandate policies have played an important role in urging researchers to make active contributions to open access and to allow scientists in biomedical disciplines to enjoy more free data than in most other academic fields. Studies have found that nearly half of recent gene expression studies "have made their data available somewhere on the internet, after accounting for datasets overlooked by the automated methods of discovery."[10] A recent mandate policy by National Science Foundation in the United States extended the requirements to fields beyond biomedical sciences.[11]

A group of studies have been conducted to evaluate the functionality, accessibility, and usability of some of the open data repositories.[12] Scholars are particularly concerned with the accuracy and entirety of query results from interoperable online resources and from any centralized repositories because the results will serve as the basis of thirty-party analyses.[13] Furthermore, researchers are interested in determining information-seeking behaviors in genomic data use and have conducted surveys and used case studies to look for data-sharing patterns as well as large-scale

TABLE 1
Examples of Open Data Repositories for Gene Expression and Genomic Hybridization Data Resources

Data Repository	URL	Data Acquisition
Serial Analysis of Gene Expression	http://www.sagenet.org/resources/index.html	In-house data
ExpressDB	http://arep.med.harvard.edu/ExpressDB/	In-house data
MAExplorer	http://www.ccrnp.ncifcrf.gov/MAExplorer/	In-house data
Public Expression Profiling Resource	http://pepr.cnmcresearch.org/browse.do	In-house data
RNA Abundance Database	http://www.cbil.upenn.edu/RAD/php/index.php	Submitted data
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/	Submitted data
Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/	Submitted data
CIBEX	http://cibex.nig.ac.jp/index.jsp	Submitted data

scientific collaborations among various types of scientists in different fields.[14]

Gene Expression Omnibus (GEO)

The Gene Expression Omnibus (GEO) project was initiated in 2000 under the supervision of the National Centre for Biotechnology Information at the National Library of Medicine.[15] Originally, GEO was formed to operate as a free data repository for high-throughput gene expression data generated mostly by microarray technologies. Over the years, the repository has expanded its content coverage to hold more data types such as genome copy number variations, genomewide profiling of DNA-binding proteins, and the next-generation sequencing technologies.

Data submitted to GEO contain three entity types: platform, a descriptive summary of the array and a data table that describes the array template; sample, an explanation of the biological objects and the experimental protocols to which it was subjected, including a data table for hybridization measures for each attribute on the matching platform; and series, a group of related samples defined as part of a research and portrays the general research objectives and strategies. The functions of GEO also include identifying and producing many related data objects to support data mining, visual presentation, and data rearrangement to alternative structures.[16]

As of summer 2011, GEO collected a total of 2,720 datasets for 9,271 platforms and 611,384 samples (www.ncbi.nlm.nih.gov/geo), in comparison to 120,000 samples found in GEO around five years ago. Researchers are allowed to submit their data to the repository and are able to use specifically designed tools and web interfaces to query and download gene expression patterns deposited by them and others via GEO-designed web interfaces and applications. GEO organizes multiple utilities to assist users in carrying out effective and accurate searches and successful downloads and then presents retrieved data in visualized forms at the level of individual genes or entire studies. On average, with current rates of submission and processing over 10,000 samples per month, GEO now receives more than 40,000 web hits and has 10,000 bulk FTP downloads in a single day.[17]

Methods

GEO also provides citation data for articles that cite the datasets and reside in PubMed. As of summer 2011, it identified a list of 13,825 recent PubMed articles that cite deposit of data in GEO, namely articles whose authors are GEO data contributors and whose dataset has been concurrently archived at publication. Although these authors are also open scientific data users, the fact that they use their own data and their roles as data creators and providers make the evidence only distantly related to our effort on seeking data usage patterns.

At the same time, GEO maintains another list of publications for a total of 1,211 journal articles representing third-party publications that use GEO data. A review of these publications shows that they either apply GEO data to validate gene expression signatures out of their own datasets or incorporate GEO data into their own analysis.[18] This list of articles serves as the basis of our analysis because these authors reuse data that is not deposited by them, thereby representing the evidence for independent use of an open data repository. Piwowar queries data against PubMed and points out the incompleteness of this list and argues that the third-party articles identified by GEO denote only 41 percent of the entire article body in PubMed that uses GEO data.[19] However, we believed that such a large number of articles in the list have been able to form an acceptable size of samples to support the statistical analysis in our study.

