A Framework for Evaluating Snippet Generation for Dataset Search

Xiaxia Wang[1], Jinchi Chen[1], Shuxin Li[1], Gong Cheng[1], Jeff Z. Pan[2,3], Evgeny Kharlamov[4,5], Yuzhong Qu[1]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China

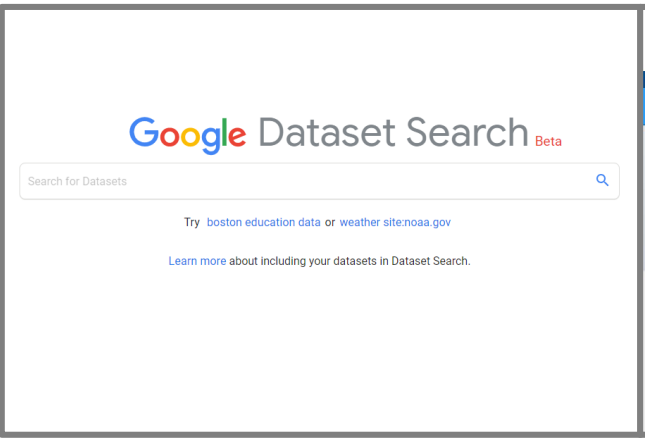
[2] Edinburgh Research Centre, Huawei, UK

[3] Department of Computing Science, University of Aberdeen, UK

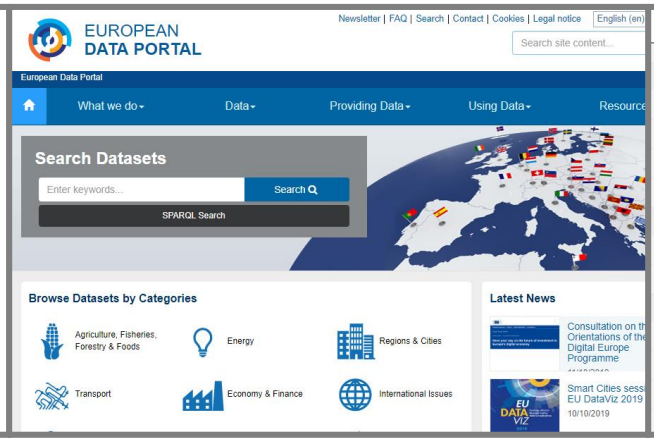
[4] Department of Informatics, University of Oslo, Norway

[5] Bosch Center for Artificial Intelligence, Robert Bosch GmbH, Germany

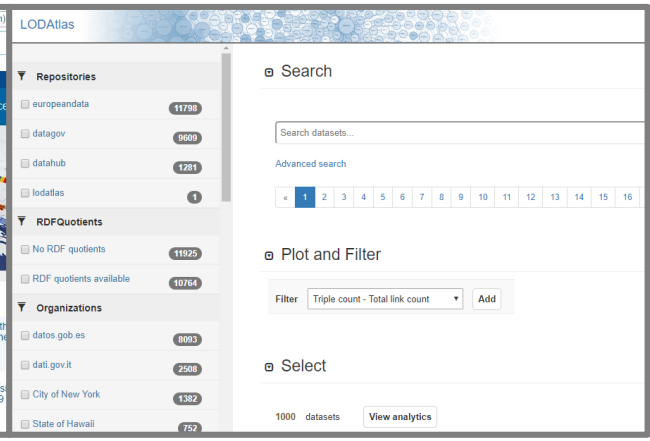
Dataset search systems: Conveniently find relevant datasets.



