

Google Dataset Search by the Numbers

Omar Benjelloun, Shiyu Chen, and Natasha Noy

Google Research, Google, USA
{benjello, shiyuc, noy}@google.com

Abstract. Scientists, governments, and companies increasingly publish datasets on the Web. Google’s Dataset Search extracts dataset metadata—expressed using schema.org and similar vocabularies—from Web pages in order to make datasets discoverable. Since we started the work on Dataset Search in 2016, the number of datasets described in schema.org has grown from about 500K to almost 30M. Thus, this corpus has become a valuable snapshot of data on the Web. To the best of our knowledge, this corpus is the largest and most diverse of its kind. We analyze this corpus and discuss where the datasets originate from, what topics they cover, which form they take, and what people searching for datasets are interested in. Based on this analysis, we identify gaps and possible future work to help make data more discoverable.

1 Dataset Search as a snapshot of datasets on the Web

We live in a data-driven world. Scientists, governments, journalists, commercial companies, and many others publish millions of datasets online. There are thousands of Web sites that publish datasets—some publish a handful, some publish hundreds of thousands [3]. Google’s Dataset Search¹ is a search engine for datasets on the Web [16]. It relies on schema.org and similar open standards to extract the semantics of dataset metadata and to make it searchable.

Arguably, the mere existence of Dataset Search and its reliance on semantic markup provided a strong incentive for dataset providers to add such markup to their Web pages. Indeed, we have seen an explosive growth of dataset metadata on the Web since we started the work on Dataset Search. In the Fall of 2016, there were about 500K Web pages that included `schema.org/Dataset` markup, with half of them coming from `data.gov`, the US Open Government portal [10]. Today, there are tens of millions of such pages, from thousands of sites.

A recent comprehensive survey highlights a variety of approaches to help users find datasets [3]: these approaches range from searching within a collection of tables with different schemas [14], to finding data in repositories, such as Figshare, Zenodo, or DataDryad, to using metadata search engines, like Dataset Search. There are a number of well respected directories of dataset publishers (e.g., DataCite [19], re3data [12], Scientific Data in Nature [15]), but they inevitably miss new datasets or repositories [4]. To the best of our knowledge, Dataset Search is the only collection of dataset metadata that includes all semantically annotated datasets on the Web.

¹ <http://datasetsearch.research.google.com>

We chose to rely primarily on `schema.org` for describing dataset metadata because both search engines and open-source tools have used it successfully to build an open ecosystem for various types of content [8]. In recent years, the scientific community has also embraced it for publishing data, by creating mappings from other metadata standards to `schema.org`. For example, Sansone and colleagues define a mapping from the DATS standard in the biomedical community to `schema.org` [20]. Wang and colleagues use `schema.org` to describe research-graph data, comprised of researchers, datasets and scholarly articles [24]. Efforts such as `bioschemas.org` [6] extend `schema.org` to include domain-specific terminology and relationships.

In this paper, we analyze the Dataset Search corpus of metadata. As of March 2020, the corpus contained 28 million datasets from more than 3,700 sites. While limited to the dataset metadata that is available in `schema.org` or DCAT, this corpus contains a sizable snapshot of the datasets on the Web. And because many researchers and scientists rely on search engines to find datasets [7], learning from this corpus can inform both the work to improve search engines for datasets and, more important, highlight the gaps in representation and coverage for the community at large. Specifically, in this paper, we make the following contributions:

- We present methods for analyzing an organically created corpus of metadata for 28 million datasets on the Web (Section 2).
- We identify a set of research questions that such a corpus can help analyze and present results of the analysis of the corpus (Section 3).
- We discuss lessons learned from the corpus analysis (Section 4).

2 Data collection methods

In this section, we describe the methods that we used to collect the metadata and to prepare it for the analysis in Section 3. In the remainder of this paper, we abbreviate the `schema.org` namespace as `so#` and the DCAT namespace as `dct#`.

2.1 From `schema.org` and DCAT on the Web to the corpus

We described the details of the Dataset Search architecture elsewhere [16]. In brief, Dataset Search relies on the Google Web crawl to find pages that contain dataset metadata and to extract the corresponding triples. A post-processing of the Web crawl data parses RDFa, Microdata, and JSON-LD into a common graph data model, broadly equivalent to W3C’s RDF triples [18]. We keep `so#Dataset`, `dct#Dataset`, and all the related entities and their properties.

We enhance, normalize, and augment this corpus in a variety of ways in order to provide users with a meaningful search experience. In this section, we focus only on those processing steps that are relevant to the subsequent data analysis. The processing happens at multiple levels of granularity: At the corpus level, we ensure that datasets are unique and attempt to remove non-datasets (i.e., pages

that include dataset markup, but do not describe datasets). At the dataset level, we augment the metadata with inferred properties. Finally, at the property level, we clean up and normalize values.

