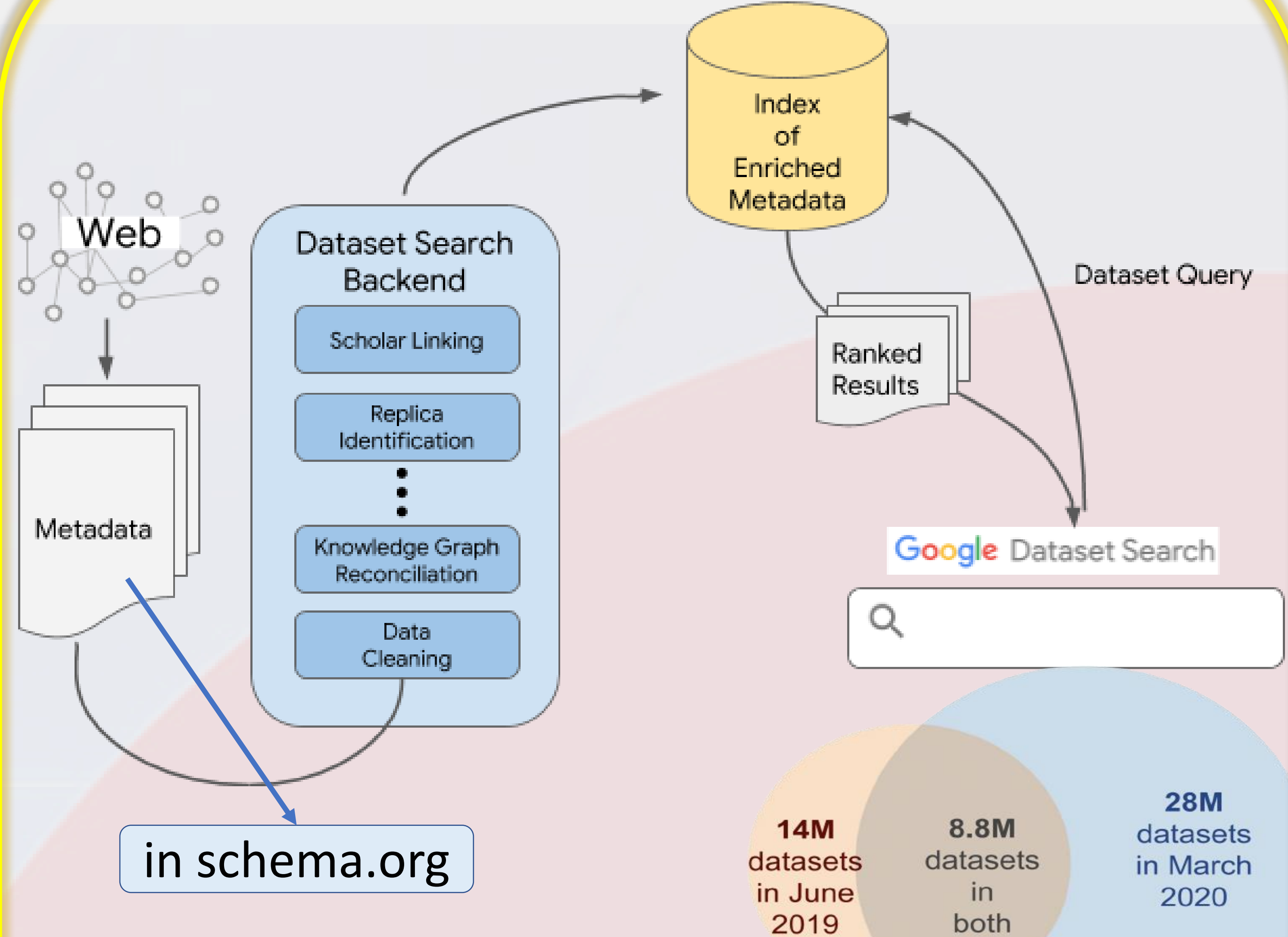


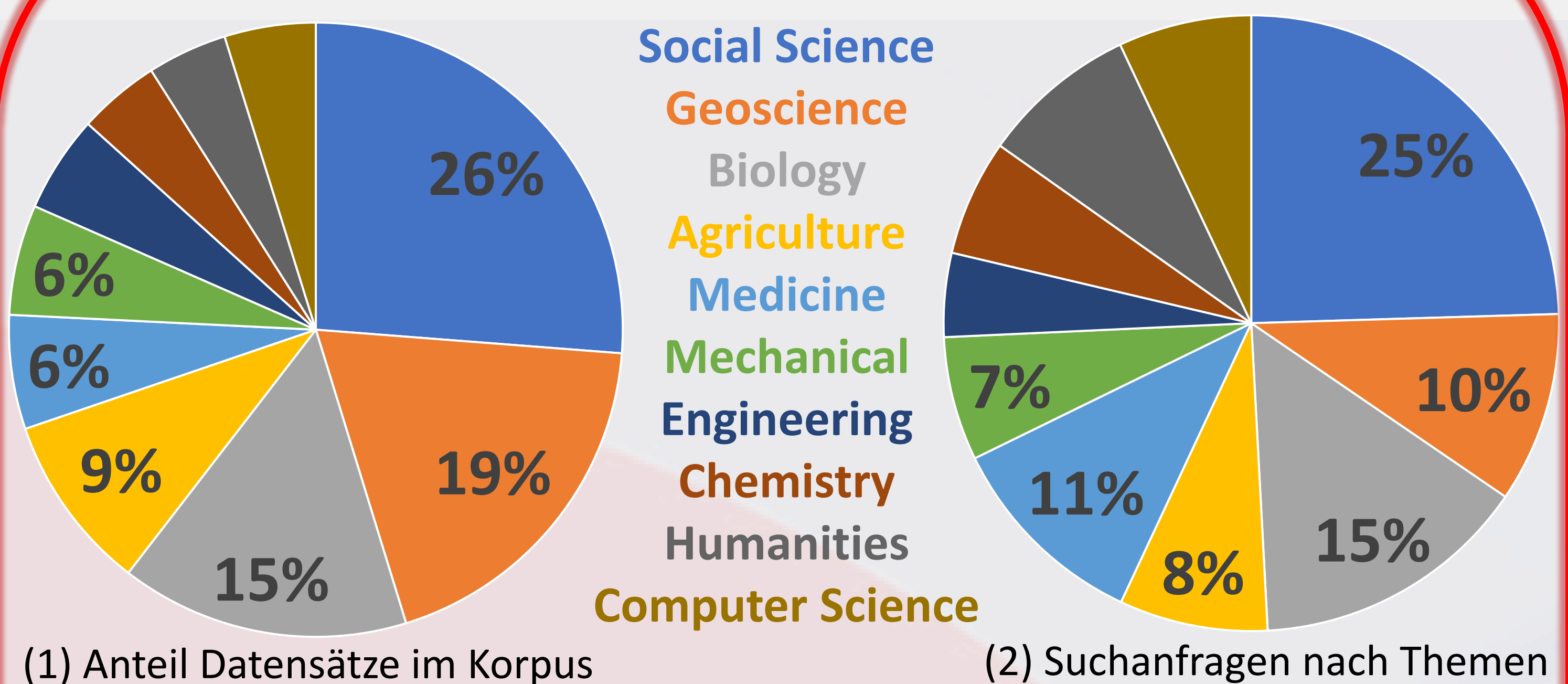
GOOGLE DATASET SEARCH BY NUMBERS

Funktionsweise: Google Dataset Search



- mehr als 31 Millionen Datensätze
- vereinfacht die Suche nach Datenpools
- in Tausenden von Repositorien im gesamten Web
- „Google Scholar for Data“
- nutzt schema.org/Dataset mark-up
 - Ziel: Datensätze für WissenschaftlerIn zugänglich und zitierbar machen

Aktueller Stand und Probleme der Suchmaschine



(1) Anteil Datensätze im Korpus

| 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% |

(3) Prozentanteil der Datensätze mit spezifischen Eigenschaften in schema.org

Probleme

- nur 34 % der Datensätze haben Lizenzinformationen
- nur 44 % stellen einen Datendownload zur Verfügung
- weniger als 1 % der Datensätze in dem Korpus liegen in verknüpften Datenformaten (z.B. RDF) vor

