

Background

- Entity recognition and linking among heterogeneous datasets are of great value in data integration and reuse
- Detecting and linking entity mentions in datasets can be facilitated by using knowledge bases on the Web, such as ontologies and vocabularies
- The number of ontologies and vocabularies, including those in the field of geosciences, has been continually increasing. Those are valuable contributions to the Web of Data



Our Aim

A Web-based entity linking and wikification service for datasets in Earth and environmental sciences

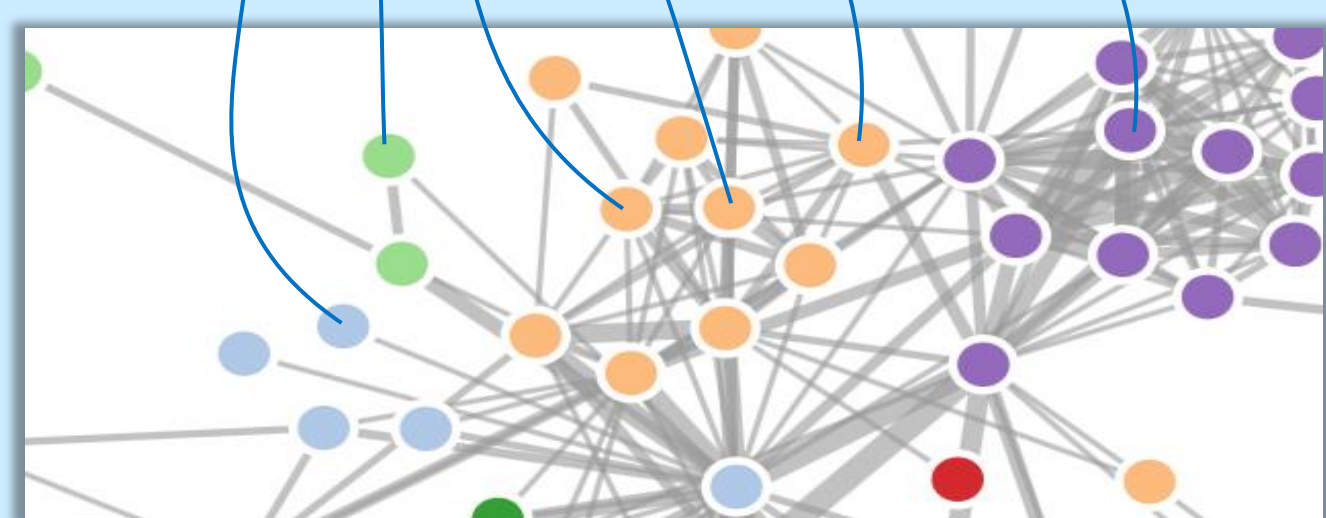
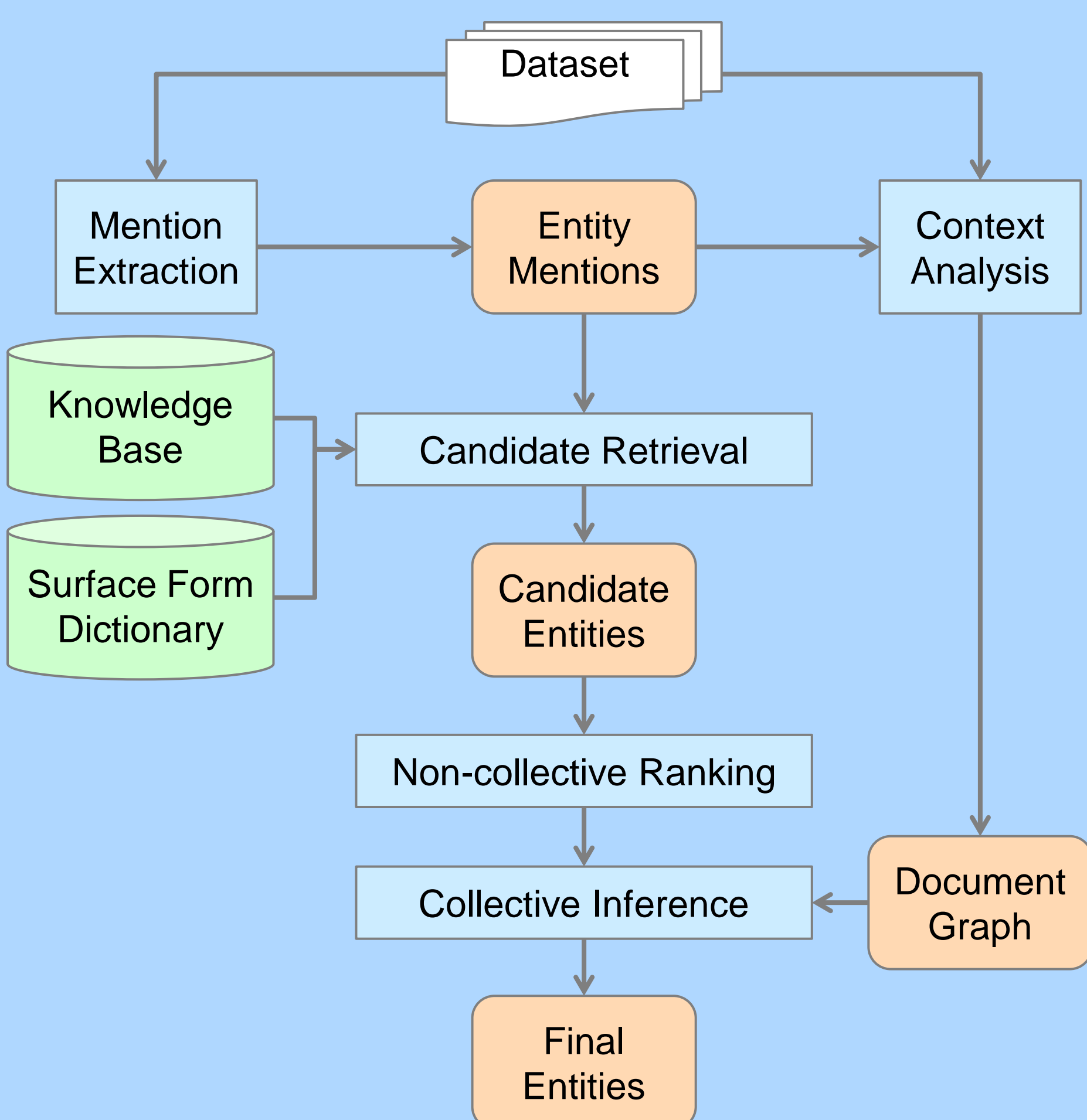


