

# Challenges and Opportunities for Journalistic Knowledge Platforms

Marc Gallofré Ocaña<sup>a</sup>, Andreas L. Opdahl<sup>a</sup>

<sup>a</sup>University of Bergen, Fosswinkelsgt. 6, Postboks 7802, 5020 Bergen, Norway

## Abstract

Journalism is under pressure from loss of advertisement and revenues, while experiencing an increase in digital consumption and user demands for quality journalism and trusted sources. Journalistic Knowledge Platforms (JKPs) are an emerging generation of platforms which combine state-of-the-art artificial intelligence (AI) techniques such as knowledge graphs, linked open data (LOD), and natural-language processing (NLP) for transforming newsrooms and leveraging information technologies to increase the quality and lower the cost of news production. In order to drive research and design better JKPs that allow journalists to get most benefits out of them, we need to understand what challenges and opportunities JKPs are facing. This paper presents an overview of the main challenges and opportunities involved in JKPs which have been manually extracted from literature with the support of natural language processing and understanding techniques. These challenges and opportunities are organised in: stakeholders, information, functionalities, components, techniques and other aspects.

## Keywords

Newsroom, Knowledge Graph, Digitalization, Overview

## 1. Introduction

Journalism is under pressure from loss of advertisement and revenues, in combination with competing online distribution channels that stream free content, while experiencing an increase in digital consumption and readers who demand quality journalism and trusted sources [1]. Information is no longer consumed from a single newspaper. Instead, readers have access to and can contrast fresh and first-hand information sources available on the internet and social media at any time.

News organisations are constantly adapting their business models to digital media innovations, to improve information quality, competitiveness and growth [2]. Journalistic Knowledge Platforms (JKPs) are an emerging type of platform that combines state-of-the-art artificial intelligence (AI) techniques such as knowledge graphs and natural-language processing (NLP); and exploit news and social media information over the net in real-time, using linked open data (LOD), encyclopaedic sources and news archives to construct knowledge graphs and provide fresh and unexpected information to journalists, helping them to dive deeply into information, events and story-lines. JKPs are increasingly driving innovation

and transforming newsrooms, leveraging information technologies to increase the quality and lower the cost of news production. In order to drive research and design JKPs that allow journalists to get most benefits out of them and support newsrooms with better solutions, we need to understand the challenges and opportunities that JKPs present for both users and developers. To do so, we have reviewed the research literature in light of our own experience with developing News Hunter [3, 4, 5], a series of JKP prototypes in collaboration with a developer of newsroom tools for the international market.

This paper presents a synthesis of the challenges and opportunities for journalistic knowledge platforms that we have found in the literature, hopefully describing the most central factors that are driving development of JKPs today. These factors have been grouped into six categories: stakeholders, information, functionalities, components, techniques and other aspects. We conclude that JKPs offer many opportunities for effective production of high-quality journalism, real-time information, enriched background information, and multilingual and cross-platform solutions for monitoring worldwide multimedia output, by offering solutions to problems such as language independence, complex newsrooms workflows, and disperse information. Central challenges include leveraging pre-news information from social media and multimedia sources, precise semantic lifting and enrichment of texts, scaling semantic technologies to big data, and detecting and reasoning over events.

*Proceedings of the CIKM 2020 Workshops,  
October 19-20, Galway, Ireland.*


EMAIL: Marc.Gallofre@uib.no (M. Gallofré Ocaña);

Andreas.Opdahl@uib.no (A.L. Opdahl)

ORCID: 0000-0001-7637-3303 (M. Gallofré Ocaña);

0000-0002-3141-1385 (A.L. Opdahl)

© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This paper is organised as follows: Section 2 summarises the methodology used for screening the challenges and opportunities. Section 3 briefly reviews the research literature. Section 4 explains the coding process. Sections 5 to 10 synthesise the main challenges and opportunities for each factor respectively – stakeholders, information, functionalities, components, techniques and other aspects.