Dataset Search: Wie die Suche nach Datensätzen vereinfacht wird

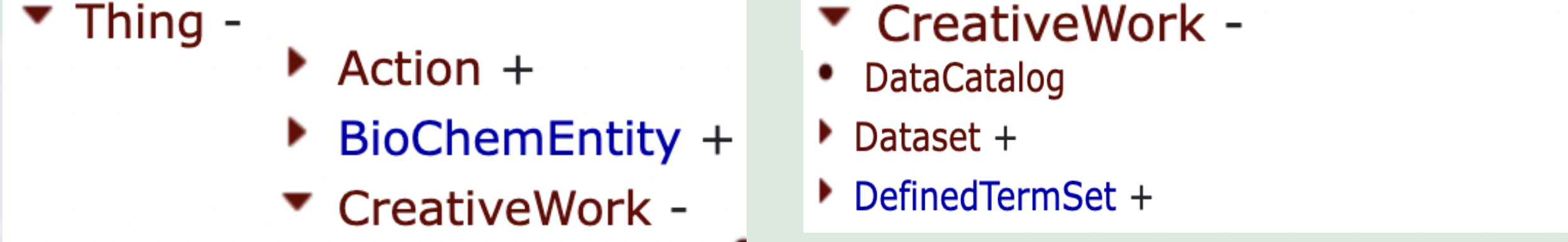
Schema.org: Die Organisation des Datenpools

Ausgewählte Eigenschaften von Schema.org/Dataset:

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

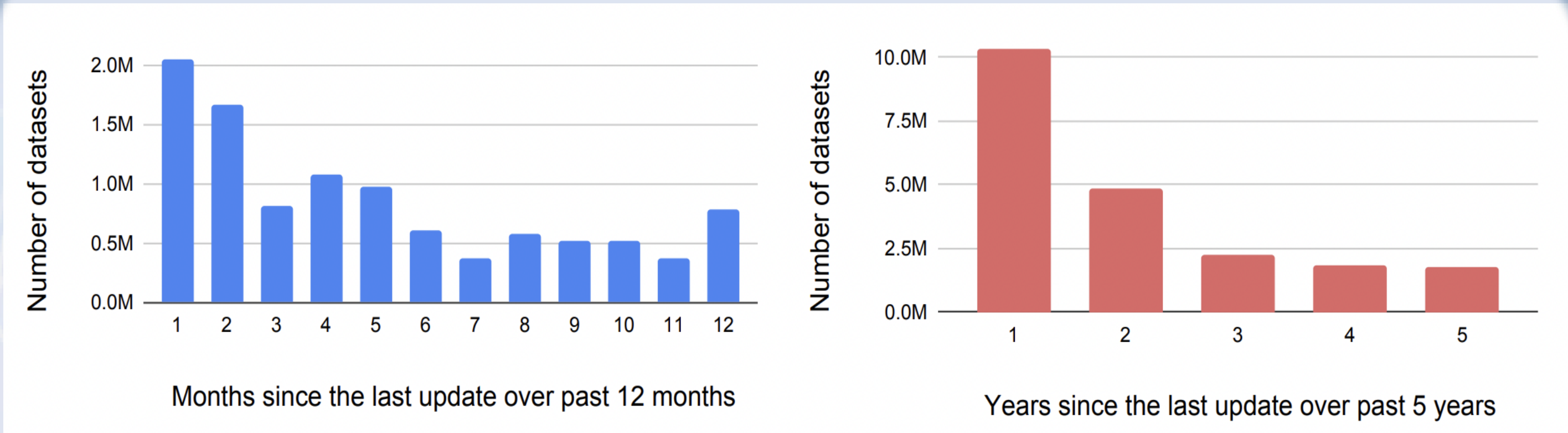
| Property | Expected Type | Description |
|--------------------------------|------------------------|--|
| Properties from Dataset | | |
| distribution | DataDownload | A downloadable form of this dataset, at a specific location, in a specific format. |
| includedInDataCatalog | DataCatalog | A data 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. |

Pfad Schema.org/Dataset:



- Auszeichnungssprache für Gliederung und Formatierung von Daten auf Webseiten
- unterstützt RDFa, Microdata und JSON-LD
- Semantik wird einheitlich und strukturiert
- Webcrawler durchsuchen Datensätze
 - aktuell existieren 792 Typen und 1447 Eigenschaften

Verbesserungen für Google Dataset Search



(a) Wann wurde ein Datensatz zuletzt aktualisiert?

- sehr dynamischer Korpus (s. Grafik a)
- repräsentiert nicht den gesamten Datenpool im Web

Ziele

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

