Panel on Use of Proprietary Data

Jamie Callan
School of Computer Science
Carnegie Mellon University
callan@cs.cmu.edu

Alistair Moffat
Department of Computing and Information Systems
The University of Melbourne
ammoffat@unimelb.edu.au

Abstract

A panel discussion on the use of proprietary data was held at SIGIR 2012 in Portland. This report summarizes the positions put forward by the six panelists and the points that arose during the wider discussion that followed.

1 Background

There has been much discussion over the last few years about the use of proprietary data – that is, data that is not made available to other researchers, even on a restricted or licensed basis – to undertake IR research that is then submitted for publication in peer-reviewed scientific conferences and journals. In the context of IR research, proprietary data is typically query logs, click logs, user demographic information and/or long-term trend information or mined analysis, user-contributed information such as social media postings, and data collected during user experiments that, under the terms of an institutional ethics approval, must be held securely. Researchers who have access to such data may be able to undertake investigations that would simply not be possible via the use of publicly available and/or licensed shared data resources.

How should such research be viewed by peer-reviewed scientific conferences and journals? Researchers sometimes make mistakes, interpret outcomes generously, or misinterpret the interactions in complex systems, thus leading to incorrect conclusions. The science disciplines have an extended tradition of encouraging scepticism of results until independently reproduced; and value the process of followup verification of experiments. Those traditions might seem to suggest that experimental studies submitted for publication should draw only on publicly-available data resources, so that repeatability and reproducibility can be assured.

However, what should be done when there is no equivalent, publicly-available data? Some types of data will *never* be publicly available to or replicable by other scientists. Laws,

regulations, privacy, and competitive reasons will restrict the sharing of some types of data. Independently gathering some types of data will require user populations or infrastructure unavailable to most scientists. Web data is the most visible example in the IR community, but the issue is more general. It includes enterprise, mobile, and medical data; user populations; and computational infrastructure. Should publication of papers derived from the examination of such data be prohibited, because they may be less reliable and perhaps less generally applicable? Or, should it be allowed, because the research community learns things it would not know otherwise?

Questions about how to handle research done with proprietary data have been occurring more prominently and with greater frequency. During the review process for the SIGIR 2012 conference, the Program Chairs noticed a number of reviews that questioned or criticized the use of proprietary data; these did not affect the decision process, but they reflect growing concerns within the IR community. The issue is also receiving greater attention outside of the IR community, for example, in a letter to the editor in *Nature* magazine [5]; a *New York Times* article [6]; and a paper in *PLoS ONE* [2]. In 2011, the U.S. National Science Foundation (NSF) amended its *Award and Administration Guide* to state:

> "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."[1]

In this context, the goal of SIGIR 2012 Panel on the Use of Proprietary Data was to raise these issues and hear diverse opinions, with a view to providing a list of possible options to the SIGIR Executive Committee in regard to developing a formal policy about the use of proprietary data in IR publications.

2 Panelists and Format

An estimated 150 people attended the panel. The moderator, Jamie Callan, presented an overview of the topic, and outlined some of the key questions that had arisen in previous discussions. The six panelists were then introduced and invited to present their perspectives on those issues, and to add further themes. Listed here in the order they spoke, the panelists were:

Norbert Fuhr	University of Duisburg-Essen
Susan Dumais	Microsoft Research
Ricardo Baeza-Yates	Yahoo! Research
William Webber	University of Maryland
Maarten de Rijke	University of Amsterdam
William Hersh	Oregon Health & Science University

Each panelist made a brief opening statement. Questions and opinions were then taken from the floor, with responses from one or more of the panelists elicited where appropriate. The next section summarizes the issues that were raised by the panelists and attendees, and the responses that emerged to them.

[1] `http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4`, accessed 8 September 2012.

3 Themes and Issues

Quality Science: One of the most oft-repeated concerns is that research that is illustrated and validated through the use of proprietary data is difficult to validate by others, including referees, whose intuitions may not be well-aligned with the results claimed by the paper. Proponents of this point of view argue that science is based on repeatable experiments; it is repeatability that gives us the confidence that the science is "true". They conclude that since IR wishes to be treated as a science, we must adhere to that standard; we do not want a "cold fusion" event in our discipline.

