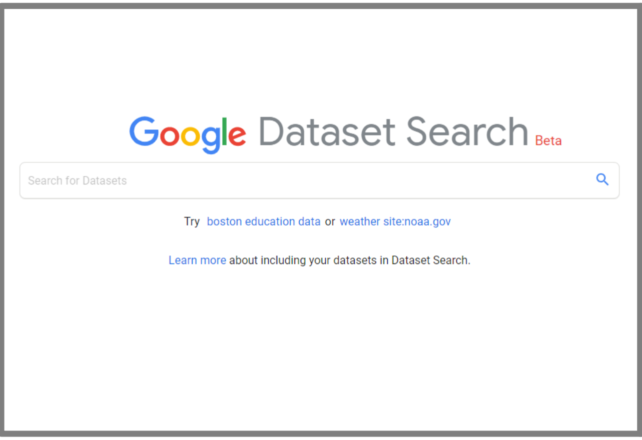
Towards Multi-Facet Snippets for Dataset Search

Xiaxia Wang[1], Gong Cheng[1], Evgeny Kharlamov[2]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Bosch Center for Artificial Intelligence, Renningen, Germany

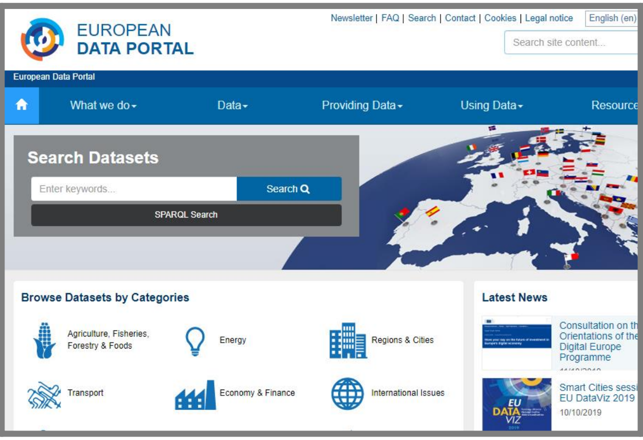
Dataset search systems: Conveniently find relevant datasets.



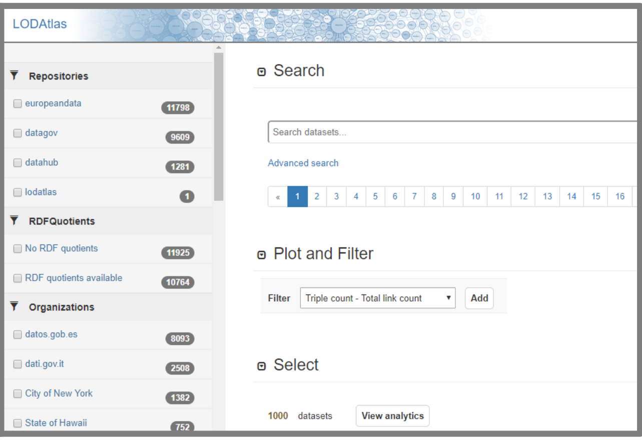
Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.

Google Dataset Search

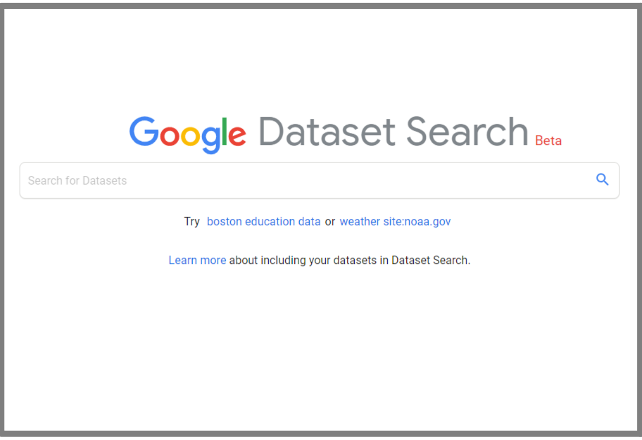


European Data portal



LODAtlas

Dataset search systems: Conveniently find relevant datasets.

