Construction of a Knowledge Graph in the Climate Research Domain

Andrija Poleksić [1,2]   Sanda Martinčić-Ipšić [1,2]

[1]Faculty of Informatics and Digital Technologies, University of Rijeka
[2]Center for Artificial Intelligence and Cybersecurity, University of Rijeka

University of Rijeka
Faculty of Informatics and Digital Technologies

Knowledge Graph: Definition and Structure

A **Knowledge Graph (KG)** is a structured representation of knowledge, formally defined as $G = E, R, T, F_k$ [1].

It consists of a set of **Entities** (E), phenomena or objects within a particular domain (e.g. *summer rainfall*, *North Pacific Ocean*) and a set of **Relations** (R) describing the interactions between these entities (e.g. *causes variability*, *derived from*).

Facts (T) are stored as triples ($(h, r, t) \in T; h, t \in E; r \in R$) (e.g. ***EASM***, ***has effect on***, ***summer rainfall***) and form a network in which entities are nodes and relations are labeled edges. This structure facilitates querying and inference over complex domain knowledge.

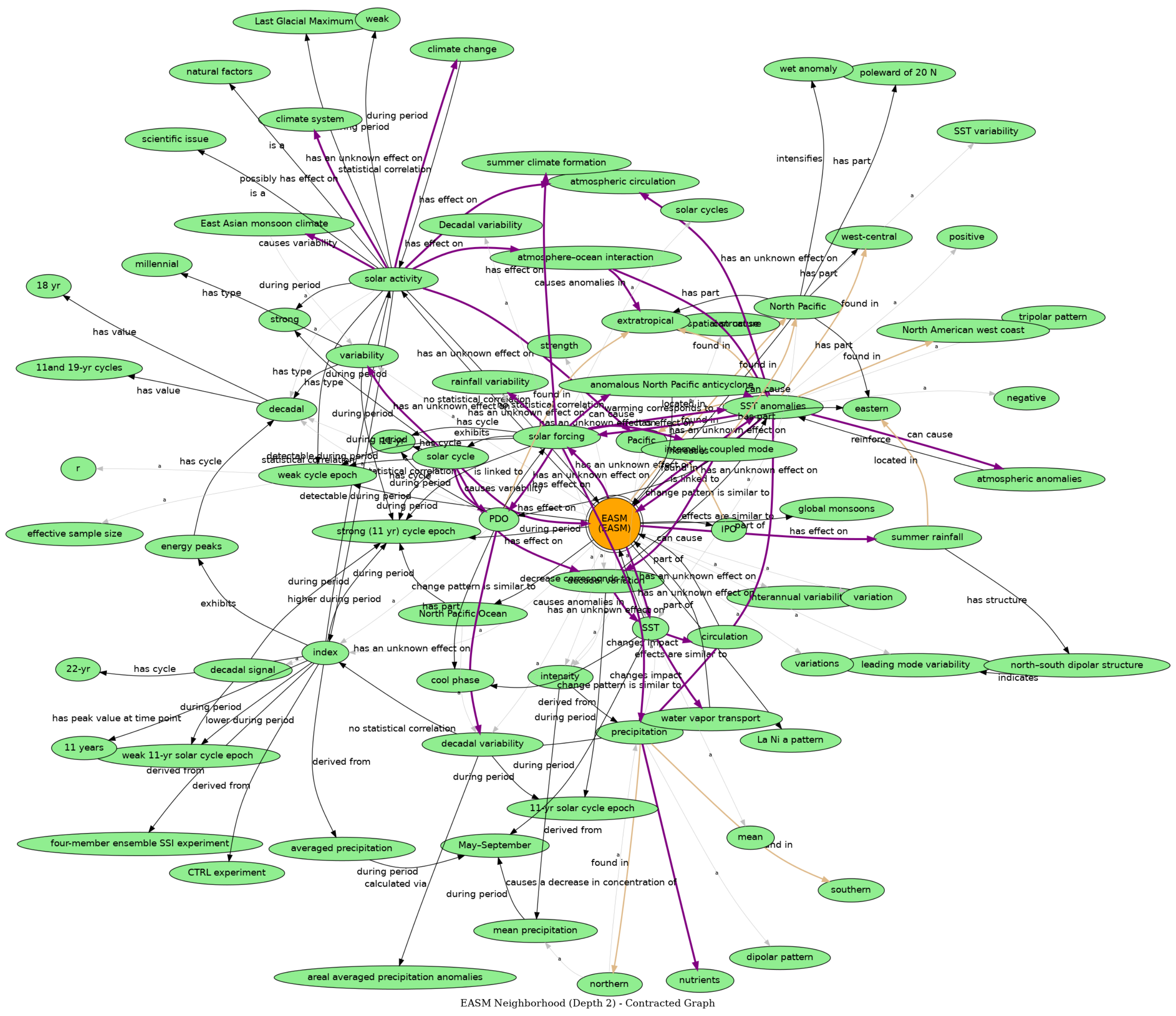

Figure 1. **Two-hop neighbourhood graph of the node "East Asian Summer Monsoon (EASM)":** This subgraph consists of 84 nodes and 156 edges, resulting in an average degree of 3.71. It was extracted from a larger knowledge graph created from two research papers [2, 3] containing 1,282 nodes and 1,278 edges with an overall average degree of 1.99.