Google Dataset Search

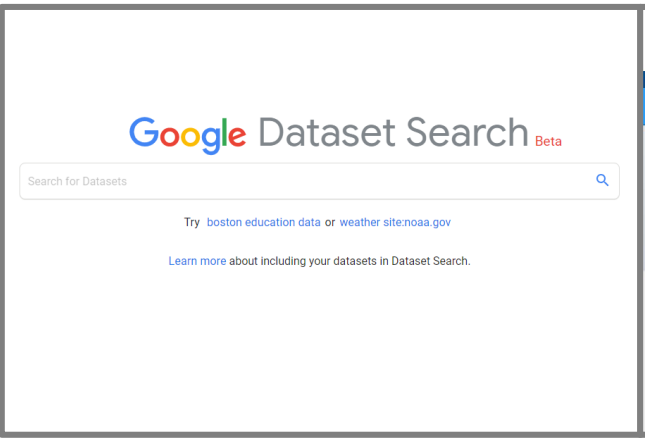


European Data portal

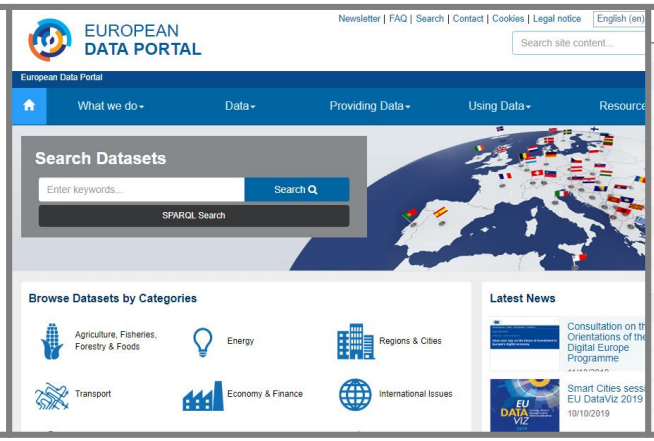


LODAtlas

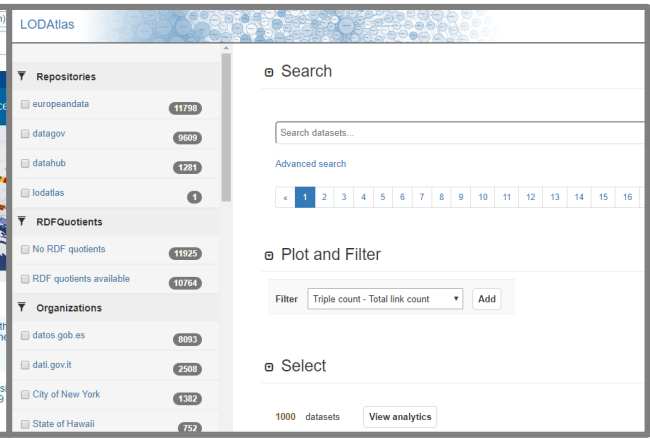
Dataset search systems: Conveniently find relevant datasets.



Google Dataset Search

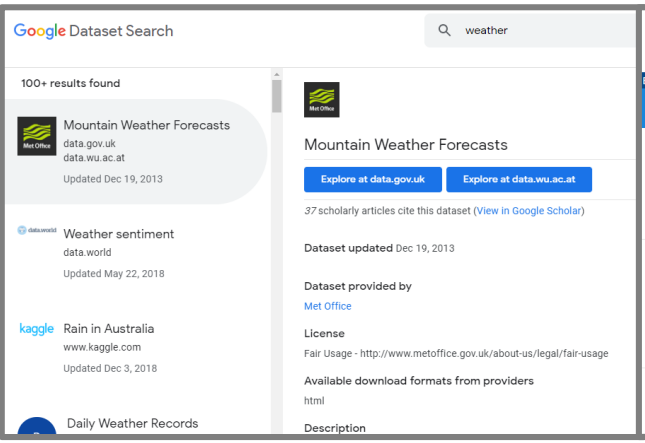


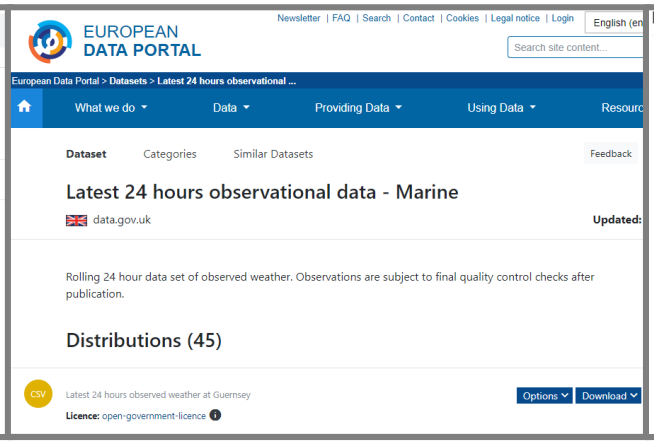
European Data portal

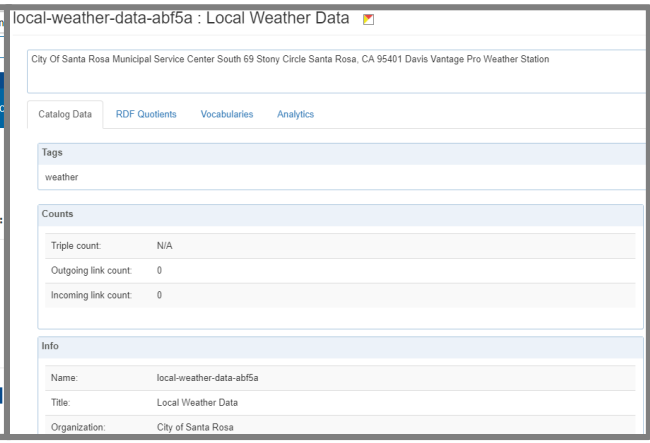


LODAtlas

Existing systems serve only **metadata** for relevance judgement.