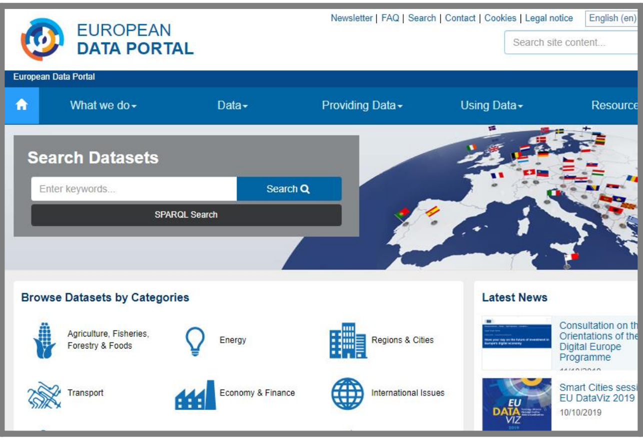


Google Dataset Search Beta

Search for Datasets

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.



EUROPEAN DATA PORTAL

Newsletter | FAQ | Search | Contact | Cookies | Legal notice | English (en)

Search site content...

European Data Portal

What we do ▾ Data ▾ Providing Data ▾ Using Data ▾ Resource

Search Datasets

Enter keywords... Search

SPARQL Search

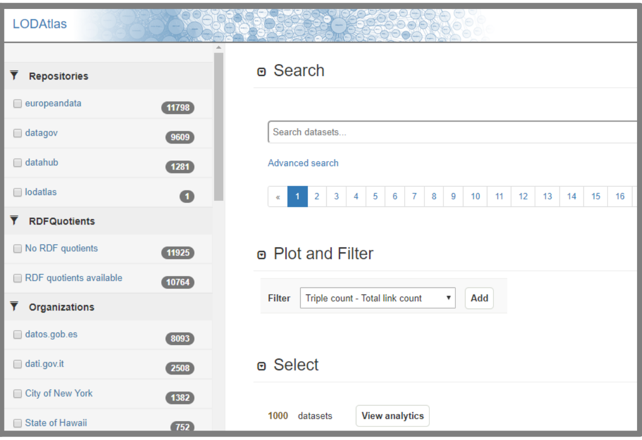
Browse Datasets by Categories

- Agriculture, Fisheries, Forestry & Foods
- Energy
- Regions & Cities
- Transport
- Economy & Finance
- International Issues

Latest News

Consultation on th Orientations of the Digital Europe Programme

Smart Cities sessi EU DataViz 2019
10/10/2019



LODAtlas

Repositories

□ europeandata	11798
□ datagov	9609
□ datahub	1281
□ lodatlas	1

RDFQuotients

□ No RDF quotients	11925
□ RDF quotients available	10764

Organizations

□ datos.gob.es	8093
□ dati.gov.it	2508
□ City of New York	1382
□ State of Hawaii	752

Search

Search datasets...

Advanced search

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Plot and Filter

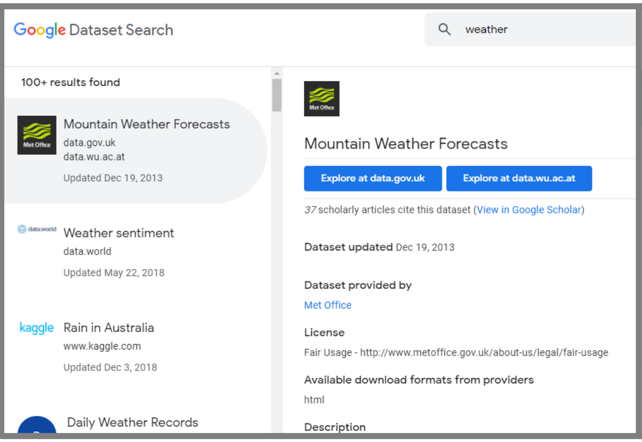
Filter: Triple count - Total link count ▾ Add

Select

1000 datasets View analytics

Google Dataset Search *European Data portal* *LODAtlas*

Existing systems provide only metadata for users.



