Google is capitalizing on a movement that they have contributed nothing to

Short title	Google and research data
Long title	Google is capitalizing on a movement that they have contributed nothing to.
Authors	Peter Kraker [1]
Author affiliation	[1] Founder and chairman of Open Knowledge Maps.
Author bios	*Peter Kraker is founder and chairman of Open Knowledge Maps, a non-profit dedicated to improving the visibility of scientific knowledge for science and society alike. He is a long-time open science advocate – he is a community coordinator of the Open Science Working Group of Open Knowledge Austria and member of the core team of the Open Access Network Austria.*
Author social links	Peter Kraker: Twitter
Date published	27 September 2018
DOI	10.5281/zenodo.1434695
Cite as (APA)	Kraker, P.. (2018). Google is capitalizing on a movement that they have contributed nothing to. *Elephant in the Lab*. DOI: https://doi.org/10.5281/zenodo.1434695

You initiated the #DontLeaveItToGoogle campaign after Google brought out a search engine for scientific datasets. What was the reason to start such an initiative?

Research data is an important scientific output and there are many benefits to research data sharing, including data reuse and aggregation. But discovery is a big problem, even bigger than in literature. In our research, we found that up to 85% of datasets are not reused. In many ways, we

cannot cash the cheques written by the open science movement, when we do not enable adequate discovery of the things we make open.

Therefore, we need better tools for data discovery. But I do not believe that Google Dataset Discovery is the right answer. It represents a proprietary and closed system on top of our own data. This is a system that benefits massively from researchers' labour, but where researchers will have no say in. Google is capitalizing on a movement that they have contributed nothing to. Therefore, we need an open alternative. However, at the moment it seems to me that funders, research administrators and infrastructures are content to leave it to Google. This is highly problematic, especially since we have discussed the problems of lock-in effects and other negative outcomes of proprietary infrastructure for years now.

#DontLeaveItToGoogle is therefore an effort to make people aware of the problems related to Google Dataset Search and to protest the inertia when it comes to funding an open alternative. It's the responsibility of the public and private funders to take initiative and provide an open alternative. We do have the European Open Science Cloud, which is to be released very soon, and data is the main focus point of this cloud, but yet I don't see, how discovery will work it these systems of federated infrastructures. I think the public and the private funders need to step up here and bootstrap such an open alternative.

Where do you see the problem in search engine for data run by Google?

I fear that Dataset Search will go the way of Google Scholar. When Google Scholar came about 15 years ago, it was a ground-breaking literature search engine. The scientific literature, however, has doubled in the meantime, and Google has not made enough investment to keep up with this growth. As a result, Google Scholar is of very limited use today and does a bad job at helping researchers to find relevant papers for their information needs. Now, this lack of innovation would not be a problem, if other tools could build on top of Google Scholar. But unfortunately they can't, because the Google Scholar index is not reusable. Innovators in this market have to first build their own index, which is not helped by the fact that Google has many special arrangements with content providers that the rest of the world does not have.

And we are all poorer for it - discovery is in many ways the departure point of research, and the results of this step decide whether research is reused or duplicated, whether new collaborations are formed or these opportunities are missed. Discovery is therefore important for efficiency, effectiveness and quality of research. Now I do fear that the same could happen for datasets, if we do not put an open alternative out there.

Do you think that public funders are not responsive enough to the demands of the research community?

I don't think that the funders don't recognize that infrastructure is generally needed, but I'm not always sure that they have the right focus. In my opinion, there is a lot of money for backend infrastructure, for creating data stores, high performance computing and that kind of thing, but there is very little money for front ends and services.

It's interesting to consider why this is so. The reason might be to same extent in the history of how scientific applications were created in the past: the interface in many cases was an afterthought. One thing that we've seen with the web and especially starting 2008-2009 is that there were innovative services that also had a very good interface. And this drew a lot of users into digital science and I think that this kind of mindset now also needs to be applied to public and non-profit infrastructures - to really think from the user's perspective, and to make sure that we take into account how things work, because that's the essence of design.

How can the focus be changed?

At the moment it's mostly about awareness - it's necessary to really bring this idea forward and also to make sure that, for example, in a conference on open infrastructures, design aspects and user interface aspects are actually taken into account. Usually, in many of these events these topics are never represented. There are always a lot of conversations about technology and interoperability and all these important things, but almost nobody talks about the interface.

What are truly open research infrastructures and why is it important to have them?

Truly open research infrastructures are those that can be reused. By that I mean the software (open source), the content and the data. They are community-driven and community-owned. In such an ecosystem, innovation thrives, because we can all build on top of each other's work. There are also no lock-in effects that we see with closed offerings - if an organisation does not work out in the way the community expects it to, the community can take it somewhere else. Therefore, truly open infrastructures are the strongest drivers of innovation in scholarly infrastructures today.

Why is it difficult for non-profit research infrastructure providers to compete in the market?

Funding for non-profit infrastructures is scarce. The VC route, meaning taking on venture capital, is not possible for non-profits and therefore, it is difficult to establish sustainable business models. In addition, commercial players such as Elsevier suck all of the money out of the market by offering expensive bundle deals. This makes it difficult for libraries to support open alternatives, even though they usually have much smaller asks than their commercial competitors. There are also free commercial services, but there users pay with their data - a highly problematic business model as we have seen in the Facebook scandal.

You founded Open Knowledge Maps a few years ago, a non-profit tool that helps to visualize scientific literature. Where are the difficulties of of such initiatives?

Well, first of all, let me say that there are many positive things to a non-profit initiative. I love working with a dedicated team of mostly volunteers, who put thousands of hours into Open Knowledge Maps in their free time. An enthusiastic community has formed around Open Knowledge Maps. We have had half a million users in the 2.5 years of our existence. It's great to

hear the many stories of people from around the world, who are now able to get an overview research topics much faster than before and discover new relationships and findings that were previously hidden from them.

But funding is indeed the moot point here. So far, we have come by on a tiny budget that would usually be barely enough for a single person. What I said earlier about scarce funding for nonprofit organisations is especially true for open source services and frontends - but they are the way researchers engage with open science. By leaving this market to proprietary and closed solutions, we are limiting innovation in how researchers - and the rest of the world - interacts with scientific knowledge.

The things we can do on a pure volunteer basis is limited. Therefore my call would be to invest in interfaces and services - this is how we make the open science revolution a reality.