Google AI research scientist announces Dataset Search

September 6 2018, by Nancy Owano

Credit: CC0 Public Domain

Google, from Day One, got big by getting into the business of finding information. Years later, Google is talking serious business about datasets. Google is launching a new search engine to help scientists find the datasets they need.

On Wednesday, Google AI research scientist Natasha Noy announced Google's launch of Dataset Search. You now get easy access to datasets, if you are scientist, or just data "geek" in another type of pursuit, looking for data for your work and for your stories and for your intellectual curiosity.

The goal is to bring you more of a single interface. Jon Fingas in *Engadget* looked at how it can benefit data searching.

"The tool provides more direct access to data presented in an open standard that makes it clear who created the info, how it was collected and how you're allowed to use it. You could not only track down climate data for a report, but make sure that it's relevant and legal to use."

This is a global (as in international) push that works in multiple languages with support for additional languages coming soon. James Vincent in *The Verge* quoted Noy: "I do think in the last several years the number of repositories has exploded."

"Simply enter what you are looking for and we will help guide you to the published dataset on the repository provider's site," she said. Currently, datasets and related data tend to be spread across multiple data repositories and one might find that information about these datasets is neither linked nor indexed by search engines. For the person doing a search, data discovery becomes tedious at best.

They are seriously into support for an ecosystem where providers of datasets themselves are being encouraged, via guidelines that Google developed, to describe their data "in a way that Google (and other search engines) can better understand the content of their pages," she said.

They used the open standard schema.org for their approach on this. On Noy's' wish list: that all data set providers get behind this common standard. It is hoped that more data repositories will use the schema.org standard to describe their datasets. That way, said Noyes, datasets are part of a "robust ecosystem."

"A search tool like this one is only as good as the metadata that data publishers are willing to provide. We hope to see many of you use the open standards to describe your data, enabling our users to find the data that they are looking for."

Jon Fingas in *Engadget*: "It's far from a definitive resource at the moment. It's a start, however, and Google is no doubt hoping that this will encourage others to make their public data more searchable."

And if all this were not enough, Google will be cutting some paths in making the most out of data about data about data.

According to *The Verge*, Jeni Tennison, chief of the Open Data Institute, said ideally Google will publish its own dataset how Dataset Search gets used. She said that Google should publish a dataset about dataset search that would be indexed by Dataset Search, added Vincent. He quoted her:

"Simply understanding how people search is important... what kind of terms they use, how they express them," says Tennison. "If we want to get to grips with how people search for data and make it more accessible, it would be great if Google opened up its own data on this." In other words, he added, Google should publish a dataset about dataset search that would be indexed by Dataset Search.

More information: www.blog.google/products/searc … r-discover-datasets/
toolbox.google.com/datasetsearch

Citation: Google AI research scientist announces Dataset Search (2018, September 6) retrieved 10 April 2024 from https://techxplore.com/news/2018-09-google-ai-scientist-dataset.html