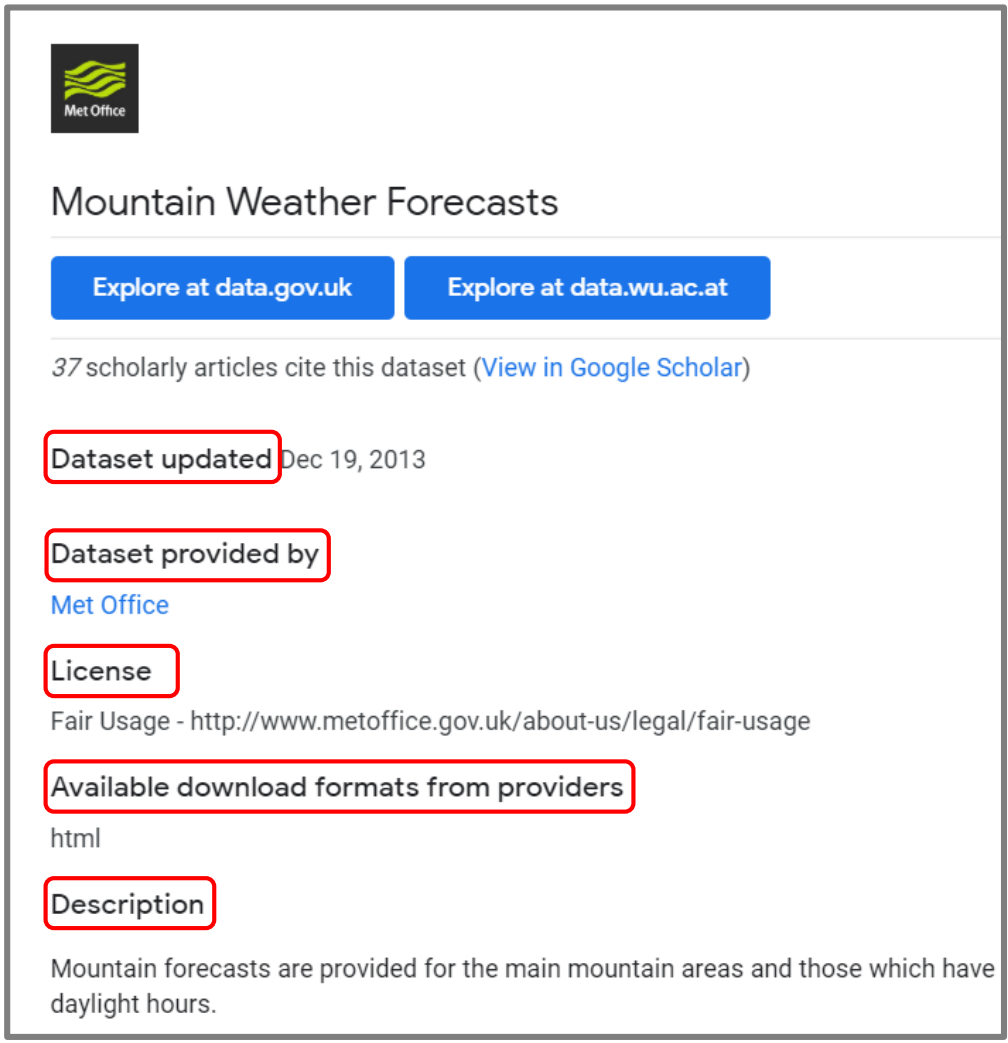




Metadata:

- No detailed information of dataset content
- Limited utility for relevance judgment

How to improve?