This list of third-party usage citations contains a standard citation for each article with such information as author(s), article title, journal name and issue, publication date, page number(s), and a PubMed ID (PMID). Clicking on a record will return the first author's affiliation and e-mail address, article abstract, a link to the free copy in PubMed (if any), publication types, MeSH terms, substances, grant support, and links to external databases that index the article. Based on the purposes of our research, some of the citation data were manually collected, including the first author's affiliation and country of origin, number of authors, publication date, journal name, and article title.

To properly evaluate the scholarly quality of the publications, all journals on the third-party list were searched against Thomson Reuter (ISI)'s Journal Citation Reports (JCR) on Web of Science to obtain their impact factors in 2010. Of the 286 journals on the list, 28 journals are not measured by JCR, which include a total of 38 articles among the 1,211 publications (approximately 0.03 percent of the research population). We believed this small fraction of missing data will not affect the analytical results for the purpose of comparing scholarly rankings of the journals.

Data were reorganized for the purposes of this analysis. A Google mapping tool was used to demonstrate a geographic distribution of the authors around the world. Both ordinal and logistic regressions as well as correlation coefficients were calculated for a series of relationships (for example, between the availability and use of datasets and between the popularity of the journals and their open access status). Some of the data and analysed results were compared to those of other related studies to seek connections between these third-party publications and the raw data cited from the GEO database.

This study did not examine the relationship between data depositors and users, although this could be an interesting piece of information for understanding data usage patterns. Such an examination will need comprehensive datasets beyond the use of one single data repository, since GEO is only one of several reputable, free biomedical data repositories. Yet we question whether quantitative data can ever present accurate information for this purpose. Alternative strategies may include distributing questionnaires among identified independent data users to comb through complex personal connections among researchers to answer the question of how and where open data sources were learned, located, accessed, identified, and used.

Results and Discussion

Date of Publication

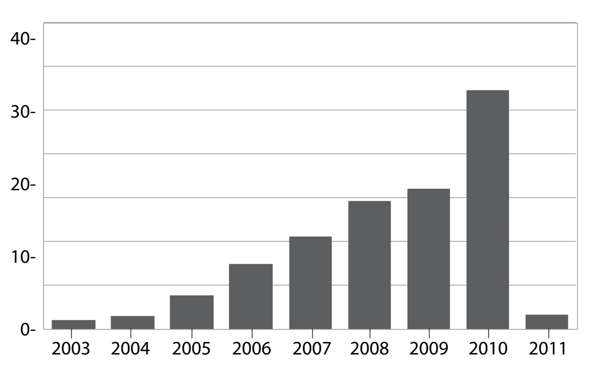
The 1,211 articles are distributed unevenly across the time period beginning in June of 2003 and ending in 2010. The list of articles was identified in mid-2011, so the data for that year is not yet complete. The first year saw five publications only, but by 2010 as many as over several hundred articles were published. An obvious trend of gradual increase in the number of publications annually is presented in figure 1. It is useful to check if this growth rate corresponds to the rate of raw data increase in the GEO database. Figure 2 is a bar chart that borrows data from figure 1 of the Barrett et al. article to display the amount of raw data submitted to GEO by year since 2000 when the repository was created.[20] These authors are the GEO staff who maintain the authoritative data for a timeline of GEO database growth. A comparison of these two figures shows a visual similarity in the chronological development of both practices. We further apply a statistical analysis to the two series of numbers by calculating their Pearson's correlation coefficients and receive the result $r = 0.98$, which is a perfect match.

It is reasonable to suggest that a large quantity of raw data can support the wealth of publications that rely on the data. The assumption that the frequency of the publications lagged behind the availability of the raw data is verified by the above comparisons. At the same time, we believe that the steady increase in both publication numbers and data numbers indicates the effect of an ongoing open access advocacy over the years. There is evidence to show that the rate of scholars' awareness of, attitudes toward, and participation in open access initiatives has climbed progressively across time and space and is spreading in most academic disciplines.[21] With regard to scientific data sharing and reuse, more researchers have recognized the value of data repositories and feel comfortable making contributions to one.[22] In the case of raw data circulations in the biomedical sciences, mandate policies have inarguably played a central role, which may serve as a successful example to inspire the open access efforts in other fields.