## 2. Method

Our research method consists of four steps: Firstly, we selected the most relevant research papers that we have identified in our previous studies on JKPs architectures and news angles [4, 6, 7, 8, 9, 10, 11, 12]. From these selected papers we manually extracted claims, i.e., sentences that express potential challenges or opportunities.

Secondly, a purposive sampling was conducted independently by two expert coders (the authors). The coders generated multiple codes for each extracted claim and the codes were cleaned with the support of NLP and NLU techniques (i.e., Damerau-Levenshtein distance [13], word2vec [14], and Wordnet [15])<sup>1</sup>. From the resulting cleaned codes, we selected the most representative ones as preliminary codes and divided them into categories.

Thirdly, based on the preliminary codes, claims were independently coded once again by both authors. This time, the coders were allowed to code each claim with multiple codes for each category. The coding agreement was estimated using Gwet’s  $AC_1$  [19] inter-rater reliability coefficient with nominal ratings. Because coders were allowed to not to code, to compute the Gwet’s  $AC_1$ , empty codes were not treated as missing values, instead, they were treated as if they were coded as “undefined”. Hence, to compute the contingency tables for multiple codes we applied the following rule: the agreement between coders A and B only happens between correctly matching codes ( $A \cap B$ ) and the other codes ( $A \Delta B$ ) were matched with missing values and treated as disagreements.

Finally, when both coders agreed on the final codes for each claim, challenges and opportunities were extracted from each claim following the assigned codes.

---

<sup>1</sup>Implemented in python with support of Scikit-learn [16], NLTK [17], SpaCy [18] and other libraries.

## 3. Reviewed papers

After a broad survey of the literature, we selected eleven papers describing five research projects related to JKPs as the starting point of our review: NEWS [20, 21], EventRegistry [22], NewsReader [23, 24, 25], SUMMA [26, 27, 28, 29] and ASRAEL [30].

NEWS is a project, in collaboration with the Spanish Agencia EFE and the Italian ANSA news agencies, that makes use of semantic technologies to improve news agencies’ workflows, productiveness and revenues by focusing on the annotation, intelligent information retrieval and user interface aspects [21]. EventRegistry is focused on collecting news articles, identifying and extracting information about events, and summarising and visualising them [22]. NewsReader extracts information about what, who, where, when from multilingual news articles and represents events in time using RDF in a knowledge graph, allowing users to find networks of actors along time [25]. SUMMA collaborates with BBC Monitoring and Deutsche Welle to develop a multilingual and multimedia platform using state-of-the-art NLP techniques to monitor internal and external media work and provide data journalism services [27]. ASRAEL aggregates news articles and leverages the Wikidata knowledge base to describe and cluster news events and provides information retrieval tools to interact with the resulting news representations [30].

## 4. Coding process

In the purposive sampling step, we extracted 322 claims from the related literature and marked them up using 406 codes. After cleaning and tidying up the initial codes, we identified six top-level categories which we divided into 62 sub-categories to be used for preliminary coding. The following six top-level categories were used:

- Stakeholder: the agent that the challenge or opportunity is for. The agent can be either a technical agent or social agent.
- Information: the information needed to meet the challenge or exploit the opportunity.
- Functionality: the service or functionality that the platform should offer to meet a challenge or exploit an opportunity.
- Component: the part of a platform that must be

created or improved to meet the challenge or exploit the opportunity.

- **Technique:** the IT solution used to meet the challenge or exploit the opportunity.
- **Other aspects:** another type of concern that the challenge or opportunity involves, such as customer heterogeneity, performance or maintenance.

We computed the inter-rater agreement for the preliminary coding with the  $AC_1$  coefficient for each category: 0.77 for Stakeholders, 0.65 for Components, 0.71 for Techniques, 0.71 for Aspects, 0.72 for Information types and 0.57 for Functionalities. The average  $AC_1$  is 0.69 with a standard deviation of 0.063, which according to Landis-Koch and Altman's benchmark scales, express an acceptable agreement among coders [19]. Finally, the assigned codes were discussed between and agreed on by the two coders.

## 5. Stakeholders

Stakeholders are agents that represent the forces and interests that drive the future of JKPs. The identified sub-categories of stakeholders are: general user, news professional, fact checker, archivist, ICT professional, audience, customer, researcher, news agency, public organisation and technical agent.