The Challenge: Climate Change, Misinformation, and Fragmented Knowledge

- **Critical Climate Impacts:** Anthropogenic climate change severely affects global ecosystems, weather and society, threatening biodiversity and sustainability of the planet [4].
- **Misinformation and Denial:** Widespread misinformation and climate denial, often from non-scientific sources, hinder effective, evidence-based responses despite strong scientific agreement [5, 6].
  Fragmented Knowledge: Climate research findings are extensive, but scattered across many publications (Figure 2e), which makes it difficult to see the full picture and understand complex connections.
- **Our Approach: Building a Knowledge Graph (KG):** We are constructing a KG from scientific literature on climate change:
  - Improve access to reliable scientific information, helping to counter misinformation.
  - Connect findings from diverse sources, reducing knowledge fragmentation.
  - Help discover new scientific insights and relationships within the data.

Data Foundation: Analyzing the Climate Change Literature

Our research is grounded in a substantial corpus of **~200,000 scientific papers** from high-impact climate change journals [7], including: *International Journal of Climatology, Journal of Climate, Journal of Geophysical Research: Atmospheres* and *Nature Climate Change*.

Initial analysis of a 10,000-paper sample revealed:

- Over **2.4 million sentences**.
- Over **15 million noun phrases** (i.e. potential entities).
- Over **5.5 million potential triples**.

A high diversity of potential entities and relations, highlights the **need for automated, domain-specific extraction methods**.

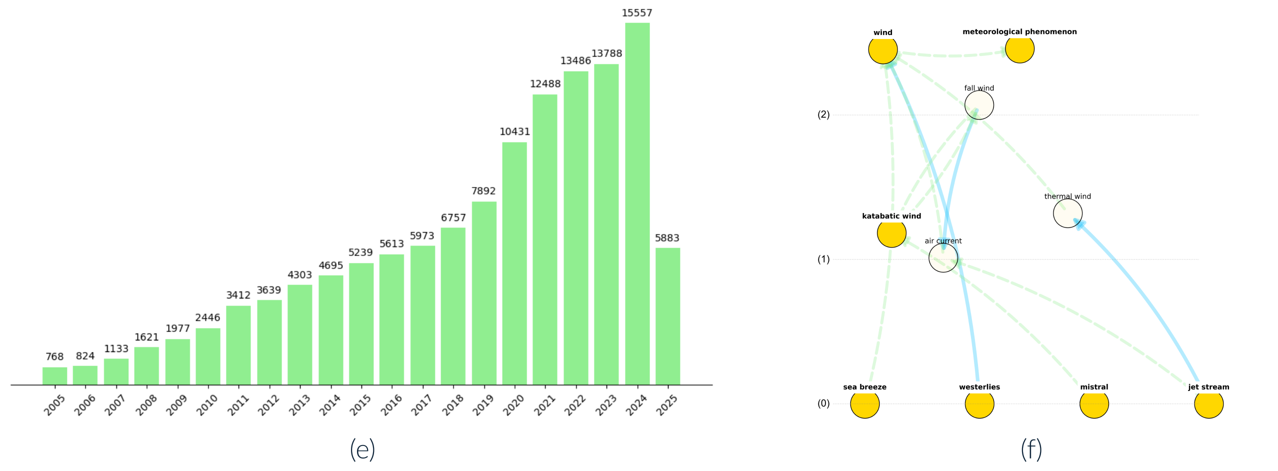

Figure 2. **WOS publication trend (e) and entity type discovery example (f)**

Methodology: Building the Climate Change Knowledge Graph

Automated Knowledge Graph construction from text centrally involves Natural Language Processing (NLP) techniques: **Named Entity Recognition (NER)** and **Relation Extraction (RE)**. Our methodology (Figure 3) involves systematic processing of climate science literature to build a structured knowledge resource:

1. **Collecting Language Resources:** Collecting climate research texts and identifying 655 core scientific terms as a foundation.
2. **Identifying Key Climate Concepts (Entities):** Automatically recognizing and categorizing important climate-related entities (e.g. *EASM*, CO_2) within texts.
3. **Discovering Relations Between Entities:** Systematically extracting how climate entities influence or interact (e.g. CO_2 *emissions* **contribute to** *global warming*).
4. **Creating an Integrated Climate Knowledge Resource:** Assembling all identified concepts and their relations into a comprehensive, queryable Knowledge Graph to support research and understanding.