Google Dataset Search system

A **dataset snippet**: A subset of RDF triples to exemplify the dataset content and its 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⟩

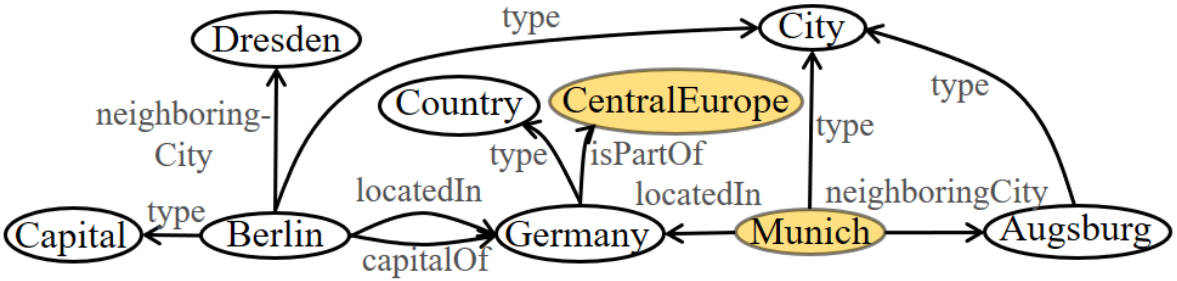
Dataset: A set of triples

A keyword query
e.g., Munich Europe

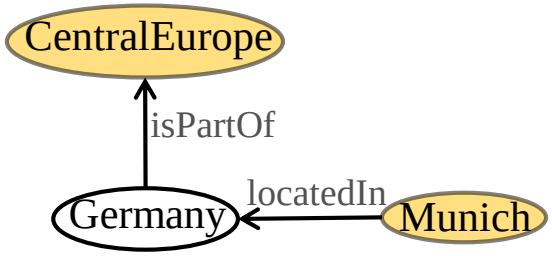
Snippet Generation

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

Snippet: A size-limited subset of triples

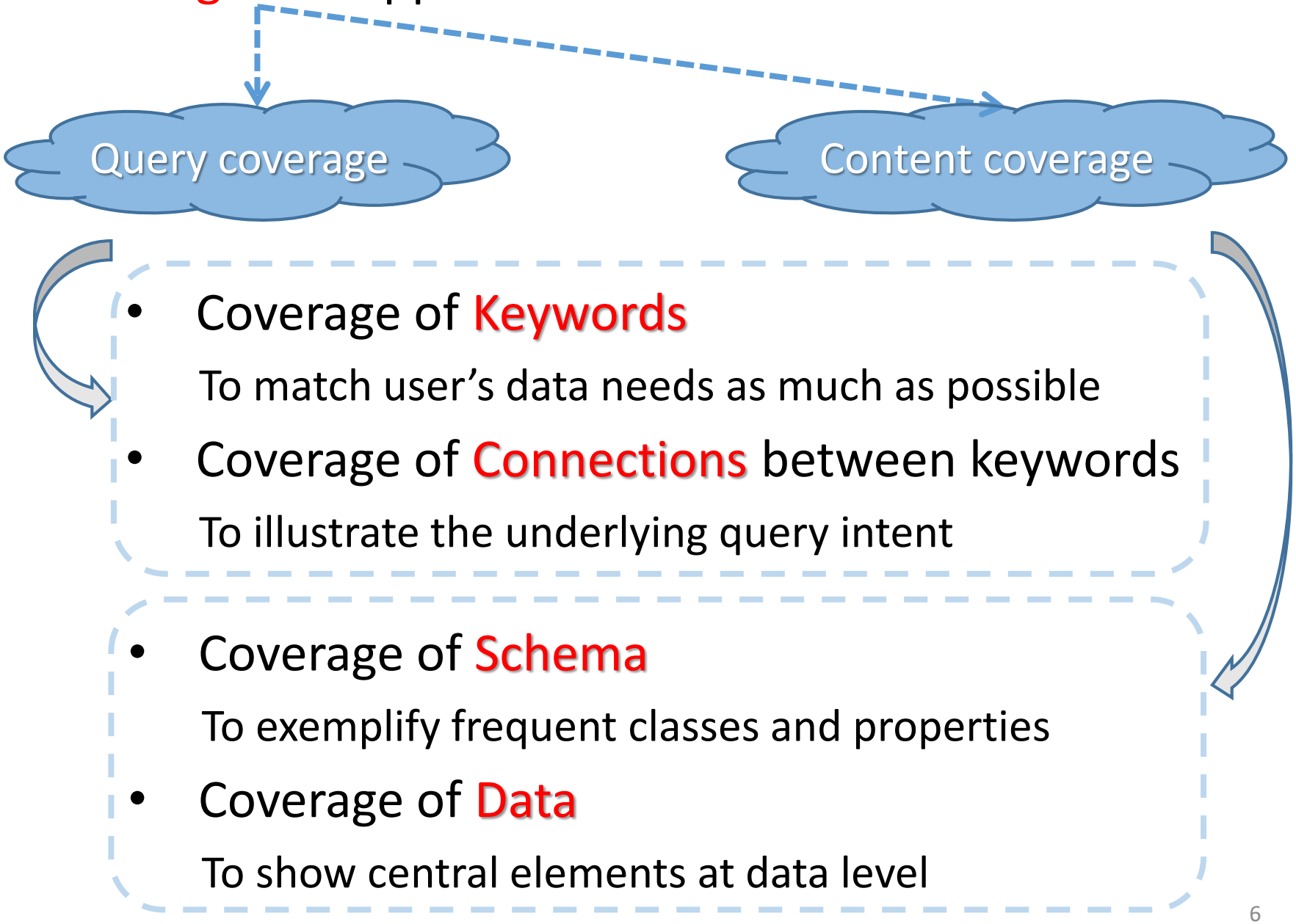


An RDF graph



A subgraph

What is a good snippet?



What is a good snippet?

Query coverage

Content coverage

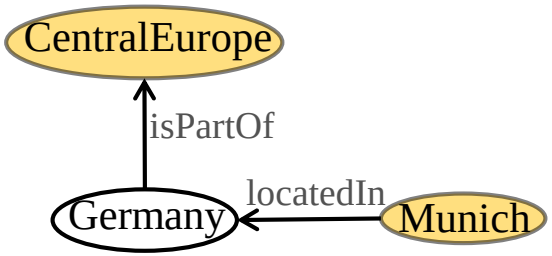
- Coverage of Keywords

  To match user's data needs as much as possible

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⟩

An RDF dataset



An RDF graph

What is a good snippet?