*General users* interact with services provided by the JKPs or newsrooms. These can be divided between the internal users that belong to newsrooms and the external ones. The internal users are *news professionals* like journalists who use JKPs for creating histories [20]; *fact checkers* who conduct an essential task in combating with fake news and misinformation [28]; *archivists* who maintain up-to-date the ontology and news archives [20]; and *ICT professionals* and knowledge engineers who represent those users involved in the development and maintenance of JKPs [21]. Whereas, the external users are the *audience* [22]; the *customers* to whom new agencies offer services; and *researchers* who investigate JKPs or analyse data, as in the SUMMA project where “[political scientists want] to perform data analyses based on large amounts of news reports” [27, p. 2].

The organisations influencing the JKPs are: the *news agencies*, including newsrooms; the *public organisations* which are those governmental agencies that interact with or consume services from newsrooms' JKPs, as in the SUMMA project which “provides media monitoring and analysis services to the BBC

own newsrooms, the British government, and other subscribers” [27, p. 1]; and the organisations that are responsible for controlling news media standards, vocabulary and ontologies (e.g., IPTC organisation<sup>2</sup>), which are indirectly influencing JKPs because the work of many news agencies and JKPs depends on those standards, as in the NEWS project where “most of the NewsCodes defined by IPTC do not have alternative versions in different languages, only in English” [20, p. 9].

Finally, the *technical agent*, which is a stakeholder that represents the JKPs and any system or technical infrastructure in newsrooms that support or interact with JKPs. A particular subtype of technical agent are the external systems that communicate with newsroom systems, like the information systems of potential customers [20].

## 6. Information

JKPs cover the whole information pipeline from gathering information and news creation to knowledge exploitation and distribution. Our study identified the following sub-categories of information to be considered in JKPs: news content, textual data, multimedia data, data format, metadata, LOD, events and information needs.

News agencies produce both textual and multimedia *news content* which have to be managed and distributed to their customers and audience [21, 20]. As *textual data* we consider the raw text in form of news articles, documents, markup files, PDF, web pages, biographies, history and geopolitical data of countries, reports, social media feeds and social blogs. Whereas, as *multimedia* we consider live broadcast, spoken content, photographs, audio and video. Besides, news agencies produce contents in different formats like plain text, Information Interchange Model (IIM), News Industry Text Format (NITF), NewsML and RDF [20].

*Metadata* is used to annotate and manage the produced content. Metadata can describe e.g., author, language, creation timestamp, location, keywords, category, provenance, priority, urgency, status, updates, rights, interest, description or media type. JKPs use *Linked Open Data* (LOD) to annotate and enrich content using semantic vocabularies and leveraging knowledge bases, as in the ASRAEL project where they “leverage the Wikidata knowledge base to produce semantic annotations of news articles” [30, p. 1].

---

<sup>2</sup><https://iptc.org/>

News agencies create stories describing *events* and deliver them to their customers and audience [21], making the events the central *information need*. Despite that, social stakeholders have other *information needs*: General users are interested in knowing who, what, with whom, where and when events took place, networks of timeline actors implications, find the events of a certain type or in a certain place, obtain facts and retrieve evidence [24]. News professionals need access to news agencies' archives and knowledge bases for documentation purposes, find connections from past events, follow histories and identify emerging topics [20, 23, 27]. While customers have different information needs mainly depending on their business or interests, e.g., "the press cabinet of a company is usually interested in news items talking about the company or its rivals, whereas a sports TV channel is interested mostly in news items describing sports events" [20].

## 7. Functionalities

JKPs provide different functionalities to their users. We identified twelve main sub-categories of functionality: news creation, verification, source selection, monitoring, knowledge discovery, trends, alert, summarisation, clustering, personalisation, business support and content management.

