Proposal to ??:
Persistent Real-Time Event Extraction from Multiple Sources
DRAFT

Co-principal investigators:
Timothy C. Haas, Director, *Profitable Biodiversity*
& ??

1 Summary

Distributed algorithms and web systems will be developed to scrape and parse events from multiple sources including the dark web, social media, and ecological monitoring systems.

Amount requested: $100,000
Grant Duration: 1 year
Number of supported students: 1

2 Deliverables

2.1 Social media scraping and parsing

The taxonomy-based social media parser of Haas and Ferreira (2018) will be improved. This system acquires events (actions) reported in social media that concern a particular topic (Haas 2021). Specific improvements will include the following.

1. The system's current shallow parsing algorithm will be improved in order to increase its chance of detecting taxonomic actions from irregularly-structured HTML, and also reduce its false detection rate. This new algorithm will execute parallel parsing tasks asynchronously so that it scales when run on a cluster computer.
2. Algorithms will be developed to extract new groups (agents) and regions to add to the system's taxonomy.

3. An inference capability will be added to the parser that allows other actions to be inferred from the presence of related actions in a social media text. For example, upon reading one of the actions `seize_crate_of_smuggled_wildlife`, or `rhinos_shot_dead`, infer the additional action `smuggle_wildlife` or `poach_for_cash`, respectively.

4. The taxonomic-action matching algorithm will be improved so that essential words need to be present in a text before a match is declared. For example, for an action to match the taxonomy's action `hyenas_maul_some_goats`, the word `hyenas` needs to be present in the text.

2.2 Testing

The tools developed in this proposal will be tested by running them on social media posts that mention illegal fishing along the Pacific coast of Mexico (Felbab-Brown 2022).

3 Budget

3.1 Salary Support

Item	Date		Amount
1.	summer 2024	Professor ??: 1 month	$1,000.00
2.	2024	Student	$5,000.00
Total Direct Costs			$0
Indirect costs		(45% of Direct Costs)	$0
Total Salary Support			$0

3.2 Natural Language Parsing (NLP) Software

Item	Date		Amount
1.	Summer 2024	purchase of NLP software from ??	$1,000.00
Total direct costs			$0
Indirect costs (45% of Direct Costs)			$0

Total software and data	$0

3.3 High Performance Computer (HPC) Time

Item	Date		Amount
1.	2024	?? hours of computer time	$1,000.00
Total direct costs			$0
Indirect costs (45% of direct Costs)			$0
Total HPC			$0

3.4 Travel

Item	Date	Description	Amount
1.	2024	Demonstrations of the new event extraction system at AI and machine learning conferences	$10,000.00
Total travel			$20,000.00

3.5 Total

Total requested	$100,000.00

4 Rationale

4.1 Social media scraping and parsing

The new approach to web scraping due to Huber et al. (2022) will be extended to persistently search and scrap the web for wildlife trafficking content.

A political-ecological action that is mentioned in a social media post consists of `date`, `source`, `country`, `actor`, `action`, `subject`. Event argument extraction is related to

political-ecological action extraction but does not attempt to extract `date, source, country, subject` from the social media text. The shallow semantic parsing framework of Luo et al. (2019), and the multiple event extraction algorithm of Liu et al. (2018) will be starting points for this new NLP system. And, the asynchronous, distributed NLP work of Zielonka et al. (2018), and the distributed event extraction framework of Kan et al. (2020) will be starting points for modifying this system so that it takes advantage of an asynchronous, parallel computing architecture.

5 References

Felbab-Brown, V. (2022), "Organized Crime is Taking Over Mexican Fisheries," *Order from Chaos,* Brookings Institution, February 21.
`https://www.brookings.edu/blog/order-from-chaos/2022/02/22/organized-crime-is-taking-`

Haas, T. C. (2021), "The First Political-Ecological Database and its Use in Episode Analysis," *Frontiers in Conservation Science, section: Planning and Decision-Making in Human-Wildlife Conflict and Coexistence*, 2:707088. DOI: 10.3389/fcosc.2021.707088
`https://www.frontiersin.org/article/10.3389/fcosc.2021.707088`

Haas T. C. and Ferreira, S. M. (2018), "Finding Politically Feasible Conservation Strategies: The Case of Wildlife Trafficking," *Ecological Applications*, 28(20): 473-494.

Huber, S., Knoll, F., and Döller, M. (2022), "A Pipeline-Oriented Processing Approach to Continuous and Long-term Web Scraping," In *Proceedings of the 17th International Conference on Software Technologies (ICSOFT 2022)*, pp. 441-448. DOI: 10.5220/0011275100003266, ISBN: 978-989-758-588-3; ISSN: 2184-2833.
`https://www.scitepress.org/PublishedPapers/2022/112751/112751.pdf`

Kan, Z., Mi, H., Yang, S., Qiao, L., Feng, D., and Li, D. (2020), "A Distributed Event Extraction Framework for Large-Scale Unstructured Text," *2020 IEEE International Conference on Joint Cloud Computing (JCC)*, 102-108. DOI: 10.1109/JCC49151.2020.00024

Liu, X., Luo, Z., and Huang, H. (2018), "Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, October 31-Novemer 4, Association for Computational Linguistics, 1247-1256.

Luo, Z., Sui, G., Zhao, H., and Li, X. (2019), "A Shallow Semantic Parsing Framework for Event Argument Extraction," KSEM 2019, LNAI, 1176: 88-96, Springer, Cam. DOI: 10.1007/978-3-030-29563-9_9

Zielonka, M., Kuchta, J., and Czarnul, P. (2018), "From Sequential to Parallel Implementation of NLP using the Actor Model," *Proceedings of 39th International Conference on Information Systems Architecture and Technology*, ISAT 2018, (eds.) L. Borzemski, J. Swiatek, and Z. Wilimowska, Springer. DOI: 10.1007/978-3-319-99981-4_15