Venue of Publication

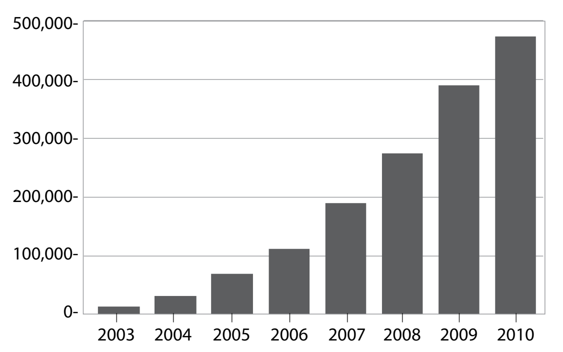
A total of 286 journals accommodate these 1,211 articles. On one hand, most of the articles cluster in a few journals: for example, ten journals publish nearly 48 percent of the total articles (see table 2). On the other hand, a large number of other journals are under the radar of the authors when selecting the venue of publication (see figure 3). This distribution pattern is the result of subject concentrations, because the popular journals found here all bear a scope matching the subjects of the published research, as well as the result of the genomic nature of the raw data in GEO.

FIGURE 1
Article Numbers by Year of Publication



The quality of these articles can be measured by counting citation rate, or the impact factor (IF) of a journal, for the past two years. The scholarly community has extensively used IF as one of the key indexes to assess the popularity and impact of publications, although its limitations have also been documented.[23] IF is especially preferred by academics in biomedical and some allied scientific disciplines.[24] Given the

FIGURE 2
Data Submitted to GEO by Year



reputation of the GEO repository in relevant fields, we expect to see high-quality publications using the GEO data.

We are not disappointed with the findings that the articles have generally been published in reputable journals. Several internationally renowned journals are visible on the list, and even the famous journals *Nature* and *Science* cannot reach the top of the ranking (see table 3). The top ten journals all have an IF value above 25.00, and there are as many as 33 journals with an IF value above 10.00. Individual journals that have published a large number of the articles and are also ranked highly include *BMC Bioinformatics* (IF=3.028, article=108), *Plos ONE* (IF=4.411, article=89), *Nucleic Acids Research* (IF=7.836, article=82), *Bioinformatics* (IF=4.877, article=79), *BMC Genomics* (IF=4.206, article=67), and *Proceedings of the National Academy of Sciences* (IF=9.771, article=49). On average, the 286 journals are scored around 5.00 in impact factor (see figure 4).

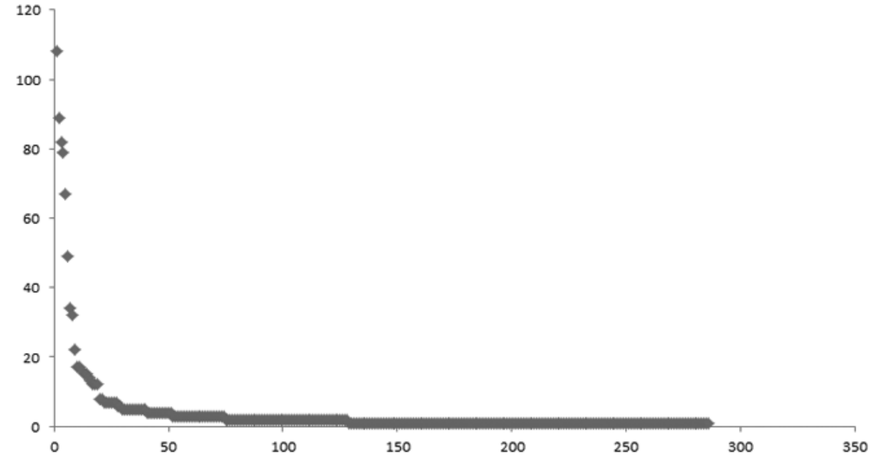
Studies have verified a positive relationship between open access article publishing and research quality,[25] although there are disagreements with this view.[26] It is also generally believed that, after an article becomes freely accessible, it will be cited more frequently. However, relatively few studies have been undertaken to test the impact of open data reuse on research quality. Piwowar et al. undertake one of these few studies examining the connection between citation rate of a publication and the public availability of its data by using data from cancer microarray clinical trials.[27] They find a significance of p=0.006 for free scientific data to be associated with a 69 percent increase in citation counts, independently of journal IFs, publication dates, and author locations. Our own study cannot validate their findings because of the different research designs and data types as well as the lack of a direct comparison of citation data.