News professionals use the JKPs for the *news creation* process. JKPs guide journalists in writing up their stories, support them with contextual background knowledge for those stories [21], provide means for comparing current events with other similar events [30], and facilitate access to previous work for creating similar content for a different audience, region or language [27]. JKPs also support news professions with *verification* tasks like fact checking, provenance [24], rights and authorship management [20, 21], which are typically time-consuming tasks for news professions as explained in "manual verification of claims is a tedious task, that consumes a lot of time and effort from journalists and professional fact-checkers" [28, p. 1].

*Source selection* and *monitoring* functionalities are two common functionalities across the studied JKPs, which harvest and store content from internal and external sources and monitor them in real-time. By doing this, JKPs relieve journalists from these time-consuming tasks, as it was happening in the BBC where "each of its ca. 300 journalist monitors up to four live broadcasts in parallel, plus several other information sources such as social media feeds" [27,

p. 1].

*Knowledge discovery* is one of the most attractive functionalities of JKPs. Knowledge discovery allows users to obtain news insights, analysis and relevant information, like in NewsReader where it "increases the user understanding of the domain, facilitates the reconstruction of news story lines, and enables users to perform exploratory investigation of news hidden facts" [24, p. 1]. Other interesting functionalities among JKPs are: *trends* used to discover emerging topics, long-term developments and changes in events over time [22, 25]; *alerts* to keep users up-to-date with the last incoming items [26]; *summarisation* of news histories and events to provide additional insights [22]; *clustering* of story-lines and events [27]; and *personalisation* of both the JKPs and its functionalities according to users' preferences and profiles [21].

JKPs provide functionalities to news agencies and newsrooms organisation and workflows. JKPs are used as *business support* systems to manage internal newsrooms output; monitor what is being broadcast, produced and covered [27]; overcome limitations in newsrooms' workflows; and improve productivity and revenues [20]. Another functionality provided by the JKPs is the *content management* which allows news agencies to produce, store, organise, manage, maintain and distribute the content and metadata produced every day [20].

## 8. Components

JKPs rely on different components to fulfil its functionalities and support users. We split JKP components into five sub-categories: input, processing, storage, interaction and output.

As input, we consider the different sources of content and information used in JKPs that are relevant for stakeholders. The textual and multimedia sources are sources of interest. However, not all analysed projects treat the information in the same way or use the same information types, like ASRAEL which only uses the title and first paragraph to represent the events [30]; and not all contents receive the same interest by news professionals, as in SUMMA which considers "entertainment programming such as movies and sitcoms, commercial breaks, and repetitions of content (e.g., on 24/7 news channels) [...] of limited interest to monitoring operations" [27, p. 1].

The processing components cover tasks from harvesting and annotating input sources to processing and lifting them, following an ETL process (i.e., Ex-



tract, Transform, Load). Input sources are harvested using different components, each with a specific purpose: *harvesting*, *translating*, *filtering* and *transcribing*. A common characteristic of the analysed projects is that source selection and monitoring functionalities are conducted in real-time by *harvesting* information sources [22, 23, 27]. The harvested content is then *translated* [27] and *filtered* according with the different stakeholders' interests and needs. Spoken content is *transcribed* [27] and images are textually described [21].

JKPs use specific components to automatically *annotate* the harvested content with metadata to support functionalities like business support, content management and personalisation [20]. The annotated content is typically processed by different components which are organised in an *NLP pipeline*. The NLP pipeline processes the content through state-of-the-art NLP and NLU modules to perform linguistic tasks [25, 24]. These tasks are focused on capturing and extracting the different information types described in section 6. Both the results of the *NLP pipeline* and the annotated content are disambiguated and represented semantically using *lifting* components. The lifting component links the semantic representation of news items to a knowledge base, for examples an RDF-based knowledge graph [25], and enriches the semantic interpretations with facts from external knowledge bases, for example from the LOD cloud [24, 30].

The JKP *storage infrastructure* is normally composed of an *archive*, a *knowledge base* and an *ontology*. The *archive* stores news articles, biographies, reports [25] and other textual and multimedia items; the *knowledge base* is where the lifted semantic representations of news items are stored and enriched with external information [24]; and the *ontology* is used to represent the structure of the news items, leveraged information, metadata and vocabulary [20].