Google Dataset Search

weather

100+ results found

Mountain Weather Forecasts
data.gov.uk
data.wu.ac.at
Updated Dec 19, 2013

Weather sentiment
data.world
Updated May 22, 2018

Rain in Australia
www.kaggle.com
Updated Dec 3, 2018

Daily Weather Records

Mountain Weather Forecasts

Explore at data.gov.uk Explore at data.wu.ac.at

37 scholarly articles cite this dataset (View in Google Scholar)

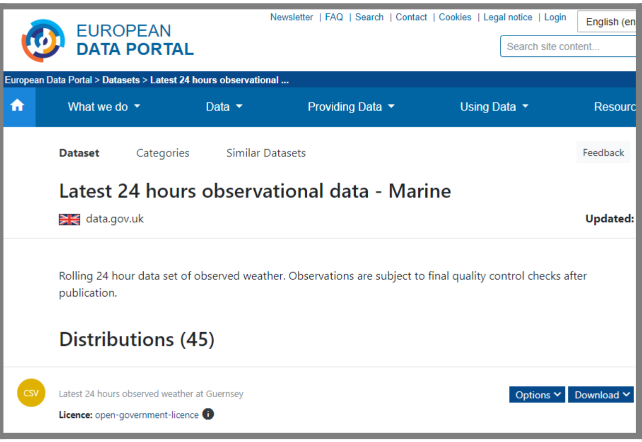
Dataset updated Dec 19, 2013

Dataset provided by
Met Office

License
Fair Usage - http://www.metoffice.gov.uk/about-us/legal/fair-usage

Available download formats from providers
html

Description



EUROPEAN DATA PORTAL

Newsletter | FAQ | Search | Contact | Cookies | Legal notice | Login | English (en)

Search site content...

European Data Portal > Datasets > Latest 24 hours observational ...

What we do ▾ Data ▾ Providing Data ▾ Using Data ▾ Resourc

Dataset Categories Similar Datasets Feedback

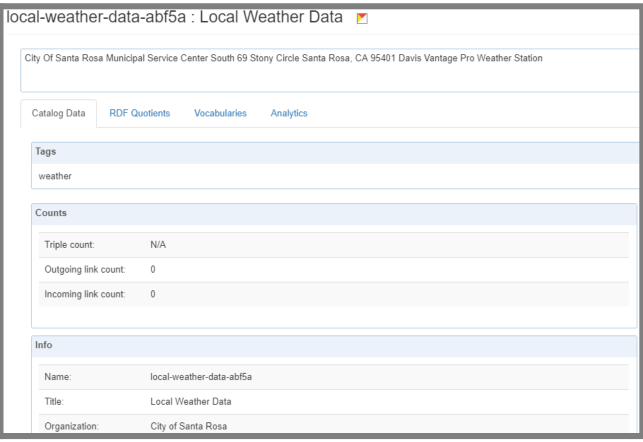
Latest 24 hours observational data - Marine

data.gov.uk Updated:

Rolling 24 hour data set of observed weather. Observations are subject to final quality control checks after publication.

Distributions (45)

CSV Latest 24 hours observed weather at Guernsey
Licence: open-government-licence
Options ▾ Download ▾



local-weather-data-abf5a : Local Weather Data

City Of Santa Rosa Municipal Service Center South 69 Stony Circle Santa Rosa, CA 95401 Davis Vantage Pro Weather Station