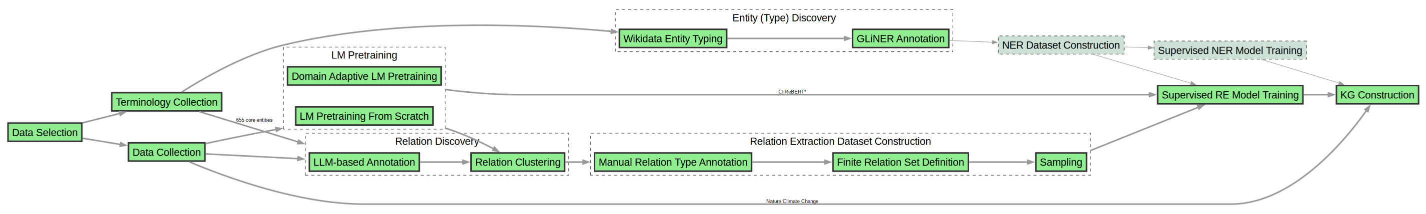

Figure 3. **Research methodology**

Key Achievement: Entity Type Discovery

We identified **21 essential Named Entity Recognition (NER) categories** tailored for climate research using Wikidata hierarchy (Figure 2f). These categories allow us to precisely identify key concepts discussed in scientific texts.

Label	Examples	Label	Examples	Label	Examples
Ecosystem	*forest, grassland*	Disease	*infection, COVID-19*	Natural Phenomenon	*mutation, runoff*
Energy Source	*electricity, fuel*	Location	*China, site*	Field of Study	*agriculture, climatology*
Natural Disaster	*drought, flood*	Measurement Unit	*km, L*	Mathematical Expression	*function, P*
Meteorological Phenomenon	*rainfall, ENSO*	Physical Phenomenon	*emission, evaporation*	Measuring Device	*rain gauge, fMRI*
Quantity	*1, concentration*	Chemical	*DNA, CO*	Geographical Feature	*land, watershed*
Astronomical Object	*star, Venus*	Time Period	*Summer, 2019*	System	*climate system, network*
Body of Water	*North Atlantic, river*	Organization	*PNAS, NIH*	Satellite	*MODIS, satellite*

Table 1. **Frequent examples of 21 NER types**

Validation with Existing Climate Ontologies: Our approach was validated against established resources. Notably, the SWEET ontology demonstrated strong alignment, covering 57.25% of our 655 core climate terms and 65.38% of the 26 initially derived candidate NER types.

Conclusion and Outlook

Developing a Knowledge Graph (KG) is key to structuring climate research. Our discovery of **21 domain-specific entity types** establishes a vital foundation using a minimally supervised method.

Next Steps: Complete the relation extraction pipeline to build the full Climate Research KG.

Impact: This comprehensive KG will support research, inform policy and education, and help counter misinformation about climate change by providing accessible, verified scientific evidence.

References

[1] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction," *ACM Comput. Surv.*, vol. 56, Nov. 2023.

[2] C. A. Simeone, G. Scott, R. A. Navarro, and D. Procter, "Clinical observations associated with phenobarbital serum monitoring to manage epilepsy in a california sea lion with domoic acid toxicosis," *Oceans*, vol. 3, no. 3, pp. 331–339, 2022.

[3] C. Jin, J. Liu, B. Wang, M. Yan, and L. Ning, "Decadal variations of the east asian summer monsoon forced by the 11-year insolation cycle," *Journal of Climate*, vol. 32, no. 10, pp. 2735 – 2745, 2019.

[4] O. Hoegh Guldberg, D. Jacob, M. Taylor, M. Bindi, S. Brown, I. A. Camilloni, A. Diedhiou, R. Djalante, K. L. Ebi, F. Engelbrecht, *et al.*, "Impacts of 1.5°c global warming on natural and human systems," in *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*, pp. 175–312, Cambridge, UK and New York, NY, USA: Cambridge University Press, 2018.

[5] P. Andre, T. Boneva, F. Chopra, and A. Falk, "Globally representative evidence on the actual and perceived support for climate action," *Nature Climate Change*, 2 2024.

[6] C. S. Areni, "Motivated reasoning and climate change: Comparing news sources, politicization, intensification, and qualification in denier versus believer subreddit comments," *Applied Cognitive Psychology*, vol. 38, no. 1, 2024.
All Open Access, Hybrid Gold Open Access.

[7] A. Poleksić and S. Martinčić-Ipšić, "Towards dataset for extracting relations in the climate-change domain," in *Joint Proceedings of the 3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG) and Data Quality Meets Machine Learning and Knowledge Graphs (DQMLKG), co-located with the Extended Semantic Web Conference (ESWC 2024)*, vol. Vol-3747, pp. 9–15, CEUR-WS, May 2024.