JKP users interact with the previous components mainly using three types of interaction components: *front-ends*, *tools* and *query engines*. JKPs provide front-end components [21] to allow stakeholders to access the system functionalities; tools which offer features to journalists when creating news articles or to general users when interacting with the system, like money converters or dictionaries [20]; and query engines that allow users to query, analyse or visualise the database through APIs [27].

News agencies use two types of distribution components for delivering content to their audience and customers [20]: *push* and *pull*. Push components offer interfaces where information consumers can select and subscribe to streams of news [20], whereas with

*pull* components, news agencies offer interfaces to access, browse and query their repositories [20].

## 9. Techniques

Techniques used in JKPs can be grouped in eight sub-categories: semantic technology, fact extraction, conceptual model, reasoning, network analysis, event analysis, NLP and training.

*Semantic technology* is used to support functionalities like knowledge discovery, news creation, verification, clustering, trends, and content management. Semantic technologies support knowledge discovery by providing means for lifting news items, and disambiguating, enriching and leveraging them with information from external knowledge bases [21, 25] – processes carried by the lifting, ontology and knowledge base components; news creation, by providing systems and vocabulary to automatically annotate news in annotation components [21]; and verification, by combining semantic technologies with the lifting and knowledge base components and linking factual claims to its sources and external knowledge bases [24, 27]. Semantic technologies and semantic representation techniques facilitate clustering news items and events [30], and detecting trends and story lines [24]. Moreover, semantic technologies provide shared semantic resources and formats which are used to support content management and facilitate conceptual interoperability [25].

*Fact extraction* techniques extract facts from news items and link them to facts in external knowledge bases (e.g., Wikidata, Wikipedia). These techniques are used to provide functionalities like verification and knowledge discovery [27] and are common features of lifting, knowledge base and query components.

*Conceptual models* provide vocabularies and ontologies which are used in conjunction with semantic technologies to support and standardise functionalities like content management and personalisation. Ontologies can be used for defining user interests and preferences based on the provided vocabulary or as shared models [20]. Conceptual models are applied in distribution, lifting, annotation, ontology, query, knowledge base and source components.

Both conceptual models and semantic technologies facilitate the usage of other techniques like reasoning, network analysis and event analysis. These techniques support functionalities like knowledge discovery, clustering and trends, and are applied in the lifting, knowledge base, ontology and annotation components. *Reasoning* techniques abstract and infer new knowledge

from news items, events and temporal aspects [24, 25]. *Network analysis* is used to find networks of actors and organisation implications through different events and time [24]. *Event analysis* is applied to detect, identify and annotate the events described in news [21, 20].

The above techniques are supported by *NLP* tasks like named entity detection, role detection, topic detection, temporal expression normalisation, temporal relation detection, factual claims extraction, natural language understanding [25, 29, 27]. These *NLP* tasks, among others, are also used in JKPs' functionalities such as knowledge discovery, content management, summarising, verification, trends, clustering, query, lifting and annotation. In order to obtain optimal results from the *NLP* tasks, different *training* techniques have to be used over extensive news corpus [30].

## 10. Other aspects

Stakeholders, information, functionalities, components and techniques are influenced or affected by additional concerns of various types. We organised these other aspects into the following sub-categories: standards, proprietary, human factors, customers heterogeneity, big data, multilingual, timeliness, quality, software architecture, performance, maintenance, and legacy.

Before moving into JKPs, news agencies used their terms, categories and vocabularies to describe their items. Yet, the interoperability between news agencies and customers was difficult. The usage of *standards* like IPTC news codes and media topics, semantic vocabularies, NAF and RDF improved the interoperability between news agencies and other stakeholders [20].

JKPs keep track of *proprietary* news information like authorship, copyrights and sources [21, 20] as a part of the content management functionalities. Property information is used as metadata in annotation components and provides provenance and reliability information [24, p. 4].

