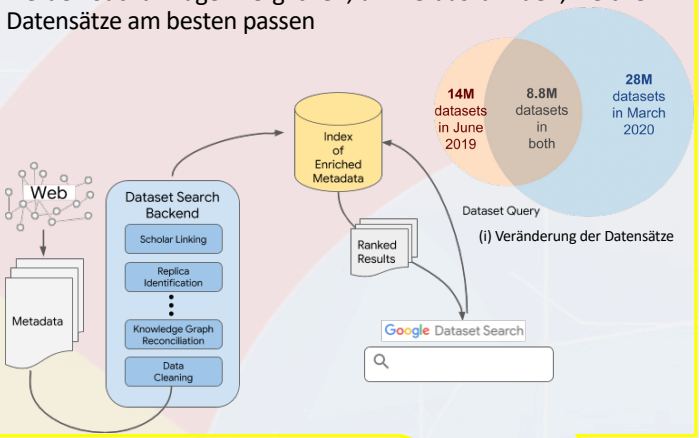
GOOGLE DATASET SEARCH BY NUMBERS

Universität Hamburg | Semantic Systems Julian Dillmann julian.dillmann@studium.uni-hamburg.de

Funktionsweise: Google Dataset Search

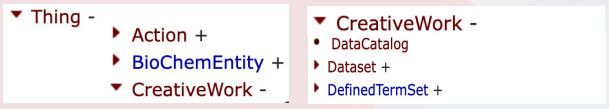
- Google Dataset Search beinhaltet mehr als 31 Millionen Datensätze
- vereinfacht die Suche nach Datenpools, die in Tausenden von Repositorien im gesamten WWW gehostet sind
- die Suchmaschine soll Datensätze allgemein für WissenschaftlerIn zugänglich und zitierbar machen („Google Scholar for Data“).
- Google nutzt schema.org/Dataset mark-up, um Datensätze zu strukturieren und Metadaten für Webcrawler lesbar zu machen
- mit den Metadaten wird der Korpus indexiert und danach werden Suchanfragen verglichen, um herauszufinden, welche Datensätze am besten passen



(i) Veränderung der Datensätze

Schema.org: Die Organisation des Datenpools

- Schema.org ist eine Auszeichnungssprache für die Gliederung und Formatierung von Daten auf Webseiten
- unterstützt die Datenformate RDFa, Microdata und JSON-LD
- Datensätze werden für Maschinen lesbar und die Semantik wird einheitlich, somit können Webcrawler die Datensätze durchsuchen
- besteht aus einer Reihe von "Typen", die jeweils mit einer Reihe von „Eigenschaften“ verbunden sind
- es existieren derzeit 792 Typen und 1447 Eigenschaften
  - Pfad Schema.org/Dataset:



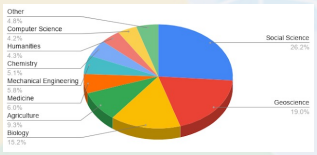
Ausgewählte Eigenschaften von Schema.org/Dataset:

Dataset
A Schema.org Type
Thing > CreativeWork > Dataset
[more...]
A body of structured information describing some topic(s) of interest.

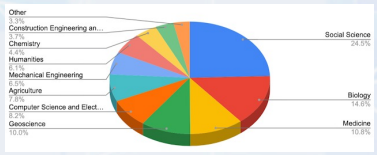
Property	Expected Type	Description
Properties from Dataset		
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format.
includedInDataCatalog	DataCatalog	A data catalog which contains this dataset. Supersedes catalog, includedDataCatalog. Inverse property: dataset
issn	Text	The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different formats of, or the linking ISSN (ISSN-L) for, this serial publication.
description	Text	A description of the item.
name	Text	The name of the item.
provider	Organization or Person	The service provider, service operator, or service performer; the goods producer. Another party (a seller) may offer those services or goods on behalf of the provider. A provider may also serve as the seller. Supersedes carrier.
url	URL	URL of the item.

Dataset Search: Wie die Suche nach Datensätzen vereinfacht wird

Aktueller Stand und Probleme der Suchmaschine



(1) Anteil der Datensätze an den jeweiligen Themen



(2) Anteil der Suchanfragen von Datensätzen an den jeweiligen Themen

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%

(3) Prozentanteil der Datensätze mit spezifischen Eigenschaften gespeichert in den Metadaten

Probleme:

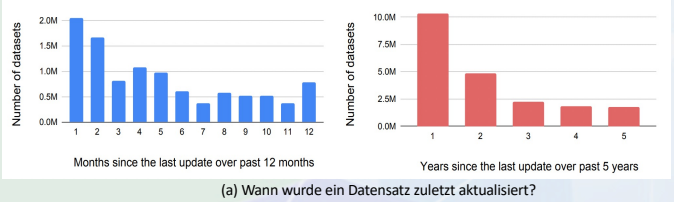
- nur 34 % der Datensätze haben Informationen über die Lizenz (3)
- nur 44 % stellen einen Download für die Daten zur Verfügung (3)
- einige Datensätze haben zu vielen Metadateneigenschaften keine Informationen (3)
- weniger als 1 % der Datensätze in dem Korpus liegen in verknüpften Datenformaten (z.B. RDF) vor

Wie man Dataset Search in der Zukunft verbessern kann

- Dataset Search hat einen sehr dynamischen Korpus (s. Grafik i)
- der Korpus repräsentiert jedoch nicht die Gesamtheit der Datenpools im Web

Ziele:

- Verbesserung der Technik zu automatisierten Bereinigung, Normalisierung und Abstimmung von Metadaten (in schema.org)
- die Provider sollen Nachrichten erhalten, wenn ihre Metadaten in schema.org unvollständig sind (s. Problem in Tabelle 3)
- Crowdsourcing: Die Nutzer von Datensätzen korrigieren oder weisen selbst auf die fehlenden Metadaten hin
- Anreize schaffen, dass Provider Lizenzinformationen veröffentlichen



(a) Wann wurde ein Datensatz zuletzt aktualisiert?

Referenzen:
Benjelloun, O., Chen, S., Noy, N.: „Google dataset search by the numbers.“; International Semantic Web Conference. Springer, Cham, 2020.
Noy, N., Benjelloun, O.: „An Analysis of Online Datasets Using Dataset Search“; Google AI Blog, 2020.
Burgess, M., Noy, N.: „Building Google Dataset Search and Fostering an Open Data Ecosystem“; Google AI Blog, 2018.