Putting aside scientific evidence, one may not find it difficult to understand why an article will attract more citations after it becomes accessible to the public. The visibility of the article on the web and convenience of access in lieu of a cost barrier are favorable conditions for wider distribution. However, it is not logical to assume a citation difference

TABLE 2
Top Ten Journals Where Articles Using the GEO Data Are Published

Journal	Article	Open Access	Impact Factor
BMC Bioinformatics	108	Yes	3.028
PLoS One	89	Yes	4.411
Nucleic Acids Research	82	Yes	7.836
Bioinformatics	79	No	4.877
BMC Genomics	67	Yes	4.206
Proc Natl Acad Sci (PNAS)	49	No	9.771
Cancer Research	34	No	8.234
Genome Biology	32	Yes	6.885
Clinical Cancer Research	22	No	7.338
BMC Cancer	17	Yes	3.153

FIGURE 3
The Distribution of Articles in Number by Journal



between publications *using* open data and publications *using* in-house data, unless the discussion is about the advantages of *providing* publicly accessible data. It is beyond the scope of this research to seek usage patterns among various types of open data repositories. Our findings have clearly revealed that open data does support high-quality research in general; in other words, researchers are able to publish high-quality articles by using freely available data contributed by others.

TABLE 3
Journals with the Highest Impact Factor

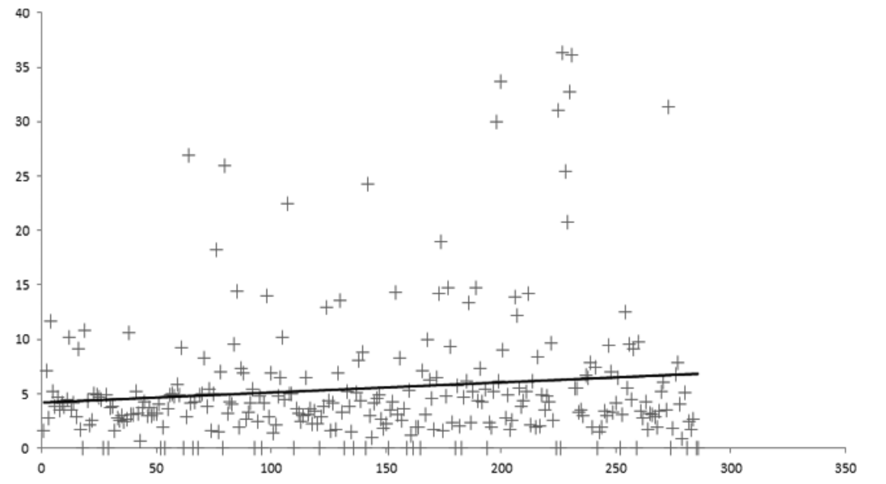
Journal	Count	Impact Factor
Nature Genetics	8	36.377
Nature	5	36.101
Lancet	2	33.633
Nature Reviews Genetics	1	32.745
Science	3	31.364
Nature Biotechnology	3	31.085
JAMA	1	30.011
Cancer Cell	2	26.925
Cell Stem Cell	2	25.943
Nature Medicine	1	25.43

Open Access Preference

Further, it is interesting to examine if the open access status of the journals on this list has a positive influence on authors' choice of a journal when seeking publication. According to the numbers in table 2, among the top five journals that contain 35 percent of the total articles, four journals are operated in an open access mode. One can, therefore, hypothesize that researchers prefer open access journal publishing after they become aware of the open access benefits through personal experience—either through self-depositing raw data to an open data repository or through sharing raw data from other data resources.

A logistic regression analysis was conducted between two variables (that is, open access status of a journal and number of our articles in the journal) to determine authors' preference for open access journals when publishing an article. The results (p>0.05) does not seem to support a relationship between the two variables; namely, authors who use GEO data may not necessarily take the open access status of a journal into consideration (see table 4). Factors other than open access (for

FIGURE 4
Distribution of Journal Impact Factors



example, the subject relevancy of a journal and the reputation of the journal on the IF index) may have played a more important role for selecting publication venues. The hypothesis is, therefore, rejected.