Catalog Data | RDF Quotients | Vocabularies | Analytics

Tags
weather

Counts

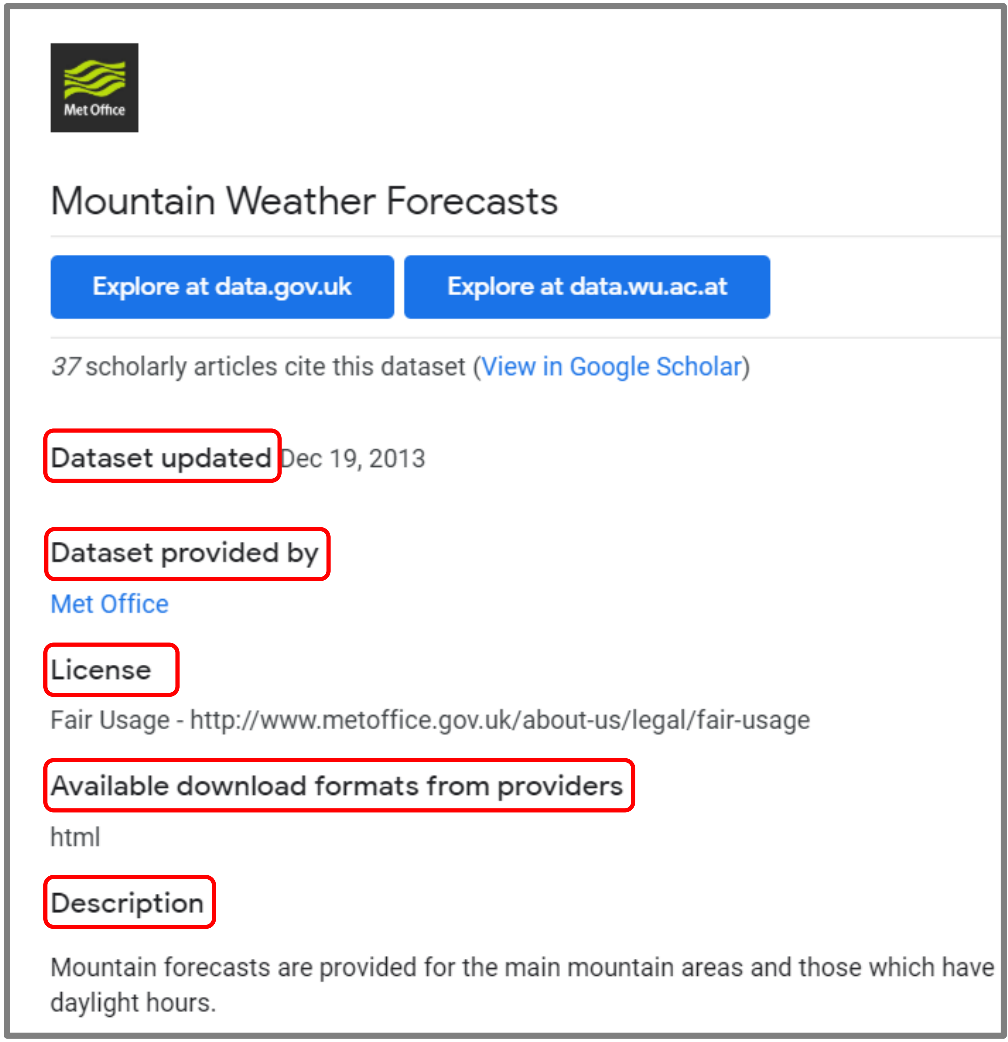
Triple count:	N/A
Outgoing link count:	0
Incoming link count:	0

Info

Name:	local-weather-data-abf5a
Title:	Local Weather Data
Organization:	City of Santa Rosa

Metadata:

- No detailed information of dataset content
- Limited utility for users to judge the relevance.



Mountain Weather Forecasts

37 scholarly articles cite this dataset (View in Google Scholar)

Dataset updated Dec 19, 2013

Dataset provided by

Met Office

License

Fair Usage - http://www.metoffice.gov.uk/about-us/legal/fair-usage

Available download formats from providers

html

Description

Mountain forecasts are provided for the main mountain areas and those which have daylight hours.