The **corpus-level analysis** starts by removing duplicates within each site [16]. We found that many dataset repositories add markup both to the dataset landing pages and to the pages that list search results within that repository. We keep only the former in the corpus through simple heuristics: When the same dataset (according to values of key properties) appears on multiple pages, we keep the page that contains only one dataset. We also remove dataset metadata that does not have values for basic properties such as title and description [5].

At the **dataset level**, we process the values for properties such as title and description as well as the terms on the Web page itself in order to identify the main topics covered by the dataset. We use the topics from re3data.org [12] and a similar set of topics from the Google Knowledge Graph [17] as our vocabulary.

In addition, our **page-level** analysis collects information from the Web page that the dataset originated from, such as the domain of the page and its language.

For **individual properties**, we normalize, clean, and reconcile values for:

- **Data downloads and formats:** We identify the patterns that data providers use to represent download information and normalize them to a single representation [16]. Providers may specify file formats through the `so#fileFormat` or `so#encodingType` properties. When both of these properties are missing, we extract a file extension from the data-download URL.
- **DOIs and compact identifiers:** Persistent citable identifiers, such as Digital Object Identifiers (DOIs) and Compact Identifiers [26], may appear in several properties, such as `so#identifier`, `so#url`, or even `so#sameAs`, and `so#alternateName`. We use regular expressions to find patterns that correspond to these identifiers, and look for known prefixes from identifiers.org in all of these properties.
- **Level of access:** Two properties define the level of access of a dataset: `so#isAccessibleForFree` is a boolean value that indicates whether or not the dataset requires a payment. `so#license` links to a license or specifies one inline. We normalize the license information into known classes of licenses, such as Creative Commons and open government licenses. Any license that allows redistribution essentially makes the dataset available for free. We count datasets with these licenses as well as datasets with `so#isAccessibleForFree` set to true as the datasets that are “open.”
- **Providers:** There is some ambiguity in schema.org on how to specify the the source of a dataset. We use the `so#publisher` and `so#provider` properties to identify the organization that provided the dataset. As with other properties, the value may be a string or an `Organization` object. Wherever possible, we reconcile the organization to the corresponding entity in the Google Knowledge Graph.
- **Updated date:** There are several date properties associated with a dataset: `so#dateCreated`, `so#datePublished`, `so#dateModified` (and similar properties in DCAT). There is little consistency in how dataset publishers distinguish between them. However, the most recent value across these dates

is usually a reliable approximation on when a dataset was last updated. We use several parsers to understand dates expressed in common formats and to normalize them. If there is no date in the metadata, we use the date when the Web page itself was last updated as a proxy.

Finally, to analyze the usage of datasets in the Dataset Search application, we look at logs for two weeks in May 2020. We extract the identifiers of the datasets that appeared in search results, and join them with their metadata to analyze search behavior in aggregate.

All the data and analyses in this paper are based on a snapshot of the Dataset Search corpus from March 26, 2020. We also compare the status of the corpus with a version from ten months prior, in June 2019.

2.2 Limitations of the analysis

While we believe that our corpus is a reasonably representative snapshot of the datasets published on the Web, we recognize that it has limitations. Indeed, we have no way of measuring how well the corpus covers all the datasets available on the Web.

First, the corpus contains only the datasets that have semantic descriptions of their metadata in schema.org or DCAT. If a dataset page does not have metadata in a machine-readable format and in a vocabulary that we recognize, it will not be in our corpus (and will not be discoverable in Dataset Search).

Second, if a dataset page is not accessible to the Google crawler or is not being crawled for some reason (e.g., because of `robots.txt` restrictions), it will not be in our corpus. When the crawler processes the page, it often needs to execute JavaScript to get the metadata. If a page is slow to render, we may not obtain dataset metadata from it.

Third, our methods for inferring new values, such as dataset topics, may be imprecise, and we have not formally evaluated their accuracy yet.

Fourth, in our analysis, we “trust” what the metadata says. For instance, if a dataset’s metadata says that the dataset is accessible for free, we count it among the open datasets. In some cases, the reality may be different when users try to download the dataset.

Finally, a significant amount of pages on the Web are designed for Search Engine Optimization or are simply spam. A page may have `so#Dataset` on it but not actually describe any dataset metadata. While we do our best to weed out such pages, our techniques are not perfect, and we cannot be certain that all the datasets in the corpus that we describe are indeed datasets.

3 Results and Observations

We start with the results of a corpus-level analysis (Section 3.1), then look at specific metadata properties (Section 3.2) and finally present our observations on the usage of datasets in Dataset Search (Section 3.3).