There is also the possibility that researchers make mistakes – as one panelist rather wryly put it:

> "If you get great experimental results, you have either made a mistake, or you are a genius. The prior on the former is probably greater than the prior on the latter."

And even if no mistakes have been made, researchers may not fully understand the interactions taking place in a complex system. Then later, when they or others re-examine the work, they find other reasons for the observed outcomes. The TREC notebook papers are sometimes reviewed and revised later, for example.

A more subtle effect suggested by one panelist is that subsequent work risks being thwarted or devalued – if a better idea is developed by someone without access to the same data as a first-mover publication, they may be unable to quantify the validity of their own ideas in comparison to now-public statements made in the original paper using the proprietary data, and hence find it difficult to achieve publication. And, while a variety of data sources is certainly a healthy situation, if there is only limited commonality of data sets across a corpus of published papers, it may be difficult for longitudinal evaluations of the form done by Armstrong et al. [3] to be carried out. Without the ability to compare results across papers, it is suggested, experimental evaluations are of little more than anecdotal value. Moreover, simple replications (even on different data) are seldom accepted for publication, creating an impetus for new algorithms to be developed.

The issue of making public a "data description" rather than the data itself was raised in this context. For example, if a high-fidelity model of some dataset is made available, it may be realistic for another researcher to use the model to generate synthetic data that has the same characteristics as the real data, but is nevertheless not the real data. Anonymizing real data before releasing it is one step towards such a data description. It might be that generic data descriptions can also yield usable outcomes; but before that can be possible fresh work will be needed in terms of how best to characterize test collections, and to establish which of the many possible parameters are the crucial ones. Certainly, the current SIGIR refereeing climate is such that papers that work purely with synthetic data are unlikely to be accepted.

It was also observed that if data is specific to one company, there may be little transferable benefit inherent in any experimental conclusions drawn from that data, and hence any paper containing those results may be of only limited interest to the broad IR community. In response, it must be noted that the IR research community includes a significant number of people employed by industry, who are very much interested in hearing about techniques carried out on other companies' proprietary data.

Another issue that was raised was that of possible publication bias – that papers that used inadequate experimental methods might in fact be more likely to be accepted, because

it might be easier to demonstrate statistically significant improvements. On the other hand, papers in which state-of-the-art baseline results are provided can, at face value, may look less "impressive". For public data such as TREC experiments, baseline effectiveness numbers are known, and so weak experimental baselines in submitted papers can be detected by alert referees. But if the data is proprietary, and the referee does not have a good understanding of the provenance and characteristics of the data, they are unlikely to be in a position to be able to say "Those baselines don't seem right".

There was also some doubt voiced about how widespread such review-time checking actually would be, even if the data were available. One comment made during the general discussion noted this suspicion:

> "People don't want data to reproduce results. They want data so that they can explore new things."

Also expressing doubt, one of the panelists commented:

> "If anyone wants to cheat with their results, they can. It has nothing to do with the data they used."

Data versus Software: Several speakers commented that a lack of sharing of software was just as much an obstacle to reproducible research as was a lack of shared data, and that where poor baselines were being used in experimental studies, it was more of a software issue than a data issue. A comment from the floor summed up the situation as:

> "We complain about things we can't control (proprietary data sharing) and don't do the things we can do, e.g. share source code ... Documenting code/data, hosting it, answering questions, etc take time that we do relatively little to encourage or recognize."

Other communities: Some disciplines and/or conferences have data disclosure and archiving requirements as part of the publication process. For example the Knowledge Discovery from Data (KDD) conference includes the following statement in its submission instructions:[2]

> "Submitted papers will be assessed based on their novelty, technical quality, potential impact, and clarity of writing. For papers that rely heavily on empirical evaluations, the experimental methods and results should be clear, well executed, and repeatable. Authors are strongly encouraged to make data and code publicly available when possible."