Google Dataset Search system

Dataset Snippet: Complementary to metadata

- a size-limited subset of triples
- exemplify dataset content
- illustrate the relevance to the query

Dataset Snippet: Complementary to metadata

- a size-limited subset of triples
- exemplify dataset content
- illustrate the relevance to the query

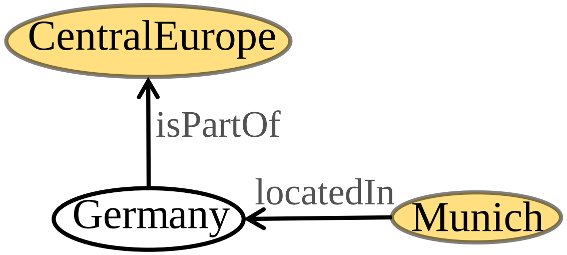
⟨Augsburg-TYPE-City⟩
⟨Berlin-capitalOf-Germany⟩
⟨Berlin-locatedIn-Germany⟩
⟨Berlin-neighboringCity-Dresden⟩
⟨Berlin-TYPE-Capital⟩
⟨Berlin-TYPE-City⟩
⟨Germany-isPartOf-CentralEurope⟩
⟨Germany-TYPE-Country⟩
⟨Munich-locatedIn-Germany⟩
⟨Munich-TYPE-City⟩
⟨Munich-neighboringCity-Augsburg⟩

An RDF dataset

Query:
Munich Europe

Snippet Generation

⟨Germany-isPartOf-CentralEurope⟩
⟨Munich-locatedIn-Germany⟩



An example snippet

Dataset Snippet: Complementary to metadata

- a size-limited subset of triples —— Size limit k
- exemplify dataset content —— Contain central elements of Schema and Instance
- illustrate the relevance to the query —— Contain query keywords

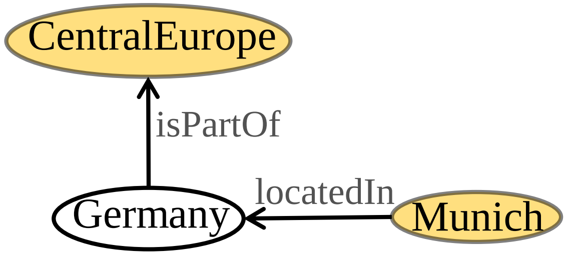
⟨Augsburg-TYPE-City⟩
⟨Berlin-capitalOf-Germany⟩
⟨Berlin-locatedIn-Germany⟩
⟨Berlin-neighboringCity-Dresden⟩
⟨Berlin-TYPE-Capital⟩
⟨Berlin-TYPE-City⟩
⟨Germany-isPartOf-CentralEurope⟩
⟨Germany-TYPE-Country⟩
⟨Munich-locatedIn-Germany⟩
⟨Munich-TYPE-City⟩
⟨Munich-neighboringCity-Augsburg⟩

An RDF dataset

Query:
Munich Europe

Snippet Generation

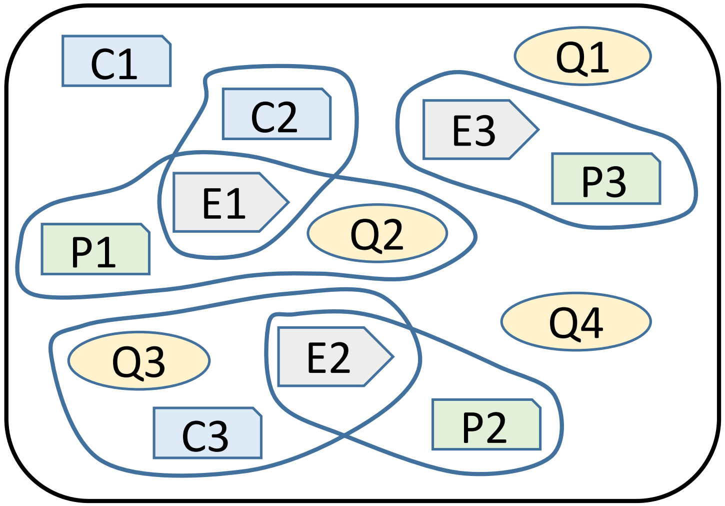
⟨Germany-isPartOf-CentralEurope⟩
⟨Munich-locatedIn-Germany⟩



An example snippet

Problem Formulation: A Weighted Maximum Coverage Problem

- Input: a collection of sets
- Select at most k sets to maximize the total weight of covered elements
- Consider *keywords, classes, properties, entities* as elements

