FireOak Strategies

FireOak Strategies, LLC
215 El Paso Avenue
Durham, NC 27703
United States

https://fireoakstrategies.com
info@fireoakstrategies.com
+1 (919) 408-7766

Insights & Articles

The Dark Side of Open Data

By Eric Hinsdale and Abby Clobridge

About this Article

The Dark Side of Open Data by Eric Hinsdale and Abby Clobridge was originally published in *Online Searcher*, Volume 42, Number 6 – November/December 2018.

All articles from the column, The Open Road, are available on the FireOak Strategies website at https://fireoakstrategies.com/theopenroad.

About FireOak Strategies

FireOak Strategies is a boutique consulting firm specializing in information and knowledge management. We are committed to helping organizations secure and share their information, data, and knowledge.

FireOak Strategies is a certified Women's Business Enterprise (WEB) and Woman-Owned Small Business (WOSB).

WOMEN OWNED™

Certified WBENC Women's Business Enterprise

For more information: https://fireoakstrategies.com.

The Dark Side of Open Data

By Eric Hinsdale and Abby Clobridge

The environment around research data management and open data has become incredibly complex—and the evolution doesn't appear to be slowing down at all. At the core of many of today's challenges are machine learning, natural language processing (NLP), and predictive analytics—the methods used for processing tremendous quantities of data for a variety of intended purposes.

On a daily basis, the news is full of stories from the private sector and government agencies that are mining massive, internally-collected, sets of data for all sorts of outcomes. Technology is making it easier for organizations to become proactive in response to patterns in data. For example, with early alert systems, it is now possible for universities to identify students who might be on the cusp of dropping out early enough for an advisor to intervene. Companies want to mine their customer data to achieve greater profitability and inventory data to forecast demand for products in a timely manner.

But these types of methods aren't restricted to closed data. In fact, one of the advantages of open data is that it allows data from disparate datasets to be combined—re-mixing or merging many "small data" sets to convert them into "big data." From the perspective of funding agencies, this type of re-use is one of the intended benefits of open data. If datasets use common variables, include well-structured and organized data elements, are deposited into interoperable repositories that can be found by harvesters, and include Creative Commons Attribution (CC-BY) licenses (or other similar license allowing for re-use), other researchers are encouraged to find, access, and re-use these datasets without restrictions.

Re-use without restrictions is what sparks fear in many researchers. Once data has been published and is out in the world, you lose all control over your dataset. It can be used, combined, and repurposed in all sorts of ways—including ways you never considered, ways that could potentially put someone else in harm's way, or for more morally-ambiguous purposes.

Although we're proponents of open data, it's useful to know about some incidents where open data has led to problems.

FITNESS TRACKING AND MILITARY INTELLIGENCE

If you or someone in your family is serious about running or cycling, or have a friend who is, you probably know about a wildly popular fitness tracking tool called Strava. Strava is a smartphone app that allows runners and cyclists to track their workouts, including generating GPS maps of workout routes. One feature that makes Strava so popular is that functions as a social network for fitness enthusiasts, allowing them to share data on their training.

On November 1, 2017, Strava announced a new feature—the Global Heatmap (strava.com/heatmap). The heatmap plotted over one billion fitness activities from 10 million users on a world map, creating a map with bright spots where most activity was taking place.

By January 2018 security analysts were taking to Twitter to report an unintended consequence of making all this data freely available. It turns out

many members of the military are devoted Strava users, and so it was possible to locate military installations—even secret facilities—by looking for light regions on Strava's heatmap. The New York Times on January 29, 2018 reported on the phenomenon ("Strava Fitness App Can Reveal Military Sites, Analysts Say", by Richard Pérez-Peña and Matthew Rosenberg).[1]

Once this unintended use came to light Strava responded by making it easier for individuals to opt out of having their data included on the heatmap, as reported by Paul Snyder in Runner's World.[2]

Another fitness application that was exploited for security intelligence was identified by Polar, a Finnish company that also makes fitness tracking devices. In July 2018, Dutch researchers were able to use data freely available through the Polar app, cross-referenced with other data freely available on the internet, to find the name and address of over 6,000 intelligence and military employees located in 69 countries, as reported by Maurits Martijn in de Corresponden.[3] Shortly after announcing its findings, Polar suspended its global activity map, according to Andrew Liptak in The Verge.[4]

ENDANGERING ENDANGERED SPECIES

In July 2015 a ranger in the Knersvlakte Nature Reserve, located in the Western Cape, South Africa, that is the home to hundreds of unique and endangered plant species, stopped a suspicious looking pickup truck. A search revealed a trove of endangered plants, along with materials used to find and collect them and information on a website where they were being offered for sale. Authorities pieced together the information and concluded the poachers were using GPS coordinates harvested from public websites—JSTOR Global Plants[5] and iSpot[6] —to locate rare plants that were collected and sold.

Researchers are just beginning to talk about the issues related to the proliferation of GPS tools now used in conservation biology. In a 2016 paper, published in Conservation Biology, Steven J. Cooke, a biologist at Carleton University in Canada, and his co-authors summed up the problem: "Animal tracking can reveal animal locations (sometimes in nearly real-time), and these data can help people locate, disturb, capture, harm, or kill tagged animals."[7]

REIDENTIFICATION OF ANONYMOUS MEDICAL DATA

Because of the sensitive nature of the data they collect, researchers in the field of medicine have been especially cautious in adopting open data policies. Recently much of the debate has been focused around a policy proposal announced by the International Committee of Medical Journal Editors (ICMJE) that would require making data public as a requirement for publication.