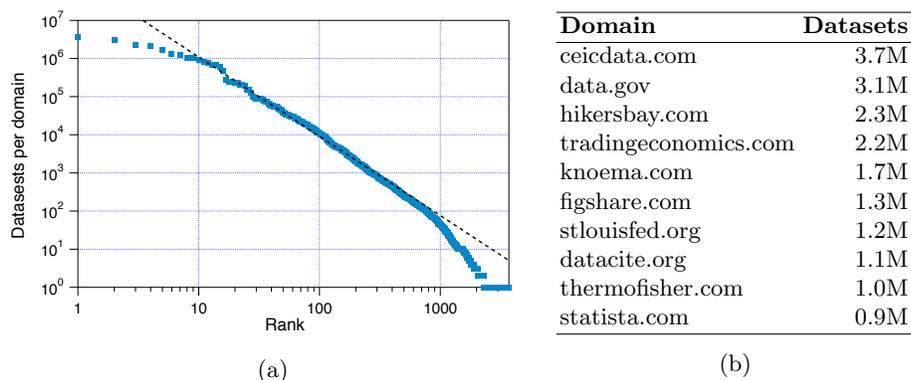


Fig. 1: Number of datasets per domain: (a) A log-log plot of domains and their relative sizes in terms of the number of datasets. The dotted line shows a fit to power-law for most of the range with the coefficient of 2.08 ± 0.01 , which is very close to quadratic fit. (b) Domains with the largest number of datasets. These ten domains are responsible for 65% of the datasets in the corpus.

3.1 Corpus-level Analysis

Looking at the corpus as a whole, as well as characteristics of the Web pages that we extracted metadata from, enables us to answer the following questions.²

Datasets and domains: How many datasets does each domain have? What is the distribution of the number of datasets by domain? What are the domains contributing the largest number of datasets? What are the most common top-level domains with datasets, and what fraction of datasets do they contribute? We know that many of the datasets come from open-government initiatives across the world. But just how many?

Dynamics of the corpus: How has the corpus of dataset metadata grown over time? Pages with datasets inevitably go offline, get moved, change their URLs. Can we trace this churn in our corpus? How prevalent is it? These numbers give a sense of how metadata on the Web is changing.

Metadata on metadata: Which fraction of datasets use schema.org vs DCAT? Which metadata fields are frequently populated, and which ones rarely have any values? These numbers give us probably the most actionable items in terms of improving metadata quality.

Datasets and domains The snapshot that we analyze in the rest of this section, taken on March 26, 2020, contains 28M datasets from 3,700 domains. The number of datasets per domain mostly follows a power law distribution, as the logarithmic scale plot in Figure 1a shows: A small number of domains publish

² Here and elsewhere, “domain” refers to “internet domain.”

Table 1: Number of datasets for the top twenty top-level Internet domains.

Top-level domain	Number of datasets	Top-level domain	Number of datasets
.com	14,956K	.ru	243K
.org	4,696K	.co	218K
.gov	3.386K	.nl	181K
.at	819K	.au	160K
.net	760K	.pl	152K
.es	524K	.uk	144K
.de	366K	.ca	139K
.edu	293K	.io	79K
.fr	281K	.world	56K
.eu	263K	.info	54K

millions or hundreds of thousands of datasets, while the long tail of domains hosts just a handful of datasets. The two domains with the largest number of datasets (`ceicdata.com` and `data.gov`) have more than 3 million datasets each. The ten largest domains (Figure 1b) account for 65% of all datasets.

While “typical” Web pages about datasets correspond to a single dataset, some pages may have multiple datasets on them. For instance, a page may describe a large dataset and break down its components as multiple datasets; or a page may be dynamically generated in response to a search in a dataset repository. In our corpus, we found that over 90% of datasets come from pages that contain exactly one dataset. Still, more than 1.6M datasets come from pages with more than ten datasets.

Table 1 shows the distribution of datasets by top-level internet domains. The vast majority of the datasets come from `.com` domains, but both `.org` and government domains are well represented. For the country-specific domains, Austria, Spain, Germany, and France are at the top of the list. If we combine all government domains across the world (`.gov`, `.gouv.*`, `.gv.*`, `.gov.*`, `.gob.*`, etc.), we find 3.7M datasets on these government domains.

To get a more complete picture of the international coverage of datasets, Table 2 breaks them down by language, as specified by or extracted from the Web pages that contain them. More than 18M datasets, or 64% are in English, followed by datasets in Chinese and Spanish, both of which are growing faster than datasets in English. Note that these numbers do not capture the nuance of specific `schema.org` property values using multiple languages.

Dynamics of the corpus The next question that we study is the change in the corpus over time. Figure 2a shows the growth in the number of datasets since the beta launch of Dataset Search, in September 2018. The number of datasets has grown steadily, from about 6M then to 28M in March, 2020. We have reported earlier [16] that, day-to-day, about 3% of the datasets are deleted from our index while 7-10% new datasets are added. Enterprise data repositories

Table 2: Number of datasets by language and the % change between June 2019 and March 2020.

Language	Number of datasets	% increase
English	18,650K	67%
Chinese	1,851K	82%
Spanish	1,485K	70%
German	743K	74%
French	492K	76%
Arabic	435K	75%
Japanese	404K	72%
Russian	354K	65%
Portuguese	304K	69%
Hindi	288K	70%

have a similarly large level of churn [9]. Figure 2b shows the results of comparing the URLs between snapshots from June 2019 and March 2020, when the corpus almost doubled in size: Only 8.8M of the 14M URLs in the June 2019 corpus, or 63%, are still there in March 2020. The other 5.4M are no longer at the same location—or may no longer be in the corpus at all. This dynamic indicates a very high level of churn.