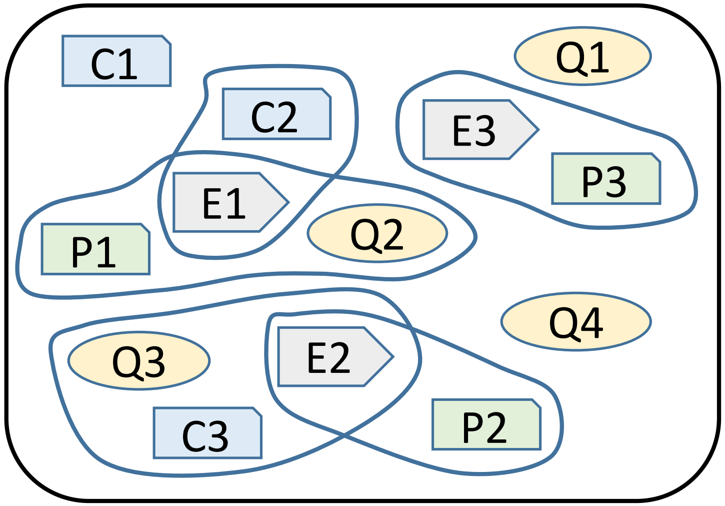


$U = Q \cup \mathrm{Cls}(T) \cup \mathrm{Prp}(T) \cup \mathrm{Ent}(T)$

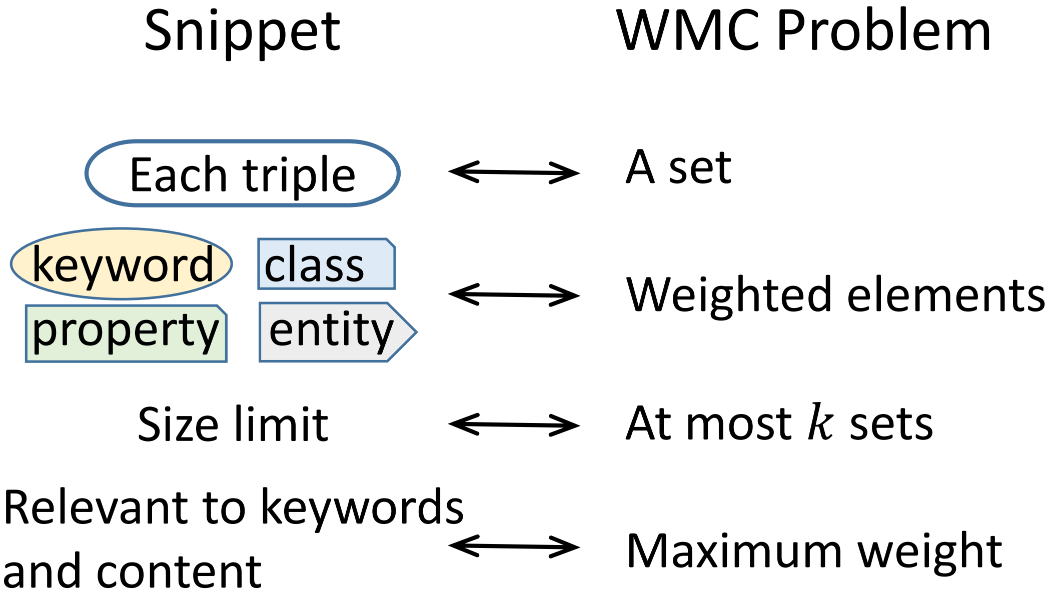
Snippet		WMC Problem
Each triple	⟷	A set
keyword class property entity	⟷	Weighted elements
Size limit	⟷	At most k sets
Relevant to keywords and content	⟷	Maximum weight

Problem Formulation: A Weighted Maximum Coverage Problem

Optimize the coverage of Keywords, Classes, Properties, and Entities.



$U = Q \cup \mathrm{Cls}(T) \cup \mathrm{Prp}(T) \cup \mathrm{Ent}(T)$



Snippet		WMC Problem
Each triple	↔	A set
keyword, class, property, entity	↔	Weighted elements
Size limit	↔	At most k sets
Relevant to keywords and content	↔	Maximum weight

Experiment

Datasets: 311 real datasets from *Datahub*

Queries:

- Real queries from *data.gov.uk*
- Artificial queries selected from *DMOZ*

Evaluation Metrics:

- Coverage of keywords
- Coverage of Paths between keywords
- Coverage of Dataset Schema
- Coverage of Data instance

Details about our Evaluation Metrics & Baselines

Tomorrow 12:00, Session 1B, FPAA level 0.