[1] nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html
[2] runnersworld.com/news/a20866081/strava-changes-heat-maps-settings
[3] decorrespondent.nl/8480/this-fitness-app-lets-anyone-find-names-and-addresses-for-thousands-of-soldiers-and-secret-agents/260810880-cc840165
[4] theverge.com/2018/7/8/17546224/polar-flow-smart-fitness-company-privacy-tracking-security
[5] plants.jstor.org
[6] ispotnature.org
[7] fecpl.ca/wp-content/uploads/2016/12/Cooke_et_al-2017-Conservation_Biology-1.pdf

While there are various arguments encouraging caution in data sharing, as explained by Shelley Wood in tctMD[8], there is one that particularly raises red flags. Even when data is anonymized, it may be possible to cross-check a data set with freely available data from other sources—social media accounts, geospatial data, online directories—to "re-identify" data, linking it back to the participants in the clinical trial. (Indeed, such a "re-identification" process was used to pinpoint the identities of the 600+ military and intelligence community members in the Strava example.) Who might be interested in going through the effort of re-identifiying clinical trial subjects? One possibility is the health insurance industry, which would use the information in making decisions about providing coverage to individuals who participated in a medical trial.

In 2017, a group of Australian researchers proved that this type is re-identification of patients through public medical records is indeed possible. The group analyzed a publicly available dataset made up of government billing records and claimed they were able to positively identify a group of prominent Australian citizens and link them to their medical histories, wrote Chris Duckett in ZDNet.[9]

IDENTIFYING SUPPORTERS OF POLITICAL CAUSES

These issues aren't new. In 2008, voters in California went to the polls to vote on Proposition 8, a measure that amended the state constitution to make same-sex marriage illegal. After the measure was approved by voters, a fierce campaign over the future of the law took shape between supporters of the measure and supporters of same-sex marriage.

One group of Proposition 8 opponents created a website called Eightmaps.com as part of the fight. The site consisted of information made public through state campaign finance disclosure laws and overlaid that information onto a map of the state. Anyone who visited the site could find the names, locations, amount donated, and employers of people who donated money to support Proposition 8. After the site was launched, many donors to Proposition 8 began experiencing threats, vandalism, intimidation, and property destruction. The full report from GovLab is on The Global Impact of Open Data website.[10] The site is no longer on the web.

TOXIC DATA

The concept of gathering disparate, publicly-available data sets and combining them to produce information far outside the scope of what was initially intended when the data was made public has become common enough to acquire a name—toxic data. In an article for Forbes[11], written by Dan Woods, one security analyst outlines a hypothetical scenario where publicly available data on travel schedules, aircraft models, and crew staffing could be combined in a way to make someone more vulnerable to a terrorist attack.

VALUE NEUTRALITY OF DATA

As these examples illustrate, data itself is value neutral: It can be put to use for positive purposes or it can be used in more nefarious ways. Most people with a stake in the open data movement realize this, which is why there is so much

[8] tctmd.com/news/open-questions-cardiology-editors-and-academics-mull-data-sharing-pitfalls-and-potential
[9] zdnet.com/article/re-identification-possible-with-australian-de-identified-medicare-and-pbs-open-data
[10] odimpact.org/case-united-states-eightmaps.html
[11] forbes.com/sites/danwoods/2018/04/30/toxic-data-a-new-challenge-for-data-governance-and-security

conversation around governance, data ethics, privacy, and security.

Two characteristics of these examples, however, reveal that discussions need to rise to a new level. First, as the volume of open data increases, the opportunities for combining disparate data sets in unexpected ways to reveal unintended information increase at an astounding rate. A second and related factor is that, just as open data advocates have always argued, open data can contribute to a faster pace of innovation. A combination of human ingenuity, computing power, and a desire to gain competitive advantage will result in data being used in ways it was never imagined when it was initially created.

TAKING A NUANCED APPROACH TO OPEN DATA

While we can't undo events that have already happened, knowing about these incidents can raise awareness of some of the pitfalls of open data, and can help us all take a more thoughtful, balanced approach moving forward.

Many of the researchers we work with have reported procrastinating in publishing their data, unless data must be openly accessible as a condition of getting a journal manuscript accepted for publication. Although researchers often stall in depositing their data into an open repository, once a funding agency asks for it, there's a flurry of activity to get a dataset published, and sometimes steps get rushed.

Instead of waiting until the last minute to publish, encourage researchers to spend some time at the end of a project to think about how this dataset might be combined with others. Consider the impact of what might happen if multiple variables were to be combined to form a new variable. How could a dataset be combined with others? What kinds of patterns might be visible?

Many of these examples involve human subjects. Within the United States, research involving human subjects and meeting specific criteria is covered by 45 CFR 36. Researchers are expected to have an Institutional Review Board (IRB) approve methodologies used in a project before research begins. If a researcher has concerns about how the release of a dataset might impact people, that researcher should raise these concerns with the relevant IRB.

Data doesn't have to be an all-or-nothing proposition-data doesn't have to be "closed" or "open." Researchers can share some variables of a complex dataset without sharing all of them. At a minimum, post metadata about a project without posting the full dataset. This option should be a last resort when research was funded (in part or in whole) by an organization with an open data policy.

It is important to note that many of the funders with open data policies do allow for exceptions, but these exceptions should be thoughtfully submitted and indicate why releasing the data could harm subjects. Not all exceptions will be granted, but, whenever possible, they should be negotiated before a grant agreement has been signed.

Making data openly-accessible requires more thinking and preparation than is often realized. Researchers creating data have an ethical responsibility to consider implications of their work. Open data should involve more preparation than simply going through the mechanics of uploading a file to an open repository.

Knowledge Management & Information Security Consultants

fireoakstrategies.com