Figure 3 captures the updates to individual datasets. 14M datasets, or 50%, have a value for at least one of the date properties in metadata (Section 2). For an additional 10.7M datasets, we were able to determine the date when the Web page was modified. Out of these 24.7M datasets with a known date of last update, 21M datasets, or 85%, have this date within the last 5 years (Figure 3). The short-term distribution (Figure 3a) shows that more datasets were last updated within the last month than in any other month in the past year. Looking at the long-term distribution (Figure 3b), 49% of datasets were last updated within the past year.

Metadata on metadata While schema.org is our primary semantic vocabulary for dataset description, we also understand and map basic DCAT properties. However, we found that fewer than 1% of datasets use the DCAT vocabulary.

Table 3 shows which fraction of datasets have values for specific properties. We require datasets to have a title and description [5]; hence, their 100% coverage in our corpus. Because we normalize and reconcile values from different properties, the properties in Table 3 do not always directly correspond to schema.org or DCAT predicates. For instance, we combine `so#publisher` and `so#creator` into “provider” because we observed that data owners do not really distinguish between the different semantics.

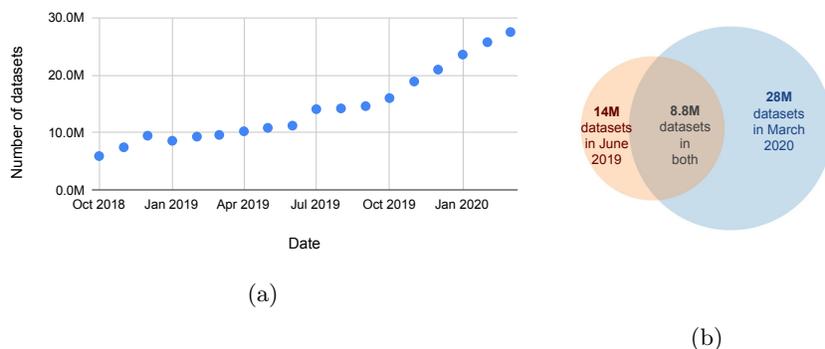


Fig. 2: Corpus dynamics: (a) Growth of datasets since the beta launch in September 2018. The number of datasets grew from 6M to 28M (b) Changed URLs in the corpus between June 2019 and March 2020: 8.8M URLs of the 14M in June 2019 remained in the corpus. The rest have either disappeared or changed.

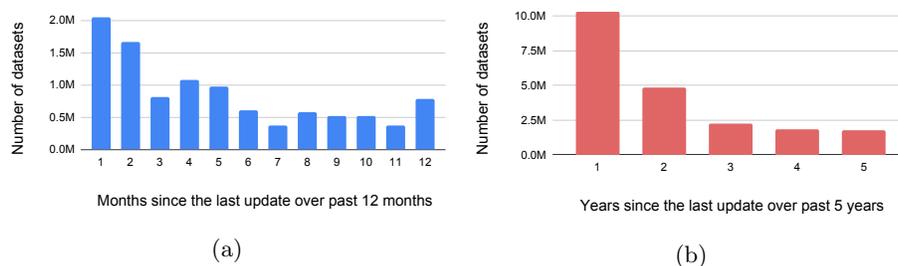


Fig. 3: Distribution of the date when a dataset was last updated: (a) at monthly granularity over the past year; (b) at yearly granularity over the last five years. Note that we have this information only for 85% of datasets.

3.2 Inside the metadata

We focus on the properties that we found most informative in understanding the corpus—or that we got asked about most often (e.g., what are the formats for the available datasets). We do not try to provide an exhaustive description of the ranges of values for every property.

Topics Figure 4 shows the distribution of the topics that datasets are associated with. We generate the topics automatically, based on dataset titles and descriptions as well as the text on the page where the dataset came from. The two largest topics are geosciences and social sciences—the areas that we focused on specifically before the beta launch in September 2018.

Table 3: Percentage of datasets with specific properties. Column 2 lists the source predicates for each property. Properties not listed in the table have values in fewer than 1% of the datasets.

