Journalistic Knowledge Platforms: from Idea to Realisation

Marc Gallofré Ocaña

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN

Journalistic Knowledge Platforms: from Idea to Realisation

Marc Gallofré Ocaña

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 23.10.2023

Year: 2023

Title: Journalistic Knowledge Platforms: from Idea to Realisation

Name: Marc Gallofré Ocaña

Print: Skipnes Kommunikasjon / University of Bergen

Marc Gallofré Ocaña
ORCID: 0000-0001-7637-3303

The most exciting phrase to hear in science,
the one that heralds new discoveries, is not
“Eureka!” (I found it!) but “That’s funny...”

credited to ***Isaac Asimov***

Scientific environment

This work was carried out at the Intelligent Information Systems (I2S) research group at the University of Bergen in collaboration with Wolftech Broadcast Solutions AS and supported by the Norwegian Research Council IKTPLUSS project 275872. It was supervised by Prof. Andreas L. Opdahl and co-supervised by Prof. Bjørnar Tessem.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Prof. Andreas L. Opdahl, for his invaluable guidance, support and advice. His expertise and mentorship have been instrumental in navigating the challenges of pursuing a PhD. I am truly grateful for his commitment to my success.

In addition, I would like to thank my colleagues, who have not only provided me with academic support but also become friends. Through our lively discussions, engaging debates, and enjoyable gatherings around food, lunches and dinners, I have been able to take a much-needed break from research and connect with others in the field, share ideas, and gain new perspectives.

I would also like to extend my thanks to past and present research fellows, colleagues, collaborators and assistants. They are the true formulas of any scientific contribution.

I would like to express my heartfelt gratitude to Stephanie and Bengji for their support and encouragement during my PhD journey. I feel truly fortunate to have them in my life, and I appreciate their support more than words can express.

Lastly, I would like to thank my family for their unconditional love, understanding and support throughout this journey. Being far away from home has been difficult and their encouragement has been a constant source of energy. I am grateful for their sacrifices, and I could not have made it this far without their unwavering belief in me.

To all of you, I am grateful for the piece of you that is reflected in this work.

Marc Gallofré Ocaña
Bergen, 2023

Abstract

Journalistic Knowledge Platforms (JKPs) are a type of intelligent information systems designed to augment news creation processes by combining big data, artificial intelligence (AI) and knowledge bases to support journalists. Despite their potential to revolutionise the field of journalism, the adoption of JKPs has been slow, with scholars and large news outlets involved in the research and development of JKPs. The slow adoption can be attributed to the technical complexity of JKPs that led news organisation to rely on multiple independent and task-specific production system. This situation can increase the resource and coordination footprint and costs, at the same time it poses a threat to lose control over data and face vendor lock-in scenarios. The technical complexities remain a major obstacle as there is no existing well-designed system architecture that would facilitate the realisation and integration of JKPs in a coherent manner over time. This PhD Thesis contributes to the theory and practice on knowledge-graph based JKPs by studying and designing a software reference architecture to facilitate the instantiation of concrete solutions and the adoption of JKPs. The first contribution of this PhD Thesis provides a thorough and comprehensible analysis of the idea of JKPs, from their origins to their current state. This analysis provides the first-ever study of the factors that have contributed to the slow adoption, including the complexity of their social and technical aspects, and identifies the major challenges and future directions of JKPs. The second contribution presents the software reference architecture that provides a generic blueprint for designing and developing concrete JKPs. The proposed reference architecture also defines two novel types of components intended to maintain and evolve AI models and knowledge representations. The third presents an instantiation example of the software reference architecture and details a process for improving the efficiency of information extraction pipelines. This framework facilitates a flexible, parallel and concurrent integration of natural language processing techniques and AI tools. Additionally, this Thesis discusses the implications of the recent AI advances on JKPs and diverse ethical aspects of using JKPs. Overall, this PhD Thesis provides a comprehensive and in-depth analysis of JKPs, from the theory to the design of their technical aspects. This research aims to facilitate the adoption of JKPs and advance research in this field.

Sammendrag

In the fulfilment of requirements of the University of Bergen for delivering the dissertation, the following text includes the abstract in Norwegian.