Result

- Our approach achieved a Balance among four evaluation metrics

	coKyw	coCnx	coSkm	coDat	Average
IlluSnip	0.1000	0.0540	0.6820	0.3850	0.3053
TA+C	0.9590	0.4703	0.0425	0.0915	0.3908
PrunedDP++	**1**	**1**	0.0898	0.2133	0.5758
CES	0.9006	0.3926	0.3668	0.2684	0.4821
coKSD	0.8352	0.3595	**0.8651**	**0.4247**	**0.6211**

- Our approach got consistent scores on different query groups

	coKyw	coCnx	coSkm	coDat	Average
data.gov.uk	0.7643	0.2882	0.8249	0.3870	0.5661
DMOZ-1	0.8977	0.7955	0.8873	0.4726	0.7633
DMOZ-2	0.8433	0.2444	0.8710	0.4569	0.6039
DMOZ-3	0.8395	0.2337	0.8693	0.4145	0.5893
DMOZ-4	0.7936	0.1877	0.8521	0.3731	0.5516

Conclusion

- Our approach (coKSD) achieved a Balance between evaluation metrics, can be used as a better dataset snippet than existing baselines.

Future Work

- Better snippet for dataset search
- Faster generation process

Thanks for your time!

Q&A

Contact: xxwang@smail.nju.edu.cn

- Keywords

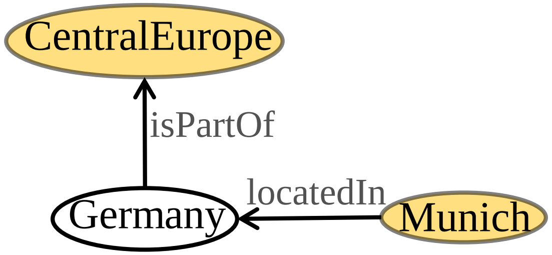
To match user’s data needs as much as possible

$$\text{maximizes } q(S) = \sum_{x \in \bigcup_{t_i \in S} \text{cov}(t_i)} \mathrm{w}(x), \qquad \text{subject to } |S| \leq k$$

$$w(x) = \begin{cases} \boxed{\alpha \cdot \frac{1}{|Q|}, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in Q} \\ \beta \cdot \text{frqCls}(x), & x \in \text{Cls}(T) \\ \beta \cdot \text{frqPrp}(x), & x \in \text{Prp}(T) \\ \gamma \cdot \left(\frac{\log(d^{+}(x)+1)}{\sum_{e \in \text{Ent}(T)} \log(d^{+}(e)+1)} + \frac{\log(d^{-}(x)+1)}{\sum_{e \in \text{Ent}(T)} \log(d^{-}(e)+1)} \right), & x \in \text{Ent}(T) \end{cases}$$

Query:
Munich Europe

⟨Augsburg-TYPE-City⟩
⟨Berlin-capitalOf-Germany⟩
⟨Berlin-locatedIn-Germany⟩
⟨Berlin-neighboringCity-Dresden⟩
⟨Berlin-TYPE-Capital⟩
⟨Berlin-TYPE-City⟩
⟨Germany-isPartOf-CentralEurope⟩
⟨Germany-TYPE-Country⟩
⟨Munich-locatedIn-Germany⟩
⟨Munich-TYPE-City⟩
⟨Munich-neighboringCity-Augsburg⟩