Image credit: sciencemag.org and gravity.com

Nature of Efforts

- Challenge:** Existing supervised approaches require a large amount of manually-labeled training data. Such training data are limited in Earth and environmental sciences.
- Approach:** An un-supervised collective inference approach for entity linking.



Technical Details

Mention Extraction:

- Uses publicly available name tagger and regular expressions to extract entity mentions

Context Analysis:

- Sentence level: If terms appear in same sentence, then we consider they are related to each other
- Paragraph level: If terms appear in same paragraph, then we consider they are related to each other

Candidate Retrieval:

- Given a surface form of entity mentions, retrieve all entities with surface form that are similar to the mentions' surface form
- Surface form is the textual appearance of entities/mentions. The Surface form dictionary is constructed based on the knowledge base

Non-collective Ranking:

- Apply to the candidate entities
- Assign entities with higher popularity a higher score, similar to PageRank

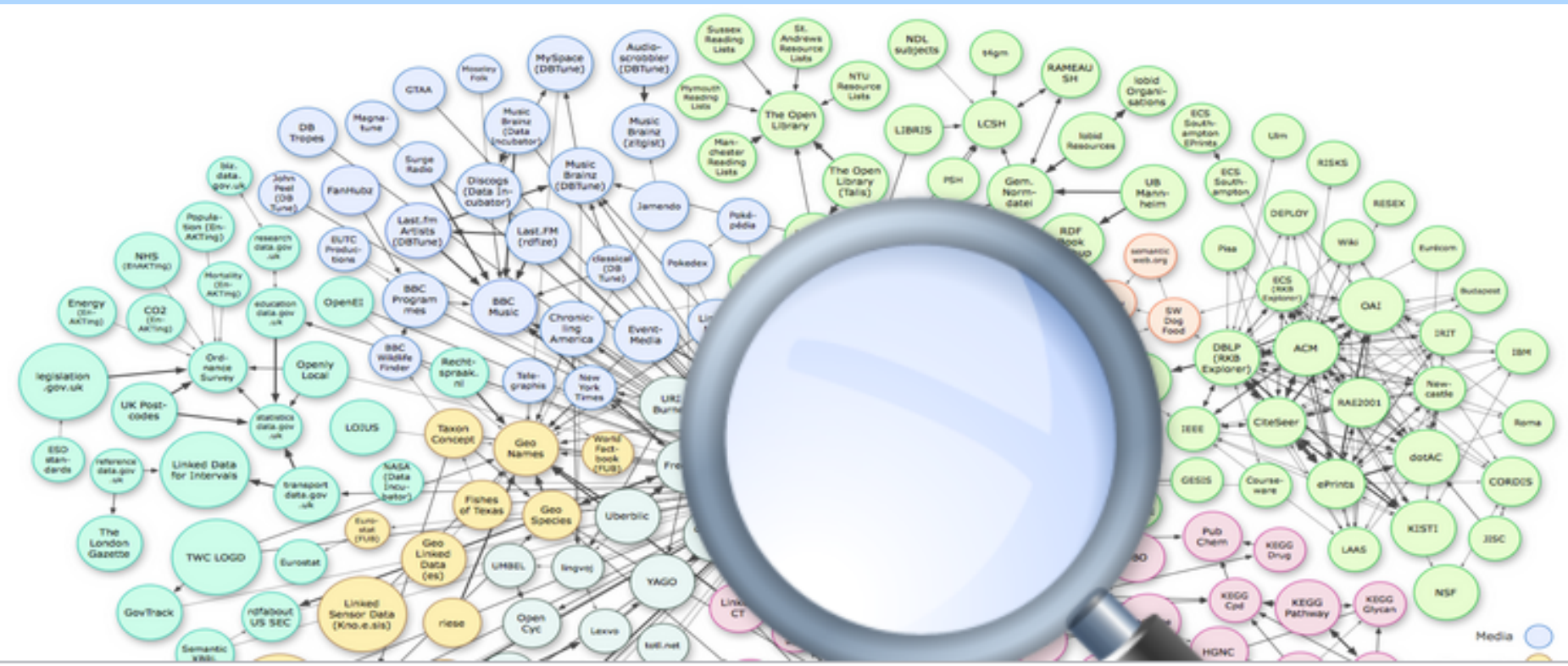
Collective Inference:

- Analyze the mentions in a context simultaneously to determine the best reference entities
- Both document graph and graph of candidate entities contain important contextual information about mentions and entities

Initial Results

Our current demo system uses DBpedia as the knowledge base.

Input Dataset/Document



The Jurassic constitutes the middle period of the Mesozoic Era, also known as the Age of Reptiles. The start of the period is marked by the major Triassic-Jurassic extinction event. Two other extinction events occurred during the period: the Late Pliensbachian/Early Toarcian event in the Early Jurassic, and the Late Tithonian event at the end; however, neither event ranks among the "Big Five" mass extinctions.