There are different *human factors* influencing JKPs and stakeholders. Before JKPs, news professionals were performing many processes by hand like news tagging, verification tasks, fact searching, finding related articles, and source monitoring. Performing these tasks manually is time-consuming, error-prone, consumes a lot of efforts, and reduces the amount and precision of the added metadata [21, 20, 28, 22]. Therefore, customers have to manually filter irrelevant content received from news agencies, creating an information overload problem which is contrary

to the information relevance that customers expect from news agencies [21, 20]. Moreover, because the difficulty of manually monitoring and finding related articles from other news providers, the audience, customers and news professions can get biased or incomplete information [22].

*Customers are heterogeneous*, they have different information needs and use different systems to interact with news agencies [20].

According to our study, JKPs deal with *big data* requirements like volume, velocity, variety: The AS-RAEL project estimated that "the number of collected articles ranges between 100.000 and 200.000 articles per day" and collected "news articles from around 75.000 news sources" [22, p. 1]. NewsReader used an archive that "contains billions of articles, biographies, and reports" [25, p. 1]. The SUMMA platform "[was] able to ingest 400 TV streams simultaneously" [27, p. 6].

Other information aspects that JKPs deal with are the *multilingual* and *timeliness* data aspects. Information and news production are created in multiple languages (e.g., Catalan, Norwegian, Spanish, English, Italian, French, Portuguese and Chinese) and need to be translated, transcribed and delivered to customers and audiences in their languages of preference [20, 27, 25, 30]. The timeliness aspect refers to the temporal aspect of events, thus news professionals, audience and customers want to receive the information as soon as it is generated [21] and reconstruct story-lines or histories over time [24, 27].

*Quality* of the results and outputs of JKPs are summarised in "news agencies are required to provide fresh, relevant, high-quality information to their customers" [21, p. 1] and ignoring these quality requirements can imply economic losses for customers [20].

Aspects concerning technical agents and their components include the *software architecture*, *performance*, *maintenance* and relation of JKPs with other systems. The software architecture of JKPs should consider scalability to deal with big data requirements [21, 24, 27], distribution to run its components and systems over multiple machines [20, 26], components independence so they can be used for other purposes [26], interoperability between components and systems [20, 25], and performance for reducing the processing and distributing time of information and live feeds [21, 24]. Manual maintenance is a time-consuming and error-prone task [20] which is automated with JKPs to keep the JKP and ontology up-to-date [26]. As JKPs communicate with customers systems, *legacy* components and other

newsroom systems, JKPs need to be designed to facilitate the integration with other technologies and systems [20, 26].

## 11. Conclusion

JKPs are a new type of platforms which offer many opportunities for newsrooms and journalists by combining AI techniques such as knowledge graphs, LOD and NLP to improve and facilitate the production of high-quality journalism. We collected challenges and opportunities that JKPs present and organised them into six categories that we assume are important for the evolution of JKPs (stakeholders, information, functionalities, components, techniques and other aspects).

JKPs offer new opportunities for consuming and interacting with news by providing enriched content from external sources like Wikipedia or Wikidata to stakeholders seeking relevant information, such as news professionals and general audiences. News texts are enriched with additional information about, e.g., involved actors, places and organisations, the connections with other news and related events. Information and data sources in JKPs are no longer split along dispersed and disconnected repositories as it happens in traditional solutions. Instead, the information pieces are connected by the knowledge graph. JKPs enhance functionalities like news creation and content management. News creation is improved with background information providing journalists with better information for their stories. Automatic metadata annotation and the usage of standards like IPTC relieve archivists from manually annotating news and improve the content management capabilities of JKPs and newsroom workflows. Knowledge graphs in JKPs bring new forms of representing news-related content and exploiting it. Techniques like network analysis, event analysis and reasoning improve the background information and knowledge discovery in JKPs while opening new research questions for researchers. JKPs can use standards such as RDF, IPTC's media topics and semantic vocabularies which simplify the interoperability and understanding between news agencies and stakeholders. The most highlighted opportunities that have been identified in the literature include event detection and analysis over time, real-time and up-to-date trustworthy information, access to enriched background information for supporting news creation, multilingual and multimedia cross-platform solutions, and tools for monitoring worldwide media output and internal newsrooms production.