Query coverage

Content coverage

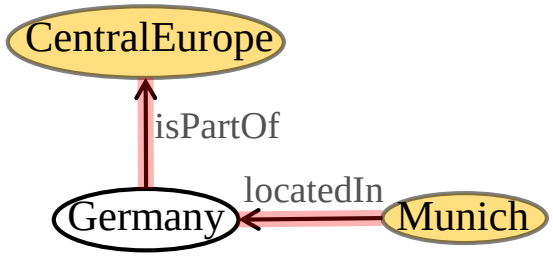
- Coverage of Keywords

  To match user's data needs as much as possible

- Coverage of Connections between keywords

  To illustrate the underlying query intent

Query:
Munich Europe



: Paths between keywords

A dataset snippet

What is a good snippet?

Query coverage

Content coverage

⟨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

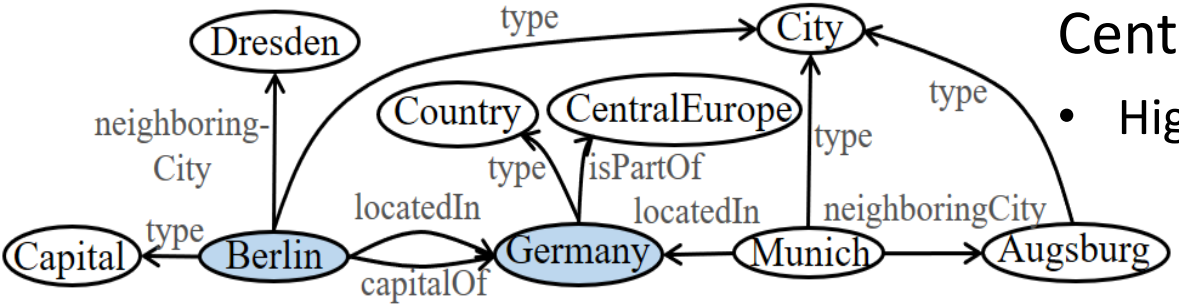
- Coverage of Schema

  To exemplify frequent classes and properties

What is a good snippet?

Query coverage

Content coverage



RDF dataset

Central **entity**:

- High in-degree and out-degree

- Coverage of Schema

  To exemplify frequent classes and properties

- Coverage of Data

  To show central elements at data level

What is a good snippet?