A possible formalization of this request might be to require that each submitted paper be accompanied by an "data and software declaration" that listed the software and data resources that had been used to generate the results in the paper, and if any of those resources are public, where they are available from, and under what conditions.

One panelist commented that the VLDB community had sought to verify computational experiments by re-running code and data during the refereeing process, and that the biggest obstacle in this process had been the code. In this sense, the software needs to be portable and sharable even before questions can be asked about data.

[2] `http://kdd2012.sigkdd.org/author_instructions.shtml`, accessed 9 September 2012.

In the bioinformatics community, data sharing is a de facto standard, and in some journals and conferences the publication process requires that the data used be added to a repository. The premier journal *Bioinformatics* stipulates:[3]

> "*Supporting Data*: All data on which the conclusions given in the publication are based must be publicly available. Bioinformatics fully supports the recommendations of the National Academies regarding data sharing (see Board on Life Sciences, Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences. Available at `www.nap.edu/books/0309088593/html`)."

Papers that primarily describe software tools face an equally precise requirement when submitted to *Bioinformatics*:

> "*Software*: If the manuscript describes new software tools or the implementation of novel algorithms the software must be freely available to non-commercial users at the time of submission, and appropriate test data should be made available. ... Authors must also ensure that the software and test data is available for a full two years following publication.

However, Alsheikh-Ali et al. [2] found that high-impact journals frequently do not enforce their data-sharing requirements.

Legal concerns: In 2006 AOL released a search log containing twenty million queries spanning a three month period. Users were identified by a "user id" rather than by name; nevertheless by collating across search queries from users, it was possible for a small number of people and their search histories to be identified. In this case use of a relatively weak anonymization process, and the fact that the release was public rather than to researchers alone (and hence there was no Data Use Agreement required), meant that what was intended as an act of generosity became instead a company embarrassment and a subject of litigation.[4]

Subsequent research has demonstrated the difficulty of masking personal information in query logs [4], and the AOL incident has caused companies to be very wary of data releases, whether accompanied by Data Use Agreements or not. As another example, a release of anonymized Wikipedia queries in September 2012 was reversed just a day later when it was discovered that some queries contained identifying data, including material that had inadvertently been pasted into the query box from users' computers' clipboards.[5]

The Lemur Toolbar project[6] was developed as a way for the academic IR community to collect their own log data, but it had relatively low uptake, and has not generated any useful resources. It seems that even the researchers who seek access to such log data have concerns about their personal privacy.

In the specific context of query logs, it was noted that academics had the option of approaching their universities' network providers, and seeking access to query logs that way. And if their internal Ethics Approval Board processes denied them access to that data, challenging that refusal might still be an easier process than seeking external resources.

[3] `http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/general.html`, accessed 9 September 2012.

[4] `http://en.wikipedia.org/wiki/AOL_search_data_leak`, accessed 8 September 2012.

[5] `http://blog.wikimedia.org/2012/09/19/what-are-readers-looking-for-wikipedia-search-data-now-available/`, accessed 23 September 2012.

[6] `http://www.lemurproject.org/querylogtoolbar/`, accessed 8 September 2012.

Some data is simply not public: It was acknowledged that some forms of data are never going to be publicly available, and yet the research that builds on that data is indisputably science. For example, in a clinical trial use is made of a pool of people, some of whom have (or are at risk of) the condition in question and some of whom – ideally demographically equivalent in all other respects – do not. The researcher then supplies these people with drugs or placebos, and monitors any changes to their health. Other researchers would not expect to have access to the same group of people, or even to know their names or addresses.

User-studies research – for which institutional Ethics Approval Board permission is almost always required – faces similar restrictions on the use of collected data. In an ethics application it is usual to name the people who will have access to the data, describe the purposes to which the data will be put, and indicate the long-term storage and security regime that will apply, even before any data is collected. Hence, in this important area of IR research it is also extremely difficult to undertake research without making use of data that might be regarded as being "proprietary" – albeit, held confidentially within a University rather than within a commercial company.

One panelist commented that many of their computer science colleagues had little understanding of ethics approval processes and why they were important, and noted that in technology-focused institutions there might also be university-level confusion in regard to what constraints should be put in place when working with human subjects.