Property	Source predicates	Percentage
description	so#description, purl#description	100.00%
title	so#name, purl#title	100.00%
provider	so#publisher, so#provider, purl#publisher	84.59%
keywords	so#keywords, dct#keyword, purl#keyword	80.08%
URL	so#url, dct#accessurl, dct#landigpage	68.30%
temporal coverage	so#temporalCoverage, so#temporal, purl#temporal	45.41%
data download	so#distribution, dct#distribution	44.34%
spatial coverage	so#spatialCoverage, so#spatial, purl#spatial	38.69%
date modified	so#dateModified, purl#modified	37.46%
license	so#license and so#license on so#distribution	34.80%
date published	so#datePublished, purl#published	30.83%
catalog	so#includedInCatalog	29.74%
variable	so#variableMeasured, dct#theme	20.90%
authors	so#author, so#creator	14.12%
same_as	so#sameAs, rdf#same_as	12.72%
date created	so#dateCreated	9.62%
alternate name	so#alternateName, rdf-schema#label	3.40%
is accessible for free	so#isAccessibleForFree	3.04%

Data downloads Most users who search for datasets ultimately want the data itself and not just its metadata. Datasets can specify means to download their data via the so#distribution property.

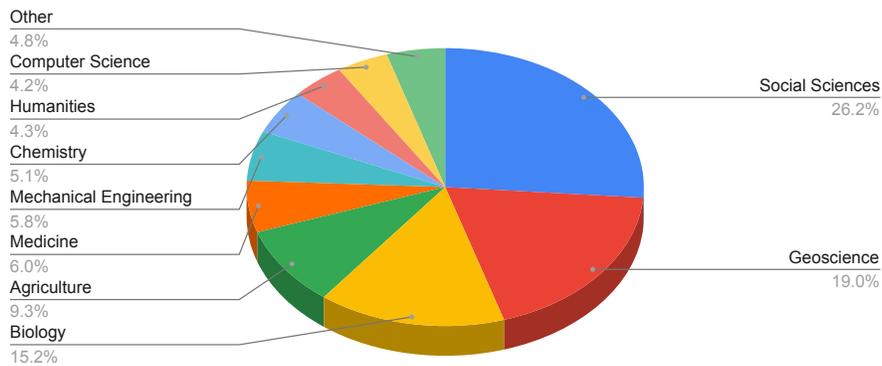


Fig. 4: Distribution of datasets by broad coverage topic, inferred from dataset metadata and the Web page itself.

Table 4: Number of datasets by content type. The counts are based on `so#fileFormat` or `so#encodingType` properties and file extensions. Note that some datasets have multiple distribution formats. Therefore, the total number of entries here is larger than the 12M datasets with data downloads.

Category	Number of datasets	% of total	Sample formats
Tables	7,822K	37%	CSV, XLS
Structured	6,312K	30%	JSON, XML, OWL, RDF
Documents	2,277K	11%	PDF, DOC, HTML
Images	1,027K	5%	JPEG, PNG, TIFF
Archives	659K	3%	ZIP, TAR, RAR
Text	623K	3%	TXT, ASCII
Geospatial	376K	2%	SHP, GEOJSON, KML
Computational biology	110K	<1%	SBML, BIOPAX2, SBGN
Audio	27K	<1%	WAV, MP3, OGG
Video	9K	<1%	AVI, MPG
Presentations	7K	<1%	PPTX
Medical imaging	4K	<1%	NII, DCM
Other categories	2,245K	11%	

Availability: Only 44% of datasets specify a data download link in their metadata (Table 3). Looking at the origin of datasets with downloads, we found that 85% of them are provided by just 10 domains.

Data types and formats: Zooming in on the subset of datasets that specify a data download, what are the broad categories of content and their relative prevalence? To answer this question, we first extract the file format of data downloads, and then bucket them into categories. For bucketing, we defined a high-level classification inspired by Elsevier DataSearch,³ and created a mapping from the file formats found in the data to the target categories (Table 4).

Tables in CSV or XLS formats are the most common type of data (37%), followed by structured formats such as JSON and XML (30%) and documents in PDF or DOC format (11%). The latter category is problematic for many applications, as it is not machine readable. Audio, video, and medical imaging formats all constitute less than 1% of the datasets.

A subset of datasets that is of interest to the Semantic Web community is the datasets that contain graph data. We can approximate this number by summing over common formats: owl, rdf, xml+rdf, sparql, and so on. Together, they represent only 0.54% of datasets with downloads. Semantic web data is largely under-represented among datasets that use Semantic Web methods to describe metadata.

³ datasearch.elsevier.com/

Table 5: Datasets with DOIs and compact identifiers

(a) Top ten providers of datasets with DOIs.		(b) Providers with more than 100 datasets with compact identifiers.	
Domain	Datasets with DOIs	Domain	Datasets with compact identifiers
figshare.com	1,300,745	neurovault.org	73,869
datacite.org	1,070,066	alliancegenome.org	29,204
narcis.nl	118,210	datacite.org	14,982
openaire.eu	109,149	openaire.eu	4,262
datadiscoverystudio.org	72,063	scicrunch.org	1,522
osti.gov	62,923	mcw.edu	517
zenodo.org	49,622	duke.edu	306
researchgate.net	41,494		
da-ra.de	39,318		

Making data citable As we hope that datasets themselves become first-class citizens of the scientific discourse, we must develop mechanisms to reference and cite them. Scientists commonly use digital object identifiers (DOIs) and compact identifiers (provided by services such as identifiers.org) for this purpose. We extract DOIs and compact identifiers from the dataset URLs or the values of the `so#url` property, as well as from `so#sameAs` and `so#identifier` properties. About 11% of the datasets in the corpus (or ~3M) have DOIs; about 2.3M of those come from two sites, datacite.org and figshare.com (Table 5a). Only a tiny fraction, 0.45% of the datasets, have compact identifiers (Table 5b).