Query coverage

Content coverage

- Coverage of Keywords

  To match user's data needs as much as possible

- Coverage of Connections between keywords

  To illustrate the underlying query intent

- Coverage of Schema

  To exemplify frequent classes and properties

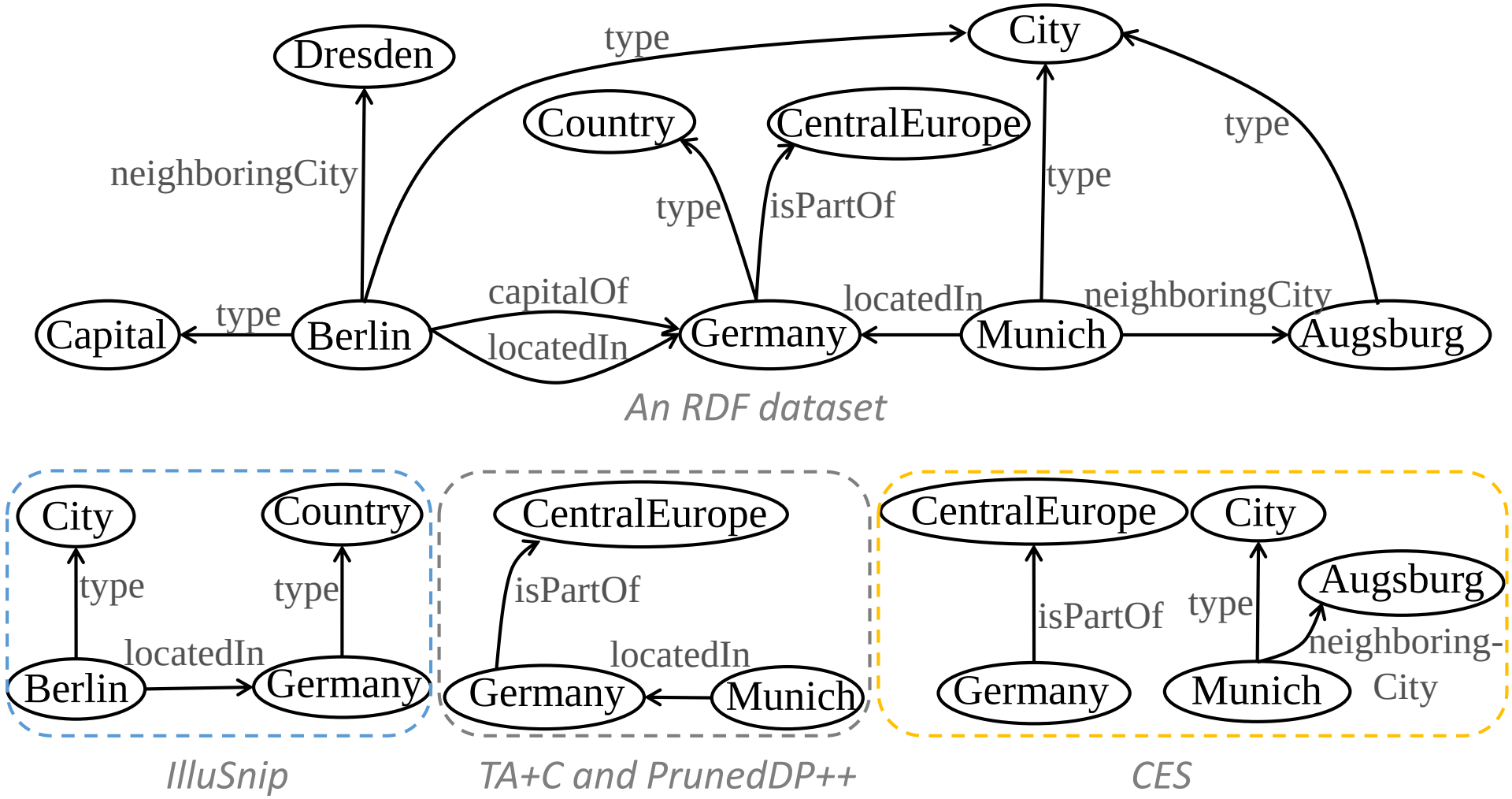
- Coverage of Data

  To show central elements at data level

No specialized Generation methods?
- Adapt from related fields.

- Snippets for **RDF Datasets** (*IlluSnip*) *[Cheng et al., WSDM, 2017]*
  select central entities, classes, properties
- Snippets for **Ontology Schemas** (*TA+C*) *[Ge et al., IPM, 2013]*
  decompose the RDF graph to tree-structured subgraphs
- Keyword search on **graphs** (*PrunedDP++*) *[Li et al., SIGMOD, 2016]*
  model as a *Group Steiner Tree* problem
  contains all the keywords, connected
- Snippets for **Documents** (*CES*) *[Feigenblat et al., SIGIR, 2017]*
  consider each triple as a sentence

