

# The News Angler Project: Exploring the Next Generation of Journalistic Knowledge Platforms

Marc Gallofré Ocaña<sup>1</sup>, Andreas L. Opdahl<sup>1</sup>, Sergej Stoppel<sup>2</sup>,  
Bjørnar Tessem<sup>1</sup>, and Csaba Veres<sup>1</sup>

<sup>1</sup> Dept. Information Science and Media Studies, University of Bergen, Norway  
{Andreas.Opdahl,Marc.Gallofre,Bjornar.Tessem,Csaba.Veres}@uib.no  
<http://i2s.uib.no>

<sup>2</sup> Wolftech Broadcast Solutions, Bergen, Norway  
[sergej@wolftech.no](mailto:sergej@wolftech.no)  
<https://wolftech.no/>

**Abstract.** The *News Angler* project aims to support journalists in finding new and unexpected connections and angles in the news. The project therefore explores how recent artificial intelligence (AI) techniques — such as knowledge graphs, natural-language processing (NLP) and machine learning (ML) — can support high-quality journalism that exploits big and open data sources. A central contribution is *News Hunter*, a series of prototype journalistic knowledge platforms (JKPs).

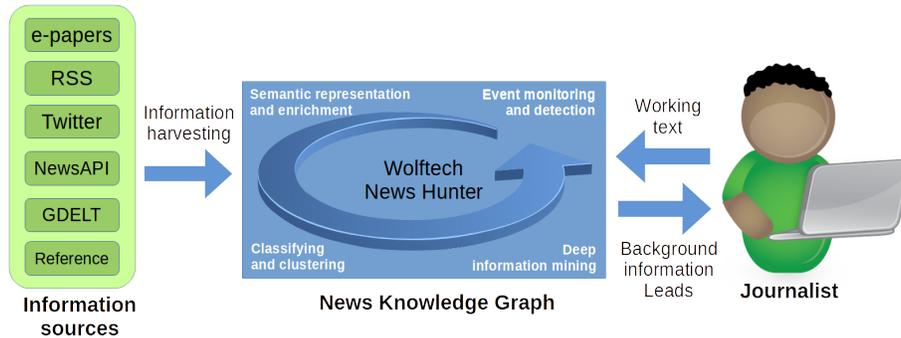
**Keywords:** Computational journalism · Journalistic Knowledge Platforms · Knowledge Graphs · Natural-Language Processing · Machine Learning · Artificial intelligence.

## 1 Background

Journalism is in crisis [15], but advances in AI offer new opportunities [11]. Journalists today have access to a wealth of digital information from news aggregators, social media, open data, and other sources. These data can be automatically analysed, integrated, organised, prepared, and stored with high semantic precision. Theories and techniques from artificial intelligence (AI) – such as knowledge graphs, natural-language processing (NLP) and machine learning (ML) – can be used to classify, label, cluster, detect events, and otherwise process streams of potentially news-relevant information in new and powerful ways.

## 2 The News Hunter Platform

The Intelligent Information Systems (I2S) group at the University of Bergen is therefore collaborating with Wolftech Broadcast Solutions AS, a supplier of newsroom systems for the international market. Together, we are developing *News Hunter* (Fig. 1), a series of prototype journalistic knowledge platforms



**Fig. 1.** The News Hunter platform uses a central news knowledge graph to make online information sources more useful for working journalists.

(JKPs) [13, 3, 4]. News Hunter continuously harvests potentially news-relevant texts from social media and other online sources in real time; analyses them semantically using NLP and ML; and represents them in a central knowledge graph. This knowledge graph can be analysed further, enriched with data from the Linked Open Data (LOD) cloud, and used to suggest leads, angles, background information and other relevant insights to working journalists.

The first generation (2000-2010) of knowledge platforms that targetted journalistic purposes focussed on exploiting semantic-web technologies for information organisation by supporting tasks such as knowledge integration and automated semantic annotation and search. The second generation (2010-2020) moved on to exploit LOD technologies for more proactive tasks, such as identifying news events, automating fact checking, and suggesting relevant background information for journalists [4].

### 3 The News Angler Project

The News Angler project aims to go beyond the state of art by exploring a third generation of big-data ready journalistic knowledge platforms [7] that support journalists in finding new and unexpected connections and angles in unfolding news events [12, 14]. The project analyses and represents the harvested texts in knowledge graphs, aiming to support deep and innovative information mining.

### 4 Techniques

The platform currently runs in a big-data ready OpenStack cloud on a cluster of more than 20 nodes managed by Terraform and Ansible, making it easy to expand on demand. It is organised as a micro-service architecture of Docker containers running as a Docker Swarm [9]. Source texts are stored in Cassandra along with their metadata represented in JSON-LD format. The central news

knowledge graph is stored in a distributed Blazegraph triple store. Kafka is used as a publication/subscription message queue and as a stream database, with metadata controlled by Zookeeper. MongoDB manages configuration and other metadata relevant for the platform and services.

Specialised services [9, 6] are implemented in Python 3.8-3.9, using FastAPI to provide external APIs [9]. Real-time harvesters have been implemented for tweets, RSS and NewsAPI messages, and for GDELT events. On the natural-language processing (NLP) side, DBpedia Spotlight is central for named-entity linking (NEL) [2]. In addition, Docker containers have been written that use Neuralcoref for co-reference resolution; Spacy NEL and Kolitsas et al's [10] end-to-end neural network for named-entity linking; and OpenNRE and SPN4RE [16] for relation extraction. A specialised service has been implemented to lift GDELT events into RDF format. The small knowledge graphs that result from these semantic analyses are represented as JSON-LD and submitted to Kafka. An updating service automatically retrieves the JSON-LD items from Kafka and inserts them into the news knowledge graph stored in Blazegraph [9, 6], which provides a SPARQL endpoint for powerful semantic information retrieval.