Data Providers While internet domains provide the conduit that brings datasets to users, the semantic provenance of datasets is more accurately captured by the notion of “provider.” Domains and providers often align, but they also may differ when a provider hosts their datasets on a platform different from their own Web site. About 84% of all datasets specify a provider and there are about 100k distinct data providers in the corpus. The top 3 providers are CEICdata.com, Knoema, and the U.S. Geological Survey. The top 20 providers account for 78% of the total datasets. Most of them are the hosts of the top domains in Figure 1b. However, 87% of the providers are “small” providers, who publish fewer than 10 datasets each.

How open is the data Finally, we analyze the licenses and availability of datasets. Dataset Search does not require the data to be open; only the metadata must be accessible to the crawler. Publishers specify access requirements for a dataset via the `so#license` property, the `so#isAccessibleForFree` boolean property, or both. About a third of the datasets (34%) specify license information and 3% of the datasets have the value for `so#isAccessibleForFree` (Table 3). Of the datasets that specify a license, we were able to recognize a known license

in 72% of the cases. Those licenses include Open Government licenses for the US and Canada, Creative Commons licenses, and several Public Domain licenses. We found that for 89.5% of these datasets either the `so#isAccessibleForFree` bit is set to true or their license is a license that allows redistribution, or both. In other words, almost 90% of these datasets are available for free. And of these open datasets, 5.6M, or 91%, allow commercial reuse.

3.3 What do users search for

Finally, we look at what Dataset Search users search for. Overall, 2.1M unique datasets from 2.6K domains appeared in the top 100 search results over 14 days in May 2020. Figure 5 shows the topics for these datasets. Note that the distribution of topics in Figure 5 is different from the one for the corpus as a whole (Figure 4), with geosciences, for instance, taking up a much smaller fraction; conversely, biology and medicine take a larger fraction relative to their share of the corpus. We are writing this paper during the Coronavirus pandemic; this timing likely explains the increased share of the biology and medicine datasets.

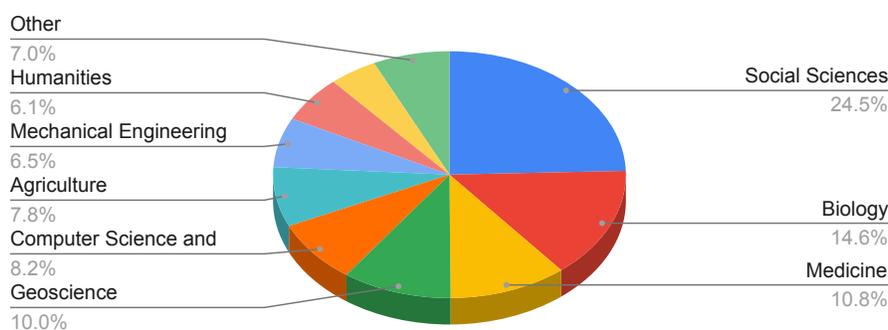


Fig. 5: Topic distribution of datasets that appeared in search results over 14 days in May 2020.

4 Discussion

We start our discussion by highlighting the results we found surprising or counter-intuitive. We then focus on the results that point to future work around improving the quality of metadata in general and provenance information in particular.

4.1 What surprised us in the data

We do not attempt to discuss every table and graph in Section 3. Rather, we focus on the results that require some explanation or discussion.

Licenses and access: Only 34% of the datasets provide any licensing information, through the `so#license` property for dataset or distribution. When no license is specified, the user technically must assume that they cannot reuse the data. Thus, adding licensing information, and, ideally, adding as open a license as possible, will greatly improve the reusability of the data [2]. At the same time, we were encouraged to see that in the vast majority of datasets that specify a license were available for free, allowed redistribution under certain conditions, and almost always allowed reuse for both commercial and non-commercial purposes. While 34% is still relatively low, this number is significantly higher than the 9% of linked-data datasets with licenses that Schmachtenberg and colleagues found in their 2014 survey [21].

Availability of data downloads: We found that only 44% of datasets specify a data download link in their metadata. This number is surprisingly low, because datasets are merely containers for data. A possible explanation is that Webmasters (or dataset-hosting platforms) fear that exposing the data-download link through Schema.org metadata may lead search engines or other applications to give their users direct access to downloading the data, thus “stealing” traffic from their Web site. Another concern may be that data needs the proper context to be used appropriately (e.g., methodology, footnotes, license information), and providers feel that only their Web pages can give the complete picture. In Dataset Search, we made the decision not to show download links as part of dataset metadata so that users can get the full context from the publisher’s Web site before downloading the data.