Snippet generation Examples



3 snippets generated by different methods w.r.t. the query ***Munich Europe***

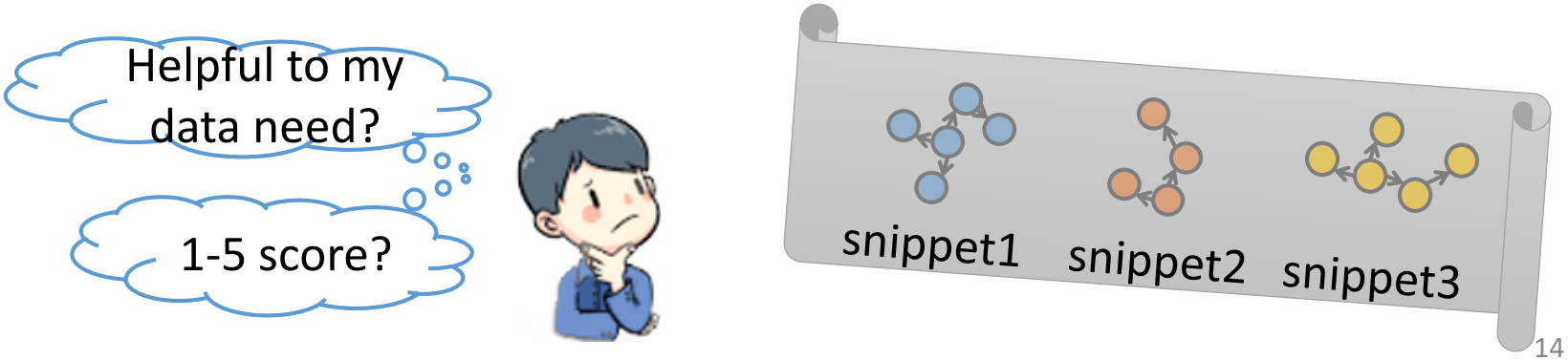
Experiments

Exp 1. Scoring the snippets with the evaluation metrics

- 311 Real Datasets: From *Datahub*
- Queries:
  - Real queries from *data.gov.uk*
  - Artificial queries selected from *DMOZ*