Journalistiske kunnskapsplattformer (JKPer) er en type intelligente informasjonssystemer designet for å forbedre nyhetsproduksjonsprosesser ved å kombinere stordata, kunstig intelligens (KI) og kunnskapsbaser for å støtte journalister. Til tross for sitt potensial for å revolusjonere journalistikkfeltet, har adopsjonen av JKPer vært treg, med forskere og store nyhetsutløp involvert i forskning og utvikling av JKPer. Den langsomme adopsjonen kan tilskrives den tekniske kompleksiteten til JKPer, som har ført til at nyhetsorganisasjoner stoler på flere uavhengige og oppgavespesifikke produksjonssystemer. Denne situasjonen kan øke ressurs- og koordineringsbehovet og kostnadene, samtidig som den utgjør en trussel om å miste kontrollen over data og havne i leverandørlåssituasjoner. De tekniske kompleksitetene forblir en stor hindring, ettersom det ikke finnes en allerede godt utformet systemarkitektur som ville lette realiseringen og integreringen av JKPer på en sammenhengende måte over tid. Denne doktoravhandlingen bidrar til teorien og praksisen rundt kunnskapsgrafbaserte JKPer ved å studere og designe en programvarearkitektur som referanse for å lette iverksettelsen av konkrete løsninger og adopsjonen av JKPer. Den første bidraget til denne doktoravhandlingen gir en grundig og forståelig analyse av ideen bak JKPer, fra deres opprinnelse til deres nåværende tilstand. Denne analysen gir den første studien noensinne av faktorene som har bidratt til den langsomme adopsjonen, inkludert kompleksiteten i deres sosiale og tekniske aspekter, og identifiserer de største utfordringene og fremtidige retninger for JKPer. Den andre bidraget presenterer programvarearkitekturen som referanse, som gir en generisk blåkopi for design og utvikling av konkrete JKPer. Den foreslåtte referansearkitekturen definerer også to nye typer komponenter ment for å opprettholde og videreutvikle KI-modeller og kunnskapsrepresentasjoner. Den tredje presenterer et eksempel på iverksettelse av programvarearkitekturen som referanse og beskriver en prosess for å forbedre effektiviteten til informasjonsekstraksjonspipelines. Denne rammen muliggjør en fleksibel, parallell og samtidig integrering av teknikker for naturlig språkbehandling og KI-verktøy. I tillegg diskuterer denne avhandlingen konsekvensene av de nyeste KI-fremgangene for JKPer og ulike etiske aspekter ved bruk av JKPer. Totalt sett gir denne PhD-avhandlingen en omfattende og grundig analyse av JKPer, fra teorien til designet av deres tekniske aspekter. Denne forskningen tar sikte på å lette vedtaket av JKPer og fremme forskning på dette feltet.

Outline

This PhD Thesis is part of the News Angler project that aimed to explore a new variety of intelligent information systems that combine artificial intelligence, big data and knowledge bases to bolster news production. The objective of the project was to provide journalists with the necessary tools to discover novel and unforeseen connections and angles in unfolding news events. As part of this project, this thesis focused on developing a software reference architecture for such software systems along with a prototype proof of concept.

This thesis represents the culmination of a design science research process, where the research contributed to 14 scientific publications and the technical development produced over 37.000 lines of code, 1.500 commits, one open-source project and inputs to various open-source libraries. The thesis is organised as follows: Chapter 1 introduces the context of the research project and the scope of the thesis. Chapter 2 provides the background theories. Chapter 3 offers an overview of the methodology and methods used in this work. Chapter 4 summarises the results and manuscripts, and Chapter 5 discusses them.

The following manuscripts are included in this thesis as part of the fulfilment of the requirements for the dissertation:

- M. Gallofré Ocaña and A. L. Opdahl, 'Supporting Newsrooms with Journalistic Knowledge Graph Platforms: Current State and Future Directions,' *Technologies*, vol. 10, no. 3, p. 68, May 2022. DOI: `10.3390/technologies10030068`

- M. Gallofré Ocaña and A. L. Opdahl, 'A Software Reference Architecture for Journalistic Knowledge Platforms,' *Knowledge-Based Systems*, vol. 276, p. 110 750, 2023, ISSN: 0950-7051. DOI: `10.1016/j.knosys.2023.110750`

- M. Gallofré Ocaña and A. L. Opdahl, 'A Blackboard Model for Parallel and Flexible Text Annotation,' Submitted

In addition to the manuscripts, this work has resulted in the following papers:

- M. Gallofré Ocaña and A. L. Opdahl, 'Challenges and opportunities for journalistic knowledge platforms,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020

- M. Gallofré Ocaña, T. Al-Moslmi and A. L. Opdahl, 'Data privacy in journalistic knowledge platforms,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020

- T. Al-Moslmi and M. Gallofré Ocaña, 'Lifting news into a journalistic knowledge platform,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020

- M. Gallofré Ocaña and A. L. Opdahl, 'Developing a software reference architecture for journalistic knowledge platforms,' in *Companion Proceedings of the 15th European Conference on Software Architecture (ECSA 2021)*, CEUR Workshop Proceedings, 2021

- M. Gallofré Ocaña, 'Identifying events from streams of rdf-graphs representing news and social media messages,' in *The Semantic Web: ESWC 2021 Satellite Events*, Cham: Springer International Publishing, 2021, pp. 186–194

