Towards Transforming Tabular Datasets into Knowledge Graphs

Nora Abdelmageed

Heinz-Nixdorf Chair for Distributed Information Systems & Computer Vision Group, Michael Stifel Center Jena
Friedrich Schiller University Jena, Germany

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

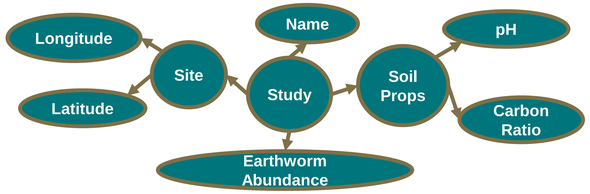
Introduction & Motivation

- Tabular datasets are not reusable or queryable!

Study Name	Earthworm Abundance	Site Name	Longitude	Latitude	pH	Carbon ratio
Szlavecz2006	2	Cylburn Arboretum	-76.6595085	39.3504699	5.11	21.31
Szlavecz2006	12	Leakin Park	-76.6968441	39.3051975	4.82	19.995

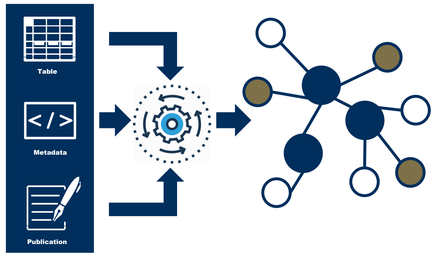
"What is the Abundance of earthworms where pH > 4?"

- Automatic Transformation to KG.



Block Diagram

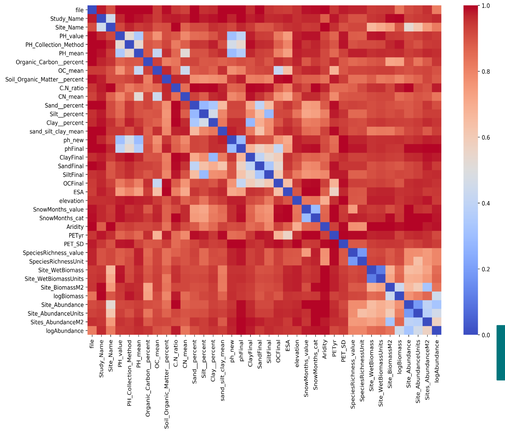
- What if table headers are missing or not descriptive enough?
- Other sources of information should be leveraged: Metadata & publications



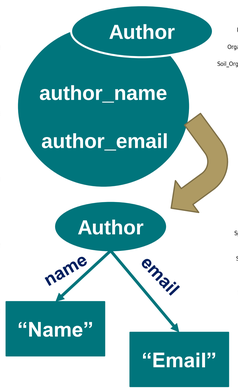
Initial Experiment

- Table headers could be clustered to infer KG concepts and properties.
- sWorm dataset is used for this experiment [1]

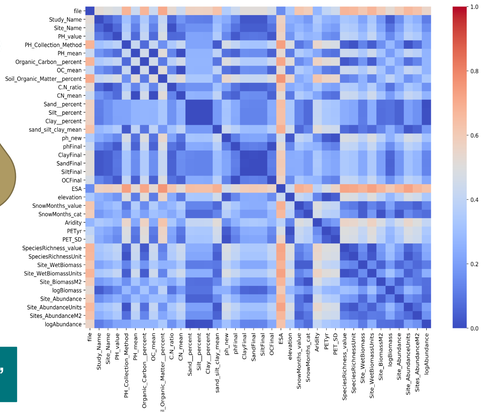
Semantic Representation [2]



Hypothesis

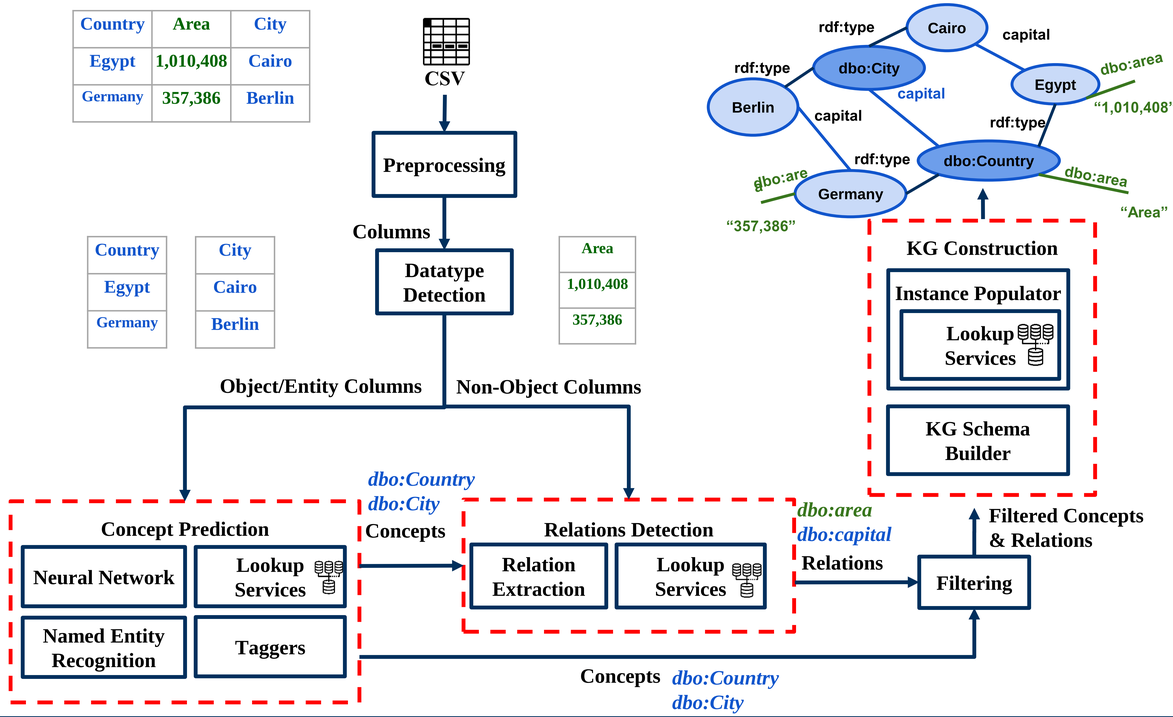


Syntactic Representation



System Architecture

Phase 1 of the project considers the CSV files only.



References

[1] Phillips, Helen RP, Carlos A. Guerra, Marie LC Bartz, Maria JI Briones, George Brown, Thomas W. Crowther, Olga Ferlian et al. "Global distribution of earthworm diversity." Science 366, no. 6464 (2019): 480-485.
[2] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Ecient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

Acknowledgments: The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (P5)" within the scope of the program line "Breakthroughs: Exploring Intelligent Systems" for "Digitization - explore the basics, use applications". I thank Birgitta Koenig-Ries, Joachim Denzler, and Sheeba Samuel for their guidance and feedback.

nora.abdelmageed@uni-jena.de
@NoraYoussef

Computer Vision Group
MichaelStifelCenterJena for Data-Driven and Simulation Science
Carl Zeiss Stiftung
ESWC20