Ongoing work includes aggregation of low-level GDELT events into composite news-level events and services for suggesting *local news angles* on the harvested and analysed news items. Earlier versions of the platform [4] also provided basic services for topic identification, classification, event detection, and a graphical user interface that we plan to include in the current platform.

## 5 Results

In parallel with technical exploration of the prototype platform, the News Angler project has made a range of related theoretical contributions, some in final form and others in progress, including: a review of challenges and opportunities for journalistic knowledge platforms (JKPs) [8]; usage scenarios for JKPs; a survey of knowledge graphs in the news; a survey of named-entity linking (NEL) techniques [2]; a proposal for combining lifting techniques [1]; a reference architecture for JKP [9]; OWL ontologies for representing news items, news events [14] and news angles [12]; analyses and representations of selected computational news angles [12]; analogical reasoning over news angles [17, 18]; and a discussion of privacy issues [5].

## Conclusion

This short paper has presented an overview of the News Angler project and the evolving News Hunter platform, including its motivation, ambition, techniques, and results, which we will continue to expand in the final year of the project. We believe the techniques and theories that the project provides have potential impact beyond computational journalism, by offering an alternative path to the surface similarity-based search and recommendation services that shape the personal and organisational information spaces that surround us today.

## References

1. Al-Moslmi, T., & Gallofré Ocaña, M. (2020). Lifting news into a journalistic knowledge platform.
2. Al-Moslmi, T., Gallofré Ocaña, M., Opdahl, A.L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8, 32862-32881.
3. Berven, A., Christensen, O.A., Moldeklev, S., Opdahl, A.L., & Villanger, K.J. (2018). News Hunter: building and mining knowledge graphs for newsroom systems. *NOKOBIT*, 26, 1-11.
4. Berven, A., Christensen, O.A., Moldeklev, S., Opdahl, A.L., & Villanger, K.J. (2020). A Knowledge-Graph Platform for Newsrooms, *Computers in Industry* 123, 103321, Elsevier.
5. Gallofré Ocaña, M. (2020). Data Privacy in Journalistic Knowledge Platforms. Proc. SKG4J 2020: the 1st International Workshop on Semantic and Knowledge Graph Advances for Journalism, CIKM 2020 workshops.
6. Gallofré Ocaña, M (2021). Identifying Events from Streams of RDF-Graphs Representing News and Social Media Messages. In *The Semantic Web: ESWC 2021 Satellite Events* (pp. 186-194). Springer, Cham.
7. Gallofré Ocaña, M., Nyre, L., Opdahl, A.L., Tessem, B., Trattner, V., & Veres, C. (2019). Towards a Big Data Platform for News Angles. In *Proc 4th Norwegian Big Data Symposium (NOBIDS), CEUR Workshop Proceedings, Vol-2316, ISSN 1613-0073*.
8. Gallofré Ocaña, M., & Opdahl, A.L. (2020). Challenges and Opportunities for Journalistic Knowledge Platforms. *Proceedings of the CIKM 2020 Workshop*, 2699, 43.
9. Gallofré Ocaña, M. & Opdahl, A.L. (2021). Developing a Software Reference Architecture for Journalistic Knowledge Platforms. Proc. SAML'2021, European Conf. on Software Architecture (ECSA 2021) Workshops.
10. Kolitsas, N., Ganea, O. E., & Hofmann, T. (2018). End-to-end neural entity linking. arXiv preprint arXiv:1808.07699.
11. Miroshnichenko, A. (2018). AI to bypass creativity. Will robots replace journalists?(The answer is “yes”). *Information*, 9(7), 183.
12. Motta, E., Daga, E., Opdahl, A. L., & Tessem, B. (2020). Analysis and Design of Computational News Angles. *IEEE Access*, 8 pp. 120613–120626.
13. Opdahl, A. L., Berven, A., Alipour, K., Christensen, O. A., & Villanger, K. J. (2016). Knowledge graphs for newsroom systems. Short paper in *Norsk konferanse for organisasjoners bruk av IT (NOKOBIT, Vol. 24, No. 1)*.
14. Opdahl, A. L., & Tessem, B. (2021). Ontologies for finding journalistic angles, *Software and Systems Modeling*, 17 pp.
15. Siles, I., & Boczkowski, P. J. (2012). Making sense of the newspaper crisis: A critical assessment of existing research and an agenda for future work. *New media & society*, 14(8), 1375-1394.
16. Sui, D., Chen, Y., Liu, K., Zhao, J., Zeng, X., & Liu, S. (2020). Joint entity and relation extraction with set prediction networks. arXiv preprint arXiv:2011.01675.
17. Tessem, B. (2019). Analogical New Angles from Text Similarity. In *Artificial Intelligence XXXVI, Lecture Notes in Computer Science (LNCS)*, Springer, pp. 449-455.
18. Tessem, B., & Opdahl, A. L. (2019). Supporting journalistic news angles with models and analogies. In *13th International Conference on Research Challenges in Information Science (RCIS 2019)* (pp. 1-7). IEEE.