- M. Gallofré Ocaña, T. Al-Moslmi and A. L. Opdahl, 'Knowledge graph semantic annotation and population with real-time events data from gdelt,' in *2022 IEEE 24th Conference on Business Informatics (CBI)*, vol. 02, 2022, pp. 65–72. DOI: `10.1109/CBI54897.2022.10050`

The following co-authored publications are related to this work:

- M. Gallofré Ocaña, L. Nyre, A. L. Opdahl, B. Tessem, C. Trattner and C. Veres, 'Towards a big data platform for news angles,' in *Norwegian Big Data Symposium*, CEUR Workshop Proceedings, 2018

- M. Albared, M. Gallofré Ocaña, A. Ghareb and T. Al-Moslmi, 'Recent progress of named entity recognition over the most popular datasets,' in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, 2019, pp. 1–9. DOI: `10.1109/ICOICE48418.2019.9035170`

- T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl and B. Tessem, 'Detecting newsworthy events in a journalistic platform,' in *The 3rd European Data and Computational Journalism Conference*, 2019, pp. 3–5

- T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl and C. Veres, 'Named entity extraction for knowledge graphs: A literature overview,' *IEEE Access*, vol. 8, pp. 32 862–32 881, 2020. DOI: `10.1109/ACCESS.2020.2973928`

- A. L. Opdahl, T. Al-Moslmi, D.-T. Dang-Nguyen, M. Gallofré Ocaña, B. Tessem and C. Veres, 'Semantic knowledge graphs for the news: A review,' *ACM Comput. Surv.*, vol. 55, no. 7, Dec. 2022. DOI: `10.1145/3543508`

- B. Tessem, M. Gallofré Ocaña and A. L. Opdahl, 'Construction of a relevance knowledge graph with application to the LOCAL news angle,' in *Nordic Artificial Intelligence Research and Development*, 2023

This work has contributed to various open-source libraries:

- RDFlib (`github.com/RDFLib/rdflib`)

- spotlight-docker (`github.com/dbpedia-spotlight/spotlight-docker`)

- pynif (`github.com/wetneb/pynif`)

- News Hunter (`git.app.uib.no/i2s/news-hunter-platform`)

Contents

Introduction

The landscape of journalism is rapidly evolving as news agencies and organisations face challenges with the decline of advertising and revenue streams [16], [17]. In addition, there is an audience that is reluctant to pay for digital content [18], [19]. The explosion of digital consumption has transformed the way people consume news. Though there has been an increase in digital consumption, traditional news sources are no longer the sole providers of news. Instead of relying on the traditional model of limited TV stations and news outlets as primary news sources, readers have now access to a vast array of free and first-hand news sources on the internet and social media platforms. Readers demand high-quality journalism [20] and seek out trusted sources [19], [21], [22], given the freedom of choice they now possess.

To meet this demand, news agencies and organisations must embrace innovation and digitalisation in order to improve news quality, competitiveness and growth [23]. Innovation and digitalisation of newsrooms are required to reduce the cost and enhance the quality of news production, transforming the way journalists and readers interact with news content and background information [24]. Consequently, newsrooms are increasingly leveraging big data and artificial intelligence (AI) techniques such as knowledge graphs and machine learning (ML) for diverse journalistic purposes [25], [26]. For instance, these and related techniques aid in identifying and contextualising newsworthy events in investigative journalism, facilitating data visualisation in digital journalism, analysing information in data journalism, automating news writing in robot journalism, and providing real-time fact-checking tools for political journalism.

With almost any type of source and data format potentially containing news-relevant information, the worldwide daily news production scales to over 100.000 articles, with social media capable of generating similar volumes in just one second. Big Data is now being utilised to support investigative journalism [25], allowing journalists access to substantial data on government and corporate activities. Data mining and visualisation tools can be used to analyse large datasets and identify patterns, trends and anomalies that may indicate misconduct or corruption. The Pandora Papers investigation[1] is one example of this, which involved the analysis of more than 11.9 million leaked documents, leading to the exposure of high-level tax evasion and money laundering schemes.

In recent years, AI-powered tools have transformed the field of journalism, as evidenced by their capacity to automate news writing, fact-checking and content creation, among other applications. With AI technology, journalists can analyse vast amounts of data in real time, flag potential sources of misinformation, and provide readers with valuable context to better understand the news. Among the most noteworthy uses of AI in journalism is the automated news writing, where AI algorithms analyse specific keywords, data, trends and events to generate news articles in real time. In addition, AI algorithms are adept at offering context to enhance readers' comprehension of the news and enabling journalists to scrutinise intricate data and social media feeds. Furthermore, AI is employed in the fight against fake news and disinformation, as it can detect and signal misleading content or be used to fact-check news with trusted knowledge bases.