Annotate

Era [59,62]: <http://dbpedia.org/resource/Jurassic>
 extinction event [164,180]: http://dbpedia.org/resource/Extinction_event
 event ranks [363,374]: [NIL](#)
 extinction events [192,209]: http://dbpedia.org/resource/Extinction_event
 Pliensbachian/Early Toarcian event [247,281]: [NIL](#)
 period [230,236]: http://dbpedia.org/resource/Orbital_period
 period [36,42]: http://dbpedia.org/resource/Orbital_period
 event [328,333]: http://dbpedia.org/resource/The_Event
 period [116,122]: http://dbpedia.org/resource/Orbital_period
 Age [82,85]: <http://dbpedia.org/resource/Cretaceous>
 end [341,344]: [http://dbpedia.org/resource/End_\(American_football\)](http://dbpedia.org/resource/End_(American_football))
 start [103,108]: http://dbpedia.org/resource/IK_Start
 mass extinctions [396,412]: http://dbpedia.org/resource/Extinction_event
 Reptiles [89,97]: <http://dbpedia.org/resource/Reptile>

Result of entity recognition and linking

Future Works

- Enrich the knowledge base: more ontologies and vocabularies in the field of Earth and environmental sciences
- Semantic parsing: to improve the result of collective inference
- Semantic reasoning: to improve the quality of linking
- Propose collaboration with the ESIP portal for ontology and vocabulary registration and knowledge base construction

Further Reading

- Zheng, J., Fu, L., Ma, X., Fox, P., 2015. SEM+: tool for discovering concept mapping in Earth science related domain. Earth Science Informatics 8 (1), 95-102.
- Zheng, J., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D.L., Hendler, J., and Ji, H. 2014. Entity Linking for Biomedical Literature. In Proceedings of ACM 8th International Workshop on Data and Text Mining in Bioinformatics, Shanghai, China.