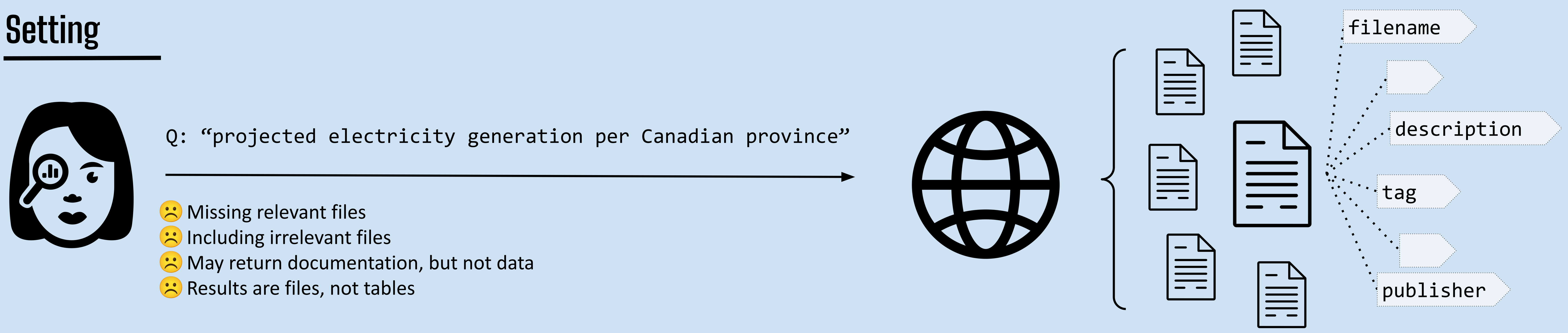
Leveraging Guides to Empower Open Data Research

Christina Christodoulakis, Moshe Gabel, Angela Demke Brown
{**christina**, mgabel, demke}@cs.toronto.edu
Department of Computer Science, University of Toronto

UNIVERSITY OF TORONTO

Setting



Q: “projected electricity generation per Canadian province”

- ☹ Missing relevant files
- ☹ Including irrelevant files
- ☹ May return documentation, but not data
- ☹ Results are files, not tables

Idea:

Empower table discovery: generate rich metadata by extracting data guides from documentation and linking them to tables found in files!

Attribute Guides:
- Name
- Title
- Description
- Unit
- Scale
- Datatype
- Domain