- Classes and properties

To exemplify central schema elements

$$\text{maximizes } q(S) = \sum_{x \in \bigcup_{t_i \in S} \operatorname{cov}(t_i)} \mathrm{w}(x), \qquad \text{subject to } |S| \leq k$$

$$w(x) = \begin{cases} \alpha \cdot \frac{1}{|Q|}, & x \in Q \\ \boxed{\beta \cdot \operatorname{frqCls}(x), \quad x \in \operatorname{Cls}(T) \\ \beta \cdot \operatorname{frqPrp}(x), \quad x \in \operatorname{Prp}(T)} \\ \gamma \cdot \left(\frac{\log(d^{+}(x)+1)}{\sum_{e \in \operatorname{Ent}(T)} \log(d^{+}(e)+1)} + \frac{\log(d^{-}(x)+1)}{\sum_{e \in \operatorname{Ent}(T)} \log(d^{-}(e)+1)} \right), & x \in \operatorname{Ent}(T) \end{cases}$$

⟨Augsburg-TYPE-**City**⟩
⟨Berlin-capitalOf-Germany⟩
⟨Berlin-**locatedIn**-Germany⟩
⟨Berlin-**neighboringCity**-Dresden⟩
⟨Berlin-TYPE-Capital⟩
⟨Berlin-TYPE-**City**⟩
⟨Germany-isPartOf-CentralEurope⟩
⟨Germany-TYPE-Country⟩
⟨Munich-**locatedIn**-Germany⟩
⟨Munich-TYPE-**City**⟩
⟨Munich-**neighboringCity**-Augsburg⟩

Central schema elements:

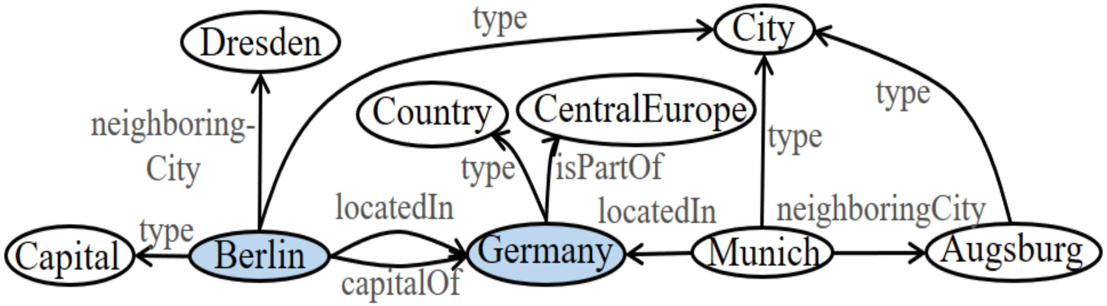
- Frequent classes & properties
  - e.g., **City**: 3 times
    locatedIn: 2 times
    neighboringCity: 2 times

- **Entities**

To show central elements at data level

$$\text{maximizes } q(S) = \sum_{x \in \bigcup_{t_i \in S} \mathrm{cov}(t_i)} \mathrm{w}(x), \qquad \text{subject to } |S| \leq k$$

$$w(x) = \begin{cases} \alpha \cdot \frac{1}{|Q|}, & x \in Q \\ \beta \cdot \mathrm{frqCls}(x), & x \in \mathrm{Cls}(T) \\ \beta \cdot \mathrm{frqPrp}(x), & x \in \mathrm{Prp}(T) \\ \boxed{\gamma \cdot \left(\frac{\log(d^+(x)+1)}{\sum_{e \in \mathrm{Ent}(T)} \log(d^+(e)+1)} + \frac{\log(d^-(x)+1)}{\sum_{e \in \mathrm{Ent}(T)} \log(d^-(e)+1)} \right), \quad x \in \mathrm{Ent}(T)} \end{cases}$$



RDF dataset

Central entity:

- High in-degree and out-degree

Approach

```
/*Greedy Algorithm: coKSD*/
Input: A dataset $T$ , a keyword query $Q$, a size bound $k$
Output: An optimum dataset snippet $S \subseteq T$
$S \leftarrow \emptyset$;
while $|S| < k$ do
    $t^* \leftarrow argmax_{t \in (T \setminus \mathrm{S})}\big(q(S \cup \{t\}) - q(S)\big)$;
    $S \leftarrow S \cup \{t^*\}$;   //select an optimal triple in each step
end while
return $S$;
```

Approximation ratio: $1 - \frac{1}{e}$