Opportunity rather than handicap: The industry-based panelists noted that there were a variety of ways to access proprietary data, and that them "handing it over" was only one of those options. Alternatives that were discussed included:

- *Take the code to the data*: It might be possible for software to be applied to proprietary data within the security perimeter of the data owner, and for only the results to be communicated outside the firewall. However, some were skeptical that companies would be willing to take on the likely complexity of nursing software compilations and executions.
- *Send an intern to the data*: A more palatable option for companies is to host a graduate student as an intern, thus giving the student researcher access to proprietary resources. One concern raised by the audience is who owns any intellectual property (IP) produced during the internship.
- *Come to the data yourself*: One of the industry-based panelists noted that their organization welcomed academic visitors. The same IP issues noted above would arise.
- *Access equivalent data from within your own institution*: This can be done, for example, by seeking permission to make use of institutional proxy logs, or by asking colleagues to install instrumented browsers.
- *Create the data yourself*: One panelist suggested that creation and maintaining a search service was a guaranteed way of accessing query and click data streams.
- *Think laterally*: A panelist commented:

  > "Do we want to be where the data is, or where the interesting questions are? Relying on shared data is very limiting, like looking for [lost] car keys under the street light",

and subsequently argued that restricting attention to only public data might fundamentally limit the kinds of questions that researchers ask, and hence give rise to unanticipated observation bias as a result of the sampling process used to generate public data sets.

- *Build simulators*: If reliable mechanisms for generating synthetic data are developed, they can be validated for usefulness against proprietary data, and then made available for unrestricted use.

Large-scale experiments in other scientific disciplines were mentioned as exemplars of shared "travel to the location" activities, including the Large Hadron Collider, and astronomical observatories in various parts of the world. (Although it was also observed that these shared endeavors were largely funded by multi-government grants rather than commercial entities). Panelists noted that as well as the "industry versus academic" gap in access to data resources that was the subject of the Panel, there was also a social-equity gap that was perhaps even more important, between researchers in developed countries who had access to the computational resources required to deal with terabyte-scale data collections, and researchers in developing countries who did not.

A theme that arose several times was that the research using proprietary data would take place anyway, regardless of whether the work was published in peer-reviewed form, and that patent filings were another way of asserting IP precedence.

4 Possible Options

The discussion by the panelists and the audience identified a number of options, not necessarily mutually exclusive, that could be considered by the SIGIR Executive Committee.

Status Quo: Change should not be made for its own sake, and it might be that the current situation is not sufficiently broken that it requires fixing.

Separate streams: The tasks studied using large-scale proprietary data are often of great practical interest, and worthy of publication in some format or another. One suggestion that arose was that "industry data" papers be collected together, and focused in one day of the conference, or in one or more sessions of the conference, so that the distinction between "reproducible science" and "non-reproducible observations" was maintained. It was observed in response that this distinction might well result in the "non-reproducible" track completely dominating the program of the conference, and that it might also need to include the majority of the "academic" papers.

Such a change might well weaken one of the key strengths of the SIGIR conference, namely that academic and industry researchers currently rub shoulders for three or more days and exchange information in a collegial manner. This risk would need to be carefully weighed; as a balancing observation it is, of course, likely that any validation of experiments carried out in the first instance on proprietary data would be done on other proprietary data, and that such replication would satisfy the scientific expectation of reproducibility.

A segregationist policy would also require great care when being framed, so that papers reporting user studies (a form of university-based proprietary data) were treated appropriately. Some primary user-study data might be sharable, for example, the screen point-positions generated as part of an eye-tracking experiment. Researchers who generated such data would

have their paper treated differently if they had made the primary data available by the time their work was submitted, or committed to making it available at the time their paper was published. But there might also be experiments undertaken in which the data is fundamentally associated with an individual and hence not capable of being released. An example of the latter would be when users agree to allow their desktop search tool to be instrumented.