Author Profile

The distribution of countries of origin by author, which is defined by self-identified affiliation of the first author, is rather diverse (see figure 5). The United States is the single largest country with regard to its number of the authors, followed by the United Kingdom, China, and several other western European countries. We take the numbers for the United States and the United Kingdom for granted because the majority of the raw data in GEO are contributed by researchers in these two countries. On the other hand, the number of studies using GEO data and the amount of data submitted to GEO for China may not have a positive correlation. However, this supposition needs to be confirmed by collecting and analyzing the numbers of data contributors for appropriate comparisons, which may be one of our future research projects.

Is there a chronological change among countries for using GEO data? To answer this question, an ordinal regression analysis was performed. Only top data-producing countries are coded for the analysis to keep the statistical output short: 1=U.S., 2=U.K., 3=China, 4=Germany, 5=Japan, and 6=other countries. Table 5 shows the analyzed result and indicates that, at the significant level (p<0.05), researchers in the United States and China have experienced a change of open data use over time, while researchers in Japan and European countries have not changed their usage pattern significantly. A possible explanation is that European and Japanese scholars are among the early adopters who participated in open access initiatives from the beginning when GEO data started being publicly available.

TABLE 4
Logistic Regression of Open Access Publishing (Variables in the Equation)

		B	S.E.	Wald	df	Sig.	Exp (*B*)
Step 1	Articles	.022	.015	2.138	1	.144	1.022
	Constant	–.778	.535	2.112	1	.146	.459

FIGURE 5
Authors' Origins by Country (Top 20 Only)

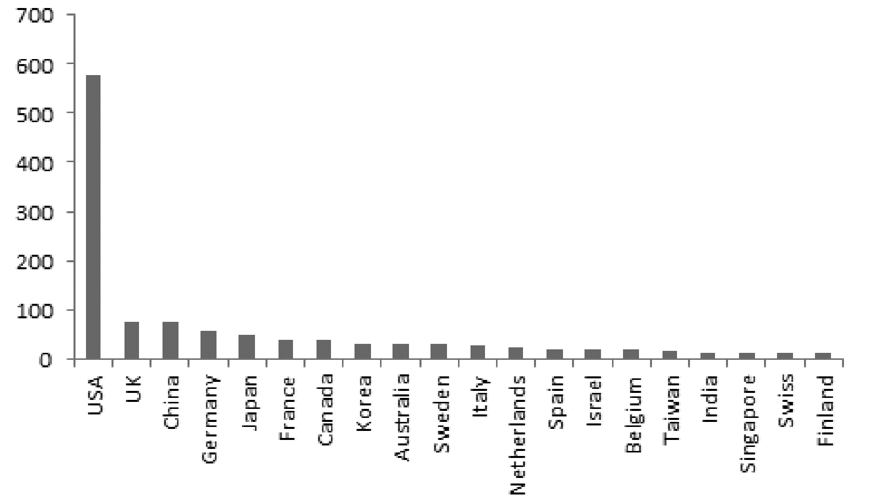


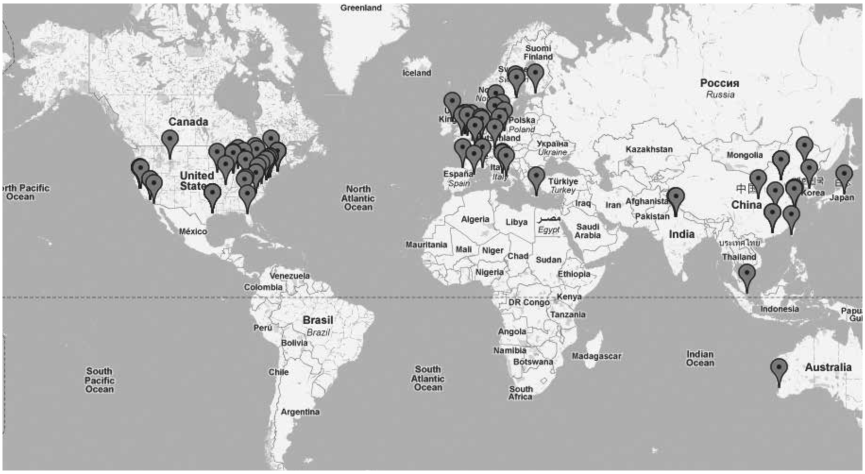
TABLE 5
Ordinal Regression of Publication Years by Country

