Google Dataset Search Webinar – everything you always wanted to know about Google Dataset Search

Natasha Noy, **Helena Cousijn**, **Mary Hirsch**, **Martin Fenner**

Blog post published December 12, 2018 in *DataCite Blog - DataCite*
https://doi.org/10.5438/4sdj-hf49

Abstract

Following the launch of Google Dataset Search we received many questions from DataCite members about schema.org, indexing by Google Dataset Search, and data citation in general. We wanted to make sure you had all the answers, so we organized a webinar and invited Natasha Noy from Google.

Keywords

Information Systems (Subfield), Google, Schema.org

Recommended citation

Noy, N., Cousijn, H., Hirsch, M., & Fenner, M. (2018, December 12). Google Dataset Search Webinar – everything you always wanted to know about Google Dataset Search. *DataCite Blog - DataCite*. https://doi.org/10.5438/4sdj-hf49

Copyright

Copyright © Natasha Noy et al. 2018. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Following the launch of Google Dataset Search we received many questions from DataCite members about schema.org, indexing by Google Dataset Search, and data citation in general. We wanted to make sure you had all the answers, so we organized a webinar and invited Natasha Noy from Google. Natasha leads the team that developed Dataset Search and during the webinar she was able to tell the people who joined everything they always wanted to know about Google Dataset Search.

For those who were unable to join, we made the recording available through the DataCite YouTube channel and are sharing the questions and answers from the webinar in this blog post. If reading this leads to even more questions, don't hesitate to reach out to us via support@datacite.org.

Questions and Answers

Schema.org Markup

Do you have an exemplary case of schema.org markup for a data source that has covered many of the things to look out for when indexing sources?

There is one fake example in the developers documentation. However, any page that you see in Google Dataset Search search is an example. Copy the link from a result that you like and paste it into the Google Structured Data Testing Tool, and you can see what kind of metadata it has and the source of the page and what is imported in in terms of metadata. You can use any tool (not just the Google tool) to see the metadata and the source and what it is translated into in terms of structured data.

There are many related entities (from schema.org, or hopefully to be added) that also would benefit by adding to search. Are there plans to create search for other scientific resource, like Linux containers (for which sometimes the line between Dataset and Container can be a bit gray!).

We're mostly working on data and datasets. But we are looking at other artifacts. Our notion of a dataset is quite inclusive. There are discussions at schema.org about describing containers in more detail but there are also cases where it would be reasonable to use "Dataset" markup for them. Also, some disciplines never publish structured data, they only publish figures and tables in papers. One thing we're looking at is extracting the metadata about the figures and tables if you ask if there is data that matches Ph levels to wave lengths in silicon, we want to find a way to point to that. At the moment we are not looking at workflows and containers.

Will there be an API exposed for developers to use for validation and query? Clarification: Validation API, as in an endpoint to send *JSON-LD* and validate it. like the web testing tool, but one that can be done programmatically (not by a human). If I generated a JSON-LD for my resource, I would want to validate it. I want to be able to do it with a testing tool (in continuous integration) If I have 100 datasets and I need to validate, I don't want to manually copy paste 100 links. I want my automated tests to pass.

Regarding the API, if there was a plan it would be more generic for structured data, validating one page in the testing tool probably gives an idea of how it will work in other pages, there might be some use cases, someone else could do this, it's an open standard.

How do you determine metadata quality?

We are still figuring it out and experimenting with different signals, from how the metadata corresponds to what is actually on the page to how much metadata is present given the type of the dataset, etc.

Indexing

Can you give us more information on how indexing works?

- *how frequently to expect updates*
- *why some datasets that have valid JSON-LD according to the Google Structured Data Testing Tool are not indexed (how to debug in this case)*
- *generally how to find out whether we've generated valid JSON-LD more immediately*

Once the dataset page is crawled (or re-crawled), it will be updated in Google Dataset Search within a few days. We (the Dataset Search team) do not control the crawl part however, and that one is subject to whatever all Google crawlers do. Your Search Console is a good place for that information. To debug, first check Structured Data Testing Tool; then make sure through the Search Console that it was crawled. If it still doesn't appear after a few days, use the Feedback button in Dataset Search to file a bug report.

Is there a size limit to the web pages that go into the index? (e.g. the structured data testing tool gives up > 2.5Mb).

This is a more general Search question. Maybe a product forum is a good place to ask.

Are you handling "spam", fake pages just implementing schema.org with dataset metadata as a fake to put spam into the search?

There are a number of quality measures that we implement, including trying to make sure that the metadata corresponds to the information on the page itself.

Data cleaning would result in a new version of the data file itself, so how do you approach file versioning? There must be a cache, so how do you handle updates and synchronization? (When the provider updates the original file)

Sitemaps help because, in sitemaps, you can see how often things get updated. Things do get re-crawled. We are downstream from crawl, once things get re-crawled we get the update within 2-3 days.

How do you currently handle "similar" datasets? We often have hundreds of datasets with very similar metadata, just slight differences in metadata (different location/platform). When searching in Google Dataset Search it mostly shows one random result, but all the similar ones are hidden.

This is not handled very well right now. It needs more thought. Any web search tries to show a diversity in results, the solution will be to group datasets and display the cluster.