Table Search

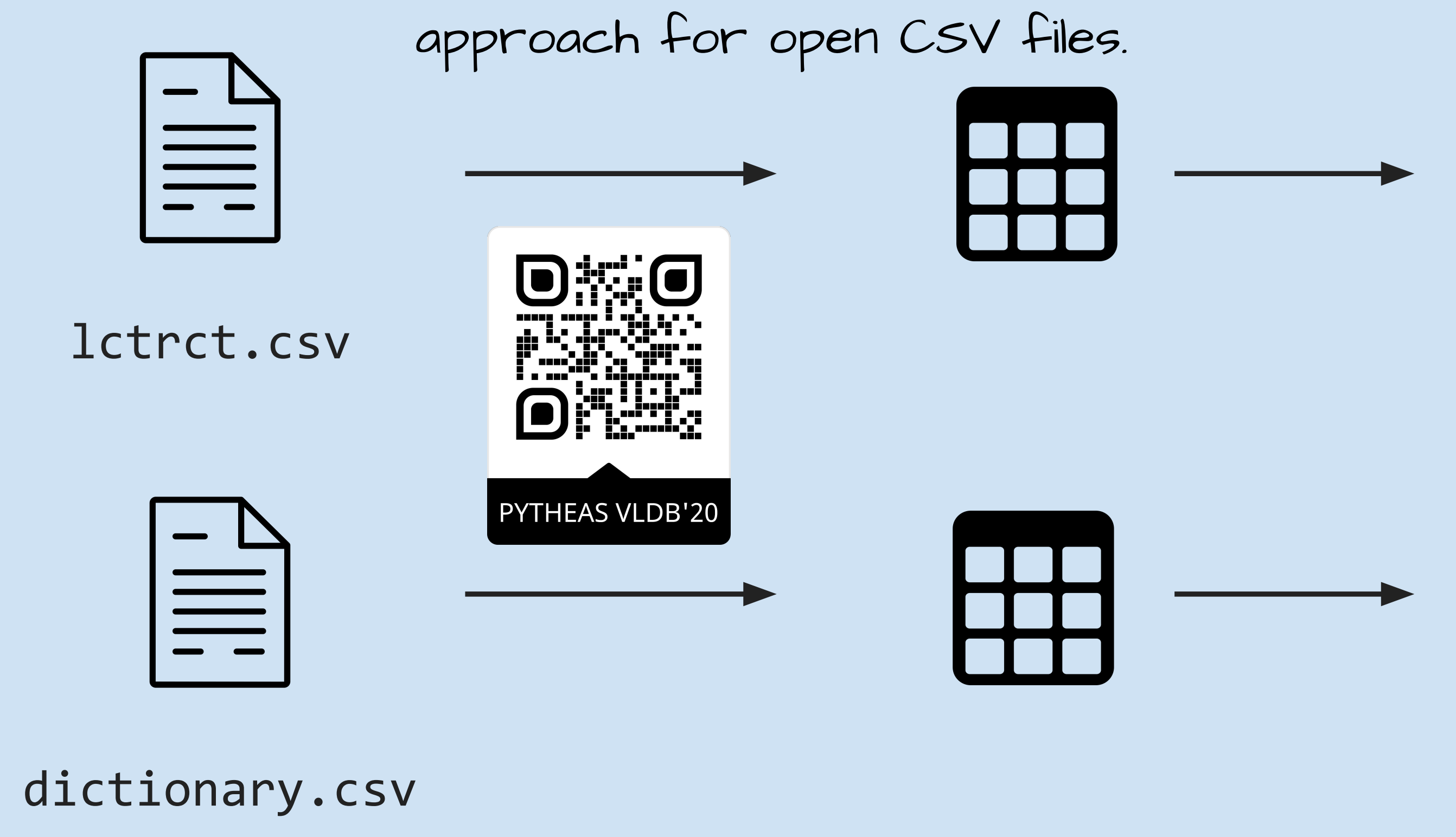
Integration

Interpretation

Value:
- Science
- Journalism
- Businesses
- Government
- ...

Example

Table discovery is challenging across file formats! We have introduced an approach for open CSV files.



Source	Area	Year	Data
Hydro	Alberta	2005	358386.9629
Biofuels/Biomass	Newfoundland and Labrador	2017	70
Natural Gas	Newfoundland and Labrador	2005	269
Solar/Wind/Geothermal	Nunavut	2040	94.6
Nuclear	Ontario	2017	90065.48