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[Year = 2003]	–5.648	.455	153.973	1	.000	–6.540	–4.755
	[Year = 2004]	–3.868	.204	359.293	1	.000	–4.268	–3.468
	[Year = 2005]	–2.729	.137	396.374	1	.000	–2.998	–2.461
	[Year = 2006]	–1.800	.111	264.211	1	.000	–2.017	–1.583
	[Year = 2007]	–1.046	.100	108.929	1	.000	–1.242	–.849
	[Year = 2008]	–.284	.096	8.797	1	.003	–.471	–.096
	[Year = 2009]	.498	.096	26.654	1	.000	.309	.687
	[Year = 2010]	3.833	.223	295.046	1	.000	3.395	4.270
Location	[Code = 1]	–.327	.118	7.694	1	.006	–.558	–.096
	[Code = 2]	–.162	.223	.524	1	.469	–.600	.276
	[Code = 3]	.557	.234	5.686	1	.017	.099	1.015
	[Code = 4]	–.069	.255	.073	1	.786	–.569	.431
	[Code = 5]	–.089	.270	.107	1	.743	–.618	.441
	[Code = 6]	0a			0			

Link function: Logit
a. This parameter is set to zero because it is redundant.

FIGURE 6
Geographic Distribution of the Authors by Country



Consequently, open access advocates should be more interested in exploiting the United States' and China's scholarly market, where potential for increased data use is noteworthy. A more thorough analysis of this topic in the future should include all countries that appear in the usage list. Furthermore, a chronological change of data *submissions* by scholars in different countries should also be examined to draw a more comprehensive picture of the open data development.

Furthermore, what is obvious as demonstrated by our numbers is a *divide* of data usages between the developed countries and the developing countries. Figure 6 is a Google map for a global view of the geographic distribution of the authors by country. Please note that points plotted on the map represent proportion of the original values for demonstration only. The shown distribution pattern suggests an overwhelming silence in the scholarly community of the third-world countries. We recommend more comprehensive studies to examine the academic infrastructure in these countries.

Conclusion

The analysis helps paint a clear picture of free data usage out of the GEO data repository and verify the assumption that the raw data do support a wealth of high-quality investigations, that the rate of open data use keeps growing over the years, and that scholars in different countries show different rates of complying with the data-sharing policies. GEO data are heterogeneous in type, which have considerably supported research of specific topics as well as research across a magnitude scale of independently submitted samples.[28] Third-party data users will continuously benefit from using ever-accumulating free data to make contributions to scholarship.

The findings can also serve as optimistic signs to highlight the influence of the open access movement among individual scholars, no matter whether they are mandated to make contributions of raw data by policies or voluntarily self-archive and use freely available data for their own scientific investigations. Success stories of open access initiatives in biomedical sciences can help various stakeholders such as administrators, open access advocates, system designers, and repository managers in other academic fields such as social sciences and humanities to adapt better strategies for promoting data sharing and reuse. Other types of open access initiatives such as e-print repositories and electronic journal publishing can also learn from the open data development. It is hoped that the significance and implications of our research can extend beyond genomic subfields.

There is no doubt that research efficiency and quality can be improved by researchers sharing and reusing primary research datasets. Among other advantages, freely available raw data "can be used to explore related or new hypotheses, particularly when combined with other available datasets."[29] The potential of open data applications is tremendous.

Notes

1. D.E. Atkins, K.K. Droegemeier, S.I. Feldman, H. Garcia-Molina, M.L. Klein, and P. Messina, *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (Washington D.C.: National Science Foundation, 2003), available online at www.nsf.gov/od/oci/reports/toc.jsp [accessed 1 September 2011]; Carole L. Palmer and M.H. Cragin, "Scholarship and Disciplinary Practices," *Annual Review of Information Science and Technology* 42, no. 1 (Nov. 2008): 213–95; D.H. Sonnenwald, "Scientific Collaboration," *Annual Review of Information Science and Technology* 41, no. 1 (Oct. 2007): 643–81.

2. T. Barrett et al., "NCBI GEO: Archive for Functional Genomics Data Sets—10 Years On," *Nucleic Acids Research* 39 (Nov. 2011): D1005–D1010; T. Barrett et al., "NCBI GEO: Mining Tens of Millions of Expression Profiles—Database and Tools Update," *Nucleic Acids Research* 35 (Nov. 2007): D760–D765; C. Brown, "The Changing Face of Scientific Discourse: Analysis of Genomic and Proteomic Database Usage and Acceptance," *Journal of the American Society for Information Science and Technology* 54, no. 10 (Aug. 2003): 926–38.