Quality and completeness of the metadata: As Table 3 shows, the majority of metadata properties have under 50% coverage. While properties such as spatial or temporal coverage apply only to specific domains, others such as authors, variables being measured, updated date, licensing and download information are general, and need to become much more prevalent for metadata to be truly descriptive. Meusel and colleagues analyzed schema.org adoption across the Web in 2015 [13] and found that data providers both adopt and drive changes in the schema.

Effect of outreach: We discussed elsewhere [16] that reaching out to specific communities and finding influencers there was key to bootstrapping the corpus in the first place. At the time, we focused on geosciences and social sciences. Since then, we have allowed the corpus to grow organically. We were surprised to see that, even after the corpus has grown manifold, the communities that we reached out to early on are still dominating the corpus: 45.2% of the datasets are from these two disciplines. Of course, this dominance may be due to other factors, such as differences in culture across communities. For instance, geosciences have been particularly successful in making their data FAIR [22].

Persistent identifiers and URLs: Many scientific disciplines have come to a consensus (or have been compelled by funding agencies and academic publishers) that it is important to publish data and to cite it [25]. There are, for example, peer-reviewed journals dedicated to publishing valuable datasets, such as Nature Scientific Data [15], and efforts such as DataCite [19], that provide

digital object identifiers (DOIs) for datasets and both encourage and enable scientists to publish their datasets. For datasets to become first-class citizens in scientific discourse, they must be citable. Unfortunately, Figure 2b shows that URLs of datasets are not persistent: 37% of URLs that had dataset metadata in June 2019 either do not have the metadata or are no longer accessible in March 2020. This high level of churn argues strongly for the use of persistent identifiers for datasets, such as DOIs and compact identifiers. This practice is now widespread for publications, and we argue that it should become just as widespread for datasets.

(Not) eating our own dogfood. Fewer than 1% of datasets in our corpus are in linked-data formats. The Dataset Search approach relies on semantic-web technologies such as DCAT and schema.org. At the same time, the Semantic Web community is either not producing enough data, not sharing it, or not adding semantic metadata to it. From profiling efforts (e.g., [1,13]), it seems that the problem is the latter: there is plenty of shared data that researchers produce, but the final step of describing it appears to be less common.

4.2 Future work

Improving metadata quality Throughout this analysis, we found many places where metadata was missing, formatted wrongly, not normalized, and so on. We discuss a few possible approaches to improve the quality of metadata:

Automated techniques: We continue to develop better techniques to automatically clean, normalize, and reconcile dataset metadata. These techniques are hard to do at scale given the heterogeneity of the data, and they will never be perfect. The benefits of these techniques are often application-specific, and there are no easy mechanisms to share them back with the community.

Feedback to publishers: We can let the owners of datasets know that their dataset metadata can be improved. Google tools such as Search Console and the Structured Data Testing tool already highlight some of these issues. One could also consider developing interactive tools to create and validate dataset metadata, which could be integrated into popular dataset management CMSs or hosting platforms.

Crowdsourcing: Why not let users of datasets fix their metadata or point to possible improvements? One option would be to provide that functionality in the Dataset Search tool, which would then funnel suggestions to the publishers. Vrandečić [23] proposed using WikiData to crowdsource the definition of metadata for popular datasets that the community cares about.

Improving provenance information Many datasets are duplicates of other datasets, or derived from other datasets. Knowing these relationships is important not just for Dataset Search, but to any user of datasets who cares about data provenance. These relationships are also critical to giving dataset publishers credit when their data is reused: derivative datasets are akin to citations of papers. While schema.org provides properties to describe these relationships,

namely `so#sameAs` and `so#isBasedOn`, the usage of these properties is low. How can we improve the coverage of this lineage information?

Automated detection: We already detect duplicate datasets and cluster them [16].

Identifying that a dataset is derived from other datasets is a much more difficult problem, but it may be feasible in restricted cases (e.g., specific forms of data, such as tables or images, and limited transformation operations).

License requirements: Most data licenses require citation when re-using a dataset, however there is no obligation to make that citation machine readable. There would be huge value in requiring the usage of `schema.org` properties when citing datasets.

Community incentives: Data provenance can help complete the picture of the usefulness and impact of datasets, together with paper citations, application usage, etc. How can we incentivize broad adoption of data provenance in the scientific and open data communities [11]?

5 Conclusions

In this paper, we analyzed the corpus of dataset metadata used in Google’s Dataset Search, a search engine over datasets on the Web. While it has limitations (Section 2.2), it is a large snapshot of datasets on the Web in a variety of disciplines. Our analysis shows that datasets on the Web are very diverse, with no one discipline truly dominating; there are datasets with semantic markup in Web sites from any country and in any language. We have observed an explosive growth over the last three years.

Yet, metadata still leaves a lot to be desired if data is truly to become a first-class citizen in scientific discourse: We need tools to ensure that the metadata is more complete and mechanisms to encourage the use of licensing information for data and persistent identifiers. And the Semantic Web community needs to eat its own dogfood by adding semantic metadata to its datasets.