On the other hand, providing one-size-fits-all JKP solutions for all possible stakeholders is challenging, because of their diversity and differing information needs. Newsworthy information comes from diverse news sources like pre-news information from social media or multimedia sources such as TV news programs. Leveraging these information sources is a complex task which requires new techniques to distinguish potentially newsworthy information from non-relevant content and extract information from multimedia items like images or videos. Summarising and presenting news-related information in JKPs like background information, events in time or actor networks to users with different information needs and skills is not a trivial task. JKPs consist of different components which interact together and with external components that need to be integrated in JKPs systems. Extracting precise semantic representations of and reasoning over relations and time remain open research questions. JKPs deal with big data, but some semantic technologies, reasoning and AI techniques are not yet ready for it. Among the reviewed JKPs, the most common challenges are problems such as language independence, multiple news channels, complex newsrooms workflows, dispersed and diverse information, lack of facts, and integration with legacy and customer systems.

After reviewing the literature, we have realised that there is not a clear definition and agreement about what constitutes an event. The event concept is used in different ways in the literature, from a handshake between two actors to bigger events like the Spanish Civil War or events in between such as a trial process. In this study, we have only reviewed five JKP-related research projects, although they are the five most central ones we have found. Hence, we may have omitted important issues that were not represented or brought up in these projects. We are therefore planning to extend the number of considered projects through a systematic literature review and contrast and expand our findings with published works on data and digital journalism. A logical continuation of this expanded study is the formal identification and modelling of goals, requirements and use cases for JKPs, which we did not find yet in the literature. Furthermore, we plan to formalise a reference framework for JKPs and continue the development of our JKP identified to validate and integrate our findings.

## Acknowledgments

This work has been supported by the Norwegian Research Council IKTPLUSS project 275872 *News Angler*, which is a collaboration with Wolftech AB, Bergen, Norway.

## References

- [1] PwC, Global entertainment & media outlook 2019–2023, 2020. URL: <https://www.pwc.com/gx/en/industries/tmt/media/outlook.html>.
- [2] J. Vázquez Herrero, S. Direito-Rebollal, A. S. Rodríguez, X. García, *Journalistic Metamorphosis: Media Transformation in the Digital Age*, Springer International Publishing, 2020. doi:10.1007/978-3-030-36315-4.
- [3] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, News hunter: building and mining knowledge graphs for newsroom systems, in: NOKOBIT–Norsk konferanse for organisasjoners bruk av informasjonsteknologi, volume 26, 2018. URL: <https://ojs.bibsys.no/index.php/Nokobit/article/view/548/467>.
- [4] M. Gallofré Ocaña, L. Nyre, A. L. Opdahl, B. Tessem, C. Trattner, C. Veres, Towards a big data platform for news angles, in: 4th Norwegian Big Data Symposium (NOBIDS) 2018, 2018, pp. 17–29. URL: <http://ceur-ws.org/Vol-2316/paper1.pdf>.
- [5] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, A knowledge graph platform for newsrooms, *Computers in Industry* (2020). To appear.
- [6] B. Tessem, A. L. Opdahl, Supporting journalistic news angles with models and analogies, in: 2019 13th International Conference on Research Challenges in Information Science (RCIS), IEEE, 2019, pp. 1–7. doi:10.1109/RCIS.2019.8877058.
- [7] A. L. Opdahl, B. Tessem, Towards ontological support for journalistic angles, in: *Enterprise, Business-Process and Information Systems Modeling*, Springer International Publishing, 2019, pp. 279–294. doi:10.1007/978-3-030-20618-5\_19.
- [8] T. A. A. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl, B. Tessem, Detecting newsworthy events in a journalistic platform, in: *The 3rd European Data and Computational Journalism Conference*, 2019, pp. 3–5.
- [9] B. Tessem, Analogical news angles from text similarity, in: *Artificial Intelligence XXXVI*, Springer International Publishing, 2019, pp. 449–455. doi:10.1007/978-3-030-34885-4\_35.
- [10] A. L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, *Software and Systems Modeling* (2020) 1–17.
- [11] E. Motta, E. Daga, A. L. Opdahl, B. Tessem, Analysis and design of computational news angles, *IEEE Access* (2020).
- [12] T. Al-Moslmi, M. Gallofré Ocaña, Lifting news into a journalistic knowledge platform, in: *Proceedings of the CIKM 2020 Workshops*, Galway, Ireland, 2020. To appear.
- [13] F. J. Damerau, A technique for computer detection and correction of spelling errors, *Commun. ACM* 7 (1964) 171–176. doi:10.1145/363958.363994.
- [14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013, pp. 3111–3119. URL: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.
- [15] G. A. Miller, Wordnet: A lexical database for english, *Commun. ACM* 38 (1995) 39–41. doi:10.1145/219717.219748.
- [16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* 12 (2011) 2825–2830. URL: <http://www.jmlr.org/papers/v12/pedregosa11a>.
- [17] S. Bird, E. Loper, E. Klein, *Natural language processing with python*, O’Reilly Media, Inc., 2009.
- [18] M. Honnibal, I. Montani, *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017. To appear.
- [19] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*, Advanced Analytics, LLC, 2014.
- [20] N. Fernández, D. Fuentes, L. Sánchez, J. A. Fisteus, The news ontology: Design and applications, *Expert Systems with Applications* 37 (2010) 8694 – 8704. doi:10.1016/j.eswa.2010.06.055.
- [21] N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, *News: Bringing se-*