3. J.H. Taylor, *Selected Papers on Molecular Genetics* (New York: Academic Press, 1965).

4. Brown, "The Changing Face of Scientific Discourse," 926.

5. A. Brazma, A. Robinson, G. Cameron, and M. Ashburner, "One-Stop Shop for Microarray Data," *Nature* 17, no. 403 (Feb. 2000): 699–700; O. Ermolaeva et al., "Data Management and Analysis for Gene Expression Arrays," *Nature Genetics* 20, no. 1 (Sept. 1998): 19–23; P. Kellam, "Microarray Gene Expression Database: Progress Towards an International Repository of Gene Expression Data," *Genome Biology* 2, no. 5 (May 2001): reports4011.1–reports4011.3.

6. D.E. Bassett Jr., M.B. Eisen, and M.S. Boguski, "Gene Expression Informatics—It's All in Your Mine," *Nature Genetics* 21, no. 1 (Jan. 1999), 51.

7. K.W. McCain, "Sharing Digitized Research-Related Information on the World Wide Web," *Journal of the American Society for Information Science* 51, no. 14 (Dec. 2000): 1321–27.

8. T. Barrett et al., "NCBI GEO: Mining Tens of Millions of Expression Profiles"; J.C. Bartlett and E.G. Toms, "Developing a Protocol for Bioinformatics Analysis: An Integrated Information Behaviour and Task Analysis Approach," *Journal of the American Society for Information Science and Technology* 56, no. 5 (Mar. 2005): 469–82; C.R. Johnson, R. MacLeod, S.G. Parker, and D. Weinstein, "Biomedical Computing and Visualization Software Environments," *Communications of the ACM* 47, no. 11 (Nov. 2004): 64–71; G.D. Schuler, J.A. Epstein, H. Ohkawa, and J.A. Kans, "Entrez: Molecular Biology Database and Retrieval System," *Methods in Enzymology* 266 (1996): 141–62.

9. L. Goodman, "Unlimited Access—Limitless Success," *Genome Research* 11 (Jan. 2001): 637–38; K. McCain, "Mandating Sharing: Journal Policies in the Natural Sciences," *Science Communication* 16, no. 4 (June 1995): 403–31; National Institutes of Health, *NOT-OD-03-032: Final NIH Statement on Sharing Research Data*; H. Parkinson et al., "ArrayExpress Update–From an Archive of Functional Genomics Experiments to the Atlas of Gene Expression," *Nucleic Acids Research* 37 (Jan. 2009): D868–72.

10. H.A. Piwowar, "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data," *PLos One* 6, no. 7 (July 2011): e18657; S.A. Ochsner, D.L. Steffen, C.J. Stoeckert, and N.J. McKenna, "Much Room for Improvement in Deposition Rates of Expression Microarray Datasets," *Nature Methods* 5, no. 12 (Dec. 2008): 991.

11. Nation Science Foundation, *Grant Proposal Guide*, Chapter II.C.2.j, 2011.

12. A.J. Butte and R. Chen, "Finding Disease-Related Genomic Experiments within an International Repository: First Steps in Translational Bioinformatics," *AMIA Annual Symposium Proceedings* 2006: 106–10; J. Dudley and A.J. Butte, "Enabling Integrative Genomic Analysis of High-Impact Human Diseases through Text Mining," *Pacific Symposium on Biocomputing* 2008: 580–91; Y.A. Lin, A. Chiang, R. Lin, P. Yao, R. Chen, and A.J. Butte, "Methodologies for Extracting Functional Pharmacogenomic Experiments from International Repository," *AMIA Annual*

Symposium Proceedings 2007: 463–67; H.A. Piwowar and W.W. Chapman, "Recall and Bias of Retrieving Gene Expression Microarray Datasets through PubMed Identifiers," *Journal of Biomedical Discovery and Collaboration* 5 (Mar. 2010): 7–20.

13. C. Letondal, "Participatory Programming: Developing Programmable Bioinformatics Tools for End-Users," in *End-User Development*, eds. H. Lieberman, F. Paterno, and V. Wulf (London: Springer, 2005); J. Massar, M. Travers, J. Elhai, and J. Shrager, "BioLingua: A Programmable Knowledge Environment for Biologists," *Bioinformatics* 21, no. 2 (2005): 199–207.