If we have only metadata records in our repository (but not the actual datasets), would those records be indexed ? And would it have an impact on the search rankings?

Metadata records will be indexed. There are so many signals that go into ranking that it is hard to identify how a single signal will affect it.

Are there plans to allow search on summaries of the data (e.g., limit search to datasets with a minimum number of rows/sample size, data with a high mean age)? How can Dataset Search learn about means and variables labels?

Means and variables could be in the metadata, there could be options to include this, including things like biases in the data. If the provider wants to publish certain stats then that could be part of the metadata. One option could be to bring this up on the schema.org GitHub issue tracker. For those that have pushed for more detailed additional properties, there's often a push back to keep things simple. This is because schema.org is designed to be easy to implement. Anything more complicated belongs in an extension, e.g BioSchemas.org.

A lot of human subject datasets are not public (for scientific use only), but the metadata is (or could be). When you start using information inside the data (variables names, labels, value labels, means) those datasets would fall under the table – unless variable metadata and data summaries become part of the metadata somehow. Any plans regarding this? I know variableMeasured is in pending, but what about data summaries?

The data does not have to be public. The metadata has to be public and crawlable. Providers would want their robots.txt file to provide information to crawlers, or to require a login in order to prevent any private data to appear in search. This is not specific to datasets, the issue is the same

with Google Search, if something is public that shouldn't be it can be found in the main search, but we have strict rules about taking this down as soon as it is reported by the provider.

All "our" datasets found in Google Dataset Search have the DataCite logo, link to search.datacite.org, "provided by Datacite", etc.? Will adding structured data markup to our own dataset landing pages change that?

Sometimes you randomly show datasets using DataCite metadata instead of native ones from the datacenter. Would it be good to have some ranking on your side, like first show the local metadata and only use DataCite as fallback?

Yes, that's the case, adding the markup will solve that. DataCite has a markup for all it's pages so you can always find them. If the primary source doesn't have the markup it won't be found, if you add the metadata it will be linked. We are working on figuring out how best to identify the primary and secondary source of a dataset to make the display more helpful to users.

As a data provider can we have a repository level profile? Not just the dataset creator/author.

That sounds like a good idea. We will consider that as we consider author/creator profiles.

Are datasets published as sparql endpoints marked up with the DCAT ontology be included in the dataset search?

Yes.

How does the dataset have to be linked in the metadata to make indexing more likely in the future?

Not sure what the question refers to. If the issue is about indexing the content of the dataset, then it needs to be linked and robots.txt should allow its crawling,

Also, we have multiple detailed distributions in our metadata for the dataset, but they only show up as available formats. Neither the links nor the names are picked up (specified via "type": "DataDownload"). Is it planned to include these metadata, or should we solve this by creating more datasets and assign them to a DataCatalog?

For the moment, we made a conscious decision not to provide direct download links so that users download the data from the original site, having all the necessary context to understand the data. It is still important to include the download links as they are part of the metadata, for Google and for other tools.

Search

What is allowed in the markdown of the description? Will you render figures (e.g., distributions of variables)? Are there plans to use logo provided in the metadata instead of the logo from the Google Knowledge Graph? The Google Knowledge Graph does for some reason not provide our logo.

There is no policy at the moment. However, I would suggest using links instead of figures. We are adding the answer on how to add your logo to the results to the FAQ.

Can Google Dataset Search do a dataset preview of its spreadsheet data since Google Search's can display spreadsheet data in search results?

Yes, that's among the features that we want to implement.

As a demo of how things could work, could you point to examples of a process where someone has discovered data w/ Dataset Search, then manipulated in Data Studio, then shared visualizations, for example?

We don't have such an example, but if someone does, and can post it on a blog somewhere, we'd love to have our users see it!

Are there any plans to show more metadata for a dataset? For example, we provide a citation for the dataset which is not shown. Since the paper must be cited when the data is used, it would be very nice to have it shown in the Dataset Search as well. Everything but the abstract seems hidden.

Yes, we do plan to extend the metadata that we are showing. We show more than just the abstract, but certainly not everything that we could be showing yet.

Do you show variables?

A few things we need to show are not shown yet.

Data Citation

How are data citations counted in Google Scholar?

Essentially we do a query for the DOI, URL, and full name of the dataset, and then try to account for special cases where the name of the dataset is generic like "public schools", because that's not very useful. We plan to start looking at the context in which data is mentioned and how people cite data. Citing data is more complicated than citing publications. We want to go beyond "is this mentioned".

Does this mean you are going into a direction of "Google Scholar for Scientific Data"

We are providing discovery: in the same way you can go into Google Scholar and find a publication on any publishers' website, you will eventually be able to go into Dataset Search and find any dataset on providers' website. Another big part of Google Scholar is the cross citations and links and for the specific links between datasets we're hoping to incorporate that as much as possible. We have talked about, further down the line, author profiles. Academics like Google Scholar because they can find all their publications and see who cites them, we are thinking in that direction for Dataset Search, but not in the short term.

Are there are plans to include data citations and their corresponding citation counts into Google Scholar Profiles?

We are in contact with the Google Scholar team discussing how and when it would make sense to cross reference in both directions. The products are clearly related and we are discussing how to provide information for all users.

The post Google Dataset Search Webinar – everything you always wanted to know about Google Dataset Search appeared first on DataCite.