- mantic web technologies into news agencies, in: *The Semantic Web - ISWC 2006*, 2006, pp. 778–791. doi:10.1007/11926078\\_56.
- [22] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: Learning about world events from news, in: *Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion*, Association for Computing Machinery, 2014, pp. 107–110. doi:10.1145/2567948.2577024.
- [23] M. Kattenberg, Z. Beloki, A. Soroa, X. Artola, A. Fokkens, P. Huygen, K. Verstoep, Two architectures for parallel processing for huge amounts of text, in: *Proceedings of Language Resources and Evaluation Conference (LREC)*, European Language Resources Association (ELRA), 2016, pp. 4513–4519. URL: <https://www.aclweb.org/anthology/L16-1714>.
- [24] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, *Journal of Web Semantics* 37-38 (2016) 132–151. doi:10.1016/j.websem.2015.12.004.
- [25] P. Vossen, R. Aggeri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, M. Rospocher, R. Segers, Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news, *Special Issue Knowledge-Based Systems*, Elsevier 110 (2016) 60–85. doi:10.1016/j.knosys.2016.07.013.
- [26] U. Germann, R. Liepins, D. Gosko, G. Barzdins, Integrating multiple NLP technologies into an open-source platform for multilingual media monitoring, in: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, Association for Computational Linguistics, 2018, pp. 47–51. doi:10.18653/v1/W18-2508.
- [27] U. Germann, R. Liepins, G. Barzdins, D. Gosko, S. Miranda, D. Nogueira, The SUMMA platform: A scalable infrastructure for multi-lingual multi-media monitoring, in: *Proceedings of ACL 2018, System Demonstrations*, Association for Computational Linguistics, 2018, pp. 99–104. doi:10.18653/v1/P18-4017.
- [28] S. a. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, Z. Marinho, Automated fact checking in the news room, in: *The World Wide Web Conference, WWW '19*, Association for Computing Machinery, 2019, pp. 3579–3583. doi:10.1145/3308558.3314135.
- [29] P. Paikens, G. Barzdins, A. Mendes, D. C. Ferreira, S. Broscheit, M. S. Almeida, S. Miranda, D. Nogueira, P. Balage, A. F. Martins, Summa at tac knowledge base population task 2016, in: *Proceedings of the Ninth Text Analysis Conference (TAC)*, 2016. URL: <https://tac.nist.gov/publications/2016/participant.papers/TAC2016.summa.proceedings.pdf>.
- [30] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, X. Tannier, Searching news articles using an event knowledge graph leveraged by wikidata, in: *Companion Proceedings of The 2019 World Wide Web Conference, WWW '19*, Association for Computing Machinery, 2019, pp. 1232–1239. doi:10.1145/3308560.3316761.