References

1. Ben Ellefi, M., Bellahsene, Z., Breslin, J.G., Demidova, E., Dietze, S., Szymański, J., Todorov, K.: RDF dataset profiling—a survey of features, methods, vocabularies and applications. *Semantic Web* **9**(5), 677–705 (2018)
2. Carbon, S., Champieux, R., McMurry, J.A., Winfree, L., Wyatt, L.R., Haendel, M.A.: An analysis and metric of reusable data licensing practices for biomedical resources. *PLOS ONE* **14**(3) (2019). <https://doi.org/10.1371/journal.pone.0213090>
3. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. *The VLDB Journal* **29**(1), 251–272 (2020)
4. Fenner, M., Crosas, M., Grethe, J., et al.: A data citation roadmap for scholarly data repositories. *bioRxiv* (2017). <https://doi.org/10.1101/097196>
5. Datasets: Search for developers. <https://developers.google.com/search/docs/data-types/dataset>
6. Gray, A.J., Goble, C.A., Jimenez, R.: Bioschemas: From potato salad to protein annotation. In: *Intl Semantic Web Conf. (Posters, Demos & Industry Tracks)* (2017)

7. Gregory, K., Groth, P., Scharnhorst, A., Wyatt, S.: Lost or found? discovering data needed for research. *Harvard Data Science Review* (4 2020). <https://doi.org/10.1162/99608f92.e38165eb>
8. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the web. *Communications of the ACM* **59**(2), 44–51 (2016)
9. Halevy, A., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang, S.E.: Goods: Organizing Google’s datasets. In: *ACM SIGMOD* (2016)
10. Hendler, J., Holm, J., Musialek, C., Thomas, G.: US Government Linked Open Data: Semantic.data.gov. *IEEE Intelligent Systems* **27**(3), 2531 (May 2012). <https://doi.org/10.1109/MIS.2012.27>
11. Herschel, M., Diestelkämper, R., Lahmar, H.B.: A survey on provenance: What for? what form? what from? *The VLDB Journal* **26**(6), 881–906 (2017)
12. Kindling, M., van de Sandt, S., Rücknagel, J., Schirnbacher, P., Pampel, H., Vierkant, P., Bertelmann, R., Kloska, G., Scholze, F., Witt, M.: The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* **23**(3/4) (2017). <https://doi.org/10.1045/march2017-kindling>
13. Meusel, R., Bizer, C., Paulheim, H.: A web-scale study of the adoption and evolution of the schema.org vocabulary over time. In: *Intl Conf on Web Intelligence, Mining and Semantics*. ACM, New York, NY, USA (2015). <https://doi.org/10.1145/2797115.2797124>
14. Nargesian, F., Zhu, E., Pu, K.Q., Miller, R.J.: Table union search on open data. *VLDB Journal* **11**(7) (Mar 2018). <https://doi.org/10.14778/3192965.3192973>
15. Nature scientific data. <https://www.nature.com/sdata> (2018)
16. Noy, N., Burgess, M., Brickley, D.: Google Dataset Search: Building a search engine for datasets in an open web ecosystem. In: *The Web Conference*. p. 13651375. ACM (2019). <https://doi.org/10.1145/3308558.3313685>
17. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons and challenges. *Commun. ACM* **62**(8), 3643 (Jul 2019). <https://doi.org/10.1145/3331166>
18. RDF 1.1 Concepts and Abstract Syntax. <https://www.w3.org/TR/rdf11-concepts/>
19. Rueda, L., Fenner, M., Cruse, P.: Datacite: Lessons learned on persistent identifiers for research data. *IJDC* **11**(2), 39–47 (2016). <https://doi.org/10.2218/ijdc.v11i2.421>
20. Sansone, S.A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J.S., Xu, H., Fore, I.M., Lyle, J., Gururaj, A.E., Chen, X., et al.: DATS, the data tag suite to enable discoverability of datasets. *Scientific data* **4**, 170059 (2017)
21. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: *Intl Semantic Web Conf (ISWC)*. pp. 245–260. Springer (2014)
22. Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., Wyborn, L.: Make scientific data FAIR (2019)
23. Vrandečić, D.: Describing datasets in Wikidata. In: *Advanced Knowledge Technologies for Science in a FAIR World, IEEE eScience Conference* (2019)
24. Wang, J., Aryani, A., Wyborn, L., Evans, B.: Providing research graph data in JSON-LD Using Schema.org. In: *26th Intl Conf on World Wide Web Companion*. pp. 1213–1218 (2017). <https://doi.org/10.1145/3041021.3053052>
25. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR guiding principles for scientific data management and stewardship. *Scientific data* **3** (2016)
26. Wimalaratne, S.M., Juty, N., Kunze, J., Janée, G., et al.: Uniform resolution of compact identifiers for biomedical data. *Scientific data* **5**, 180029 (2018)