Statements of data resources used: The option of requiring all authors to lodge a "Data and Software Declaration" (DASD) as part of their submission came up several times. The DASD would disclose what data and software were used, whether or not either were public, and if they were, how others could access them. In the case of shared data and standard tasks, such as TREC-style effectiveness evaluations, the DASD could also require that authors state previous results that have been attained on the same data, including by the systems that participated in the original blind evaluation task the data was created for.

Putting a level of required disclosure in place would provide referees with a better standard of information than is currently the case, and might encourage authors to think more carefully about their experimentation and baselines. Reviewers could be invited to comment on the level of disclosure provided, via a question added to the review form. The requirement to complete a DASD would not be so prescriptive as to suggest that papers that use only non-shared data should be penalized in any way. But it might be appropriate to allow that papers be summarily rejected if a DASD was not completed, or contained incorrect information. The latter option would then put DASD compliance on a par with the current expectation that suitable steps be taken to anonymize the authorship details. Completion of a DASD might also help focus the intentions of owners of proprietary data, and catalyze future data releases.

Greater emphasis on describing experimental conditions: Shared resources may enable a rapid research cycle, but they are not the only approach to reproducible scientific results. Scientists in fields such as psychology and medicine do not have shared user populations and experimental conditions. Instead, each study describes population features (for example, weight, age, heredity) that the scientist believes may affect outcomes (for example, diabetes before age 40). Other scientists form their own populations with the same (or different) characteristics and test whether they get the same outcomes.

Information retrieval has had the luxury of high-quality shared resources for more than two decades. However these resources have allowed most of the field to avoid serious study of corpus, user, and task characteristics, and these can have significant effects on outcomes. There is anecdotal evidence that different datasets behave differently under some conditions, but little understanding or investigation of what causes those differences.

The field might place greater emphasis on requiring papers to describe much more thoroughly the experimental conditions thought to affect outcomes for research done on both public and proprietary datasets. Over time, more detailed discussion of data characteristics might help reviewers develop intuitions that apply not just to known data, but also to proprietary data.

Renewed community approaches to building and using shared resources: The previous Lemur Toolbar project could be revived, possibly including ground-truth user-based relevance labeling. In terms of software reproducibility, it was suggested by a panelist that SIGIR could promote a library of standard implementations, including shared data samples and evaluation tools. For this to be a viable option, non-trivial long-term funding

would be required. A report from the recent SWIRL 2012 workshop [1, pp 13-14] offered a similar suggestion.

Mandatory use of public resources: A somewhat stricter expectation might be to require that all papers include at least some use of "best available fit" public resources. Authors would be free to note that the effect they had demonstrated showed up only on the non-shared data and was not observed on the shared data they had used; but would not be permitted to ignore shared data that might be regarded by referees and other researchers as being able to provide a reasonable confirmation of the claims being made. Authors would be required to defend "there is no suitable best available fit shared data" statements if they wished to make that case. That is, the default position would be that referees could expect to see at least some experimentation on shared data, even if it failed to show the same effect as was illustrated using proprietary data.

Mandatory use of nothing but public resources: At the extreme end of the spectrum, the community might take the position that use of proprietary data be banned, with papers that make use of any non-sharable data (and presumably non-shared code) resources to be rejected without consideration of their merits. Such a policy would need to be applied without fear or favor to researchers in universities as well as to researchers in industry contexts.

Acknowledgment

We thank the six panelists for making their time available during the panel and through the subsequent email discussion, and for being willing to share their views.

References

[1] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson (eds). Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum*, 46(1):2–32, June 2012.

[2] Alawi A. Alsheikh-Ali, Waqas Qureshi, Mouaz H. Al-Mallah, and John P. A. Ioannidis. Public availability of published research data in high-impact journals. *PLoS ONE*, 6(9), 2011.

[3] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM CIKM International Conference on Information and Knowledge Management*, pages 601–610. ACM, 2008.

[4] Alissa Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Transactions on the Web*, 2(4), October 2008.

[5] Bernardo Huberman. Big data deserve a bigger audience. *Nature*, 482, 2012.

[6] John Markoff. Troves of personal data, forbidden to researchers. *The New York Times*, May 21, 2012.