14. P.K. Chilana, C.L. Palmer, and A.J. Ko, "Comparing Bioinformatics Software Development by Computer Scientists and Biologists: An Exploratory Study," *31st International Conference on Software Engineering* (Vancouver, May 2009); M. Umarji and C. Seaman, "Informing Design of a Search Tool for Bioinformatics," *Proceedings of the ICSE Workshop on Software Engineering for Computational Science and Engineering* (2008).

15. R. Edgar, M. Domrachev, and A.E. Lash, "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository," *Nucleic Acids Research* 30, no. 1 (Jan. 2002): 207–10.

16. Barrett et al., "NCBI GEO: Mining Tens of Millions of Expression Profiles"; Edgar, Domrachev, and Lash, "Gene Expression Omnibus."

17. Barrett et al., "NCBI GEO: Archive for Functional Genomics Data Sets."

18. Ibid.

19. H.A. Piwowar, "A Method to Track Dataset Reuse in Biomedicine: Filtered GEO Accession Numbers in PubMed Central," *ASIS&T Annual Meeting* (2008).

20. Barrett et al., "NCBI GEO: Archive for Functional Genomics Data Sets."

21. J. Xia, "A Longitudinal Study of Scholars Attitudes and Behaviours toward Open-Access Journal Publishing," *Journal of the American Society for Information Science and Technology* 61, no. 3 (Mar. 2010): 615–24; J. Xia, "Diffusionism and Open Access,*" Journal of Documentation* 68, no. 1 (Jan. 2010): 72–99.

22. Letondal, "Participatory Programming."

23. E.T. Funkhouser, "The Evaluative Use of Citation Analysis for Communications Journals," *Human Communication Research* 22, no. 4 (June 1996): 563–74; L.I. Meho and K. Yang, "Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar," *Journal of the American Society for Information Science and Technology* 58, no. 13 (Nov. 2007): 2105–25.

24. M. Bordons, M.T. Fernández, and I. Gómez, "Advantages and Limitations in the Use of Impact Factor Measures for the Assessment of Research Performance," *Scientometrics* 53, no. 2 (2004): 195–206; A.M. Diamond Jr., "What Is a Citation Worth?" *Journal of Human Resources* 21 (1986): 200–15.

25. K. Antelman, "Do Open Access Articles Have a Greater Research Impact?" *College and Research Libraries* 65, no. 5 (2004): 372–82; Y. Gargouri, C. Hajjem, V. Larivière, Y. Gingras, L. Carr, T. Brody, and S. Harnad, "Self-Selected or Mandates, Open Access Increases Citation Impact for Higher Quality Research," *Plos One* 5, no. 10 (2010): e13636; S. Lawrence, "Free Online Availability Substantially Increases a Paper's Impact," *Nature* 411, no. 6837 (May 2001): 521; A. Swan, *The Open Access Citation Advantage: Studies and Results to Date* (technical report, University of Southampton, 2010), available online at http://eprints.ecs.soton.ac.uk/18516 [accessed 1 September 2011]; J. Xia, S.K. Wilhoite, and R.L. Myers, "Multiple Open Access Availability and Citation Impact," *Journal of Information Science* 37, no. 3 (June 2011): 19–28.

26. P.M. Davis, B.V. Lewenstein, D.H. Simon, J.G. Booth, and M.J.L. Connolly, "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial," *BMJ* 337 (July 2008): a568.

27. H.A. Piwowar, R.S. Day, and D.B. Fridsma, "Sharing Detailed Research Data is Associated with Increased Citation Rate," *Plos One* 2, no. 3 (2007): e308.

28. H. Huang, C.C. Liu, and X.J. Zhou, "Bayesian Approach to Transforming Public Gene Expression Repositories into Disease Diagnosis Databases," *PNAS USA* 107, no. 15 (Apr. 2010): 6823–28; S. Suthram, J.T. Dudley, A.P. Chiang, R. Chen, T.J. Hastie, and A.J. Butte, "Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets," *PLoS Computational Biology* 6, no. 2 (Feb. 2010): e1000662.

29. Piwowar, "Who Shares? Who Doesn't?" e18657.