Natural Language Processing (NLP), which "is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content" [27], plays a crucial role in AI-powered tools for journalism. It involves the development of algorithms, methods and models that enable computers to automatically analyse and process natural language data, such as text and speech. For example, this enables machines to generate responses, mine data, derive meaning, semantics and linguistic structures, and classify, synthesise and translate content. Recent cutting-edge advancements in NLP models such as BERT [28] and GPT-3 [29] have opened the field to new possibilities that are yet to be fully explored.

Today's NLP is mostly driven by embedding techniques, compared to the first decades of NLP which are mostly driven by pre-defined sets of vocabularies and rules [27]. Embedding techniques are used in machine learning and deep learning (DL) to represent concepts as vectors in a low-dimensional space. These vector spaces provide sub-symbolic representations through mathematical models that position concepts in the space according to similarity or other relations. Embeddings are sub-symbolic representations with a stochastic component. They require large amounts of data to be meaningful and are hard to explain to humans, but even large vector spaces and stores can be efficiently managed by computers. Well-known techniques for word and text embedding are word2vec [30] and transformers [31] like BERT [28] and GPT-3 [29].

Knowledge graphs are topical and popular choice for representing the derived news-relevant information. Knowledge graphs capture and abstract knowledge using graph-based data models. Nodes in the graph represent entities of interest and edges signify the relationships between them [32]. With its focus on concepts and relations, knowledge graphs provide powerful symbolic representations of the world around us. To achieve this

[1] Pandora Papers investigation, available at: `www.icij.org/investigations/pandora-papers`

level of sophistication, knowledge graphs rely on ontologies and logic rules that define the semantics and terms of the graph. This makes it possible to reason about the data and integrate it seamlessly with other sources of knowledge. And because it relies on symbolic representations rather than raw data, the knowledge graph is particularly well-suited for handling heterogeneous and dynamic data. Unlike other approaches that require vast amounts of data to be meaningful, the symbolic representations created by the knowledge graph are immediately comprehensible and easy for humans to understand.

To news production, knowledge graphs offer a valuable tool for capturing and organising news-relevant information about topics, people, organisations, events and other entities of interest [14]. Knowledge graphs provide contextual information to news stories, enabling journalists to gain a more comprehensive understanding of a topic or event by integrating information from various sources into a single graph. One of the remarkable features of knowledge graphs is their ability to enable semantic integration, thereby enabling flexible data and schema evolution. This entails that knowledge graphs can be adapted over time to accommodate new data and changing requirements, without the need for a complete overhaul of the underlying data model. In addition, the use of graph query languages simplifies the exploration of intricate relationships through arbitrary-length paths. To illustrate this, a knowledge graph may encompass details concerning the companies involved in a merger, the key actors in a political scandal, and the contextual background and related developments of a breaking news story.

As one might expect, both research and industry are in agreement regarding the significance and challenges posed by the upcoming artificial intelligence systems, encompassing a vast array of domains [33]. In particular, future AI systems must be semantically coherent, explainable and trustworthy. To achieve these objectives, they must be able to integrate sub-symbolic deep learning, symbolic knowledge representation and logical reasoning [34]. Notably, knowledge graphs are an apt choice for knowledge representation and reasoning [32], in conjunction with neural networks for sub-symbolic AI. Since the world is constantly evolving, these systems must integrate continuous-learning techniques, ensuring that the deep-learning and machine-learning models remain current and updated.

A new variety of intelligent information system has emerged that melds artificial intelligence (AI), big data and knowledge bases in order to bolster news production [1]. I refer to this type of system as *Journalistic Knowledge Platform* (JKP). JKPs are capable of harvesting and parsing news and social media data in real time, while simultaneously leveraging vast encyclopaedic sources. The news-relevant information is represented semantically and incorporated into knowledge bases, utilising linked open data (LOD) [35] and employing AI techniques such as natural language processing [6]. These knowledge bases are subsequently leveraged with data analysis, reasoning and information retrieval techniques, providing meaningful background and newsworthy information to journalists [12]. Furthermore, JKPs assist both journalists and readers in delving deeper into the intricacies of news-relevant information, events and storylines [36]. JKPs are generally built with a variety of mechanisms for interacting with the system, such as live feeds, alerts and search capabilities. Since JKPs aggregate and represent personal data from a multitude of sources, they must also implement privacy policies and protocols [5]. In addition, to combat the proliferation of fake news and misinformation, JKPs must manage the provenance of news sources and facilitate the identification thereof. As a topical choice, I focused in particular on the JKPs that utilise knowledge graphs [32] and semantic technologies [37] for representing knowledge.

In the research literature, several semantic knowledge-graph based JKPs have been proposed. In broad terms, these JKPs can be divided into two groups: the earlier JKPs (which prevailed until circa 2010