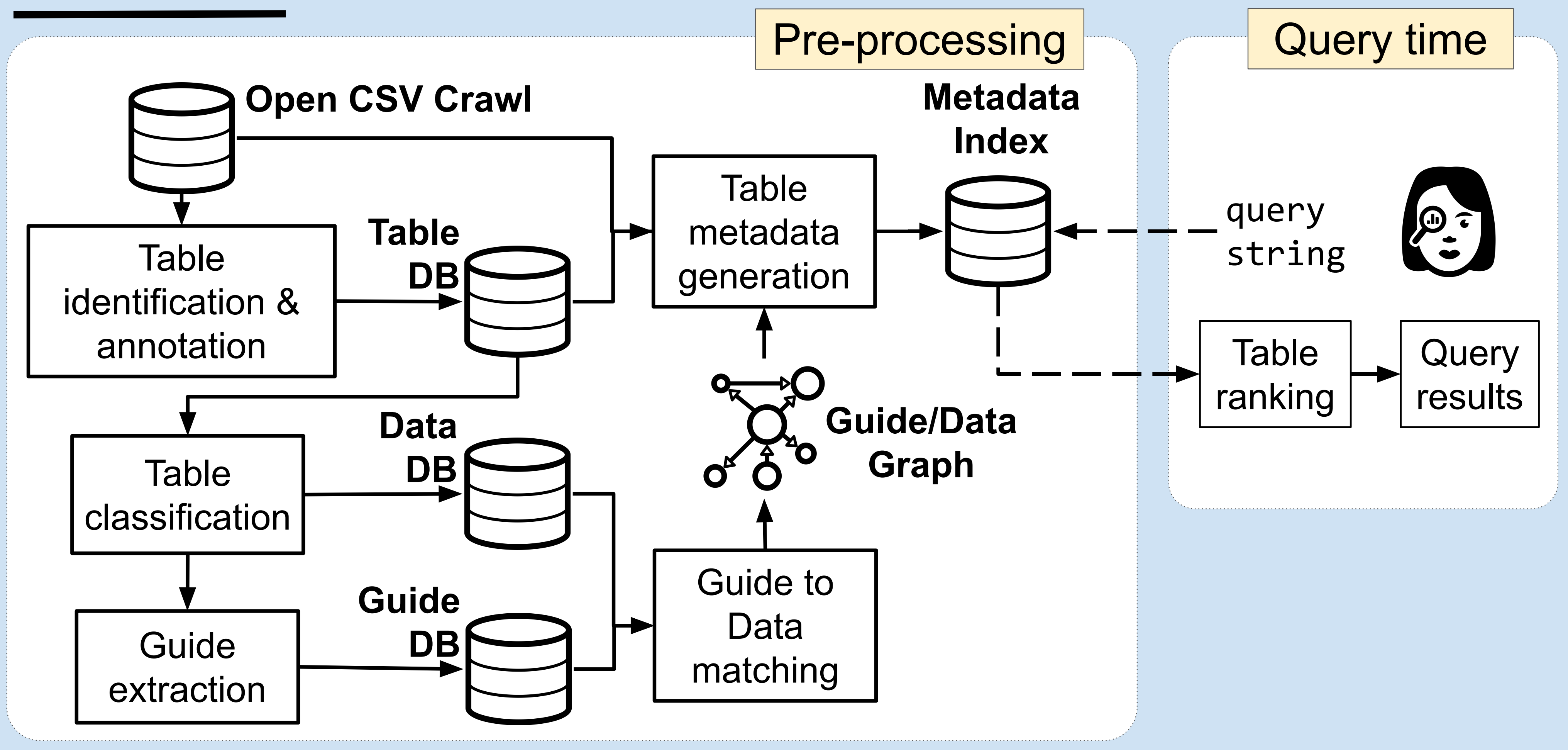
“Source refers to the various energy types used to produce electricity.”

“Province or territory.”

“Year that the data refers to. For end-use demand, 2005 to 2013 are historical numbers, while 2014 to 2040 are projected values.”

“Electric energy measured in GW.h is a billion (109) watt hours of electric energy per year. One GW.h is equal to 0.0036 petajoules.”

Design



References

1. Apache Lucene, https://lucene.apache.org/
2. Capgemini Consulting: Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources (2015), accessed: 2019-09-23
3. **Christodoulakis**, C., Munson, E., Gabel, M., Brown, A.D., Miller, R.J.: Pytheas: Pattern-based table discovery in CSV files. PVLDB 13 (11), 2075–2089 (2020)
4. Liu, Y., Bai, K., Mitra, P., Giles, C.L.: Tableseer: Automatic table metadata extraction and searching in digital libraries. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. p. 91–100. JCDL ’07, Association for Computing Machinery, New York, NY, USA (2007)
5. Miller, R.J., Nargesian, F., Zhu, E., **Christodoulakis**, C., Pu, K.Q., Andritsos, P.: Making open data transparent: Data discovery on open data. IEEE Data Eng. Bull. 41 (2), 59–70 (2018)
6. Machova, R., Hub, M., Lnenicka, M.: Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders’ perspective. Aslib Journal of Information Management 70 (05 2018)