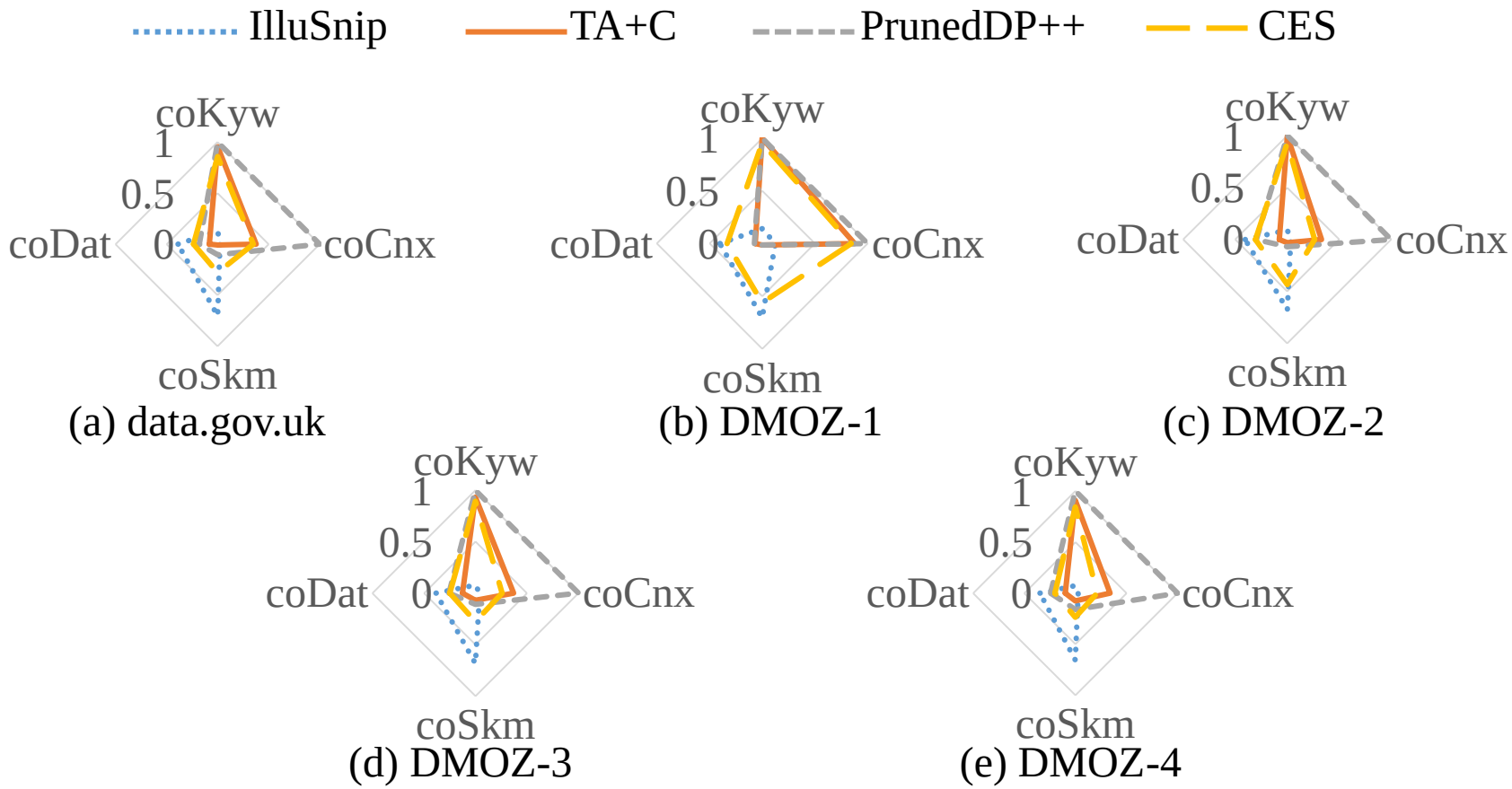
Exp 2. User study

- 20 participants, rating on a 1-5 scale



Result

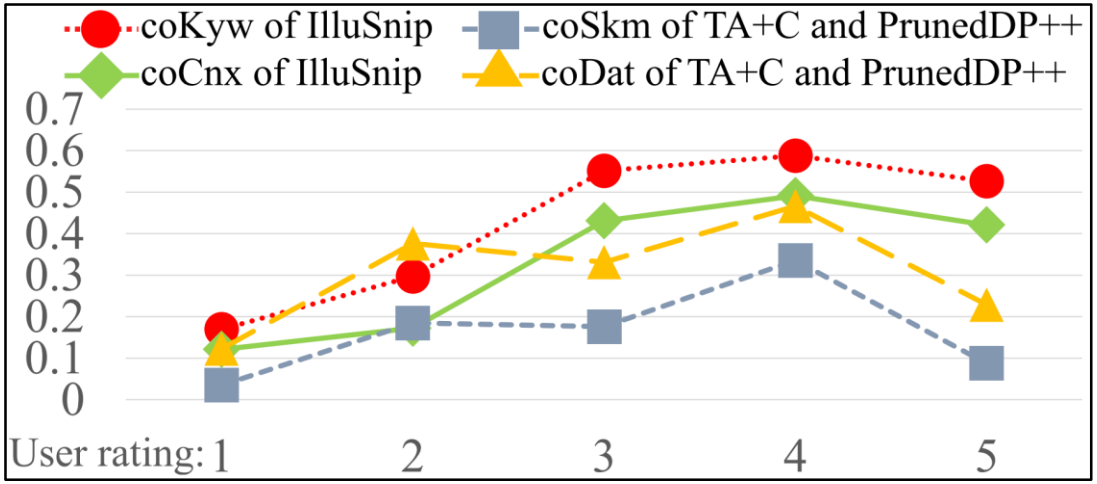
- Different methods have different preferences.
- NO one achieves a balance between all metrics.



Average scores of evaluation metrics on each group of query-dataset pairs

Result

- The positive correlation between User ratings and Metrics scores shows the validity of the evaluation framework.
- Users are NOT satisfied with existing methods.



Correlation between Metrics and User Ratings

	Human-rated scores (1 – 5)
IlluSnip	3.10 ± 1.28
TA+C	2.36 ± 1.29
PrunedDP++	1.92 ± 1.19

Human-rated Usefulness of Snippets

➢ Contribution

- Propose an evaluation framework for dataset snippet
- Adapt 4 SOTA generation methods from related fields
- Evaluate these methods on real-world datasets

➢ Conclusion

- Existing methods have different preferences, but no one achieves a balance between all aspects.
- The user ratings are consistent with metrics scores, indicating the validity of our evaluation framework.

➢ Future Work

- More metrics
- Snippet for other data formats
- Better snippets

Thanks for your time!

Q&A

Contact: xxwang@smail.nju.edu.cn

Differences: Existing Methods vs Our Intent

- Snippets for **RDF Datasets** (*IlluSnip*)
  NOT query-biased vs Query-related
- Snippets for **Ontology Schemas** (*TA+C*)
  Don't care about Schema or Data instance vs Cover main content
- Keyword search on **graphs** (*PrunedDP++*)
  Focus ONLY on Keywords vs Also illustrate Data Content
- Snippets for **Documents** (*CES*)
  Diversified triples are usually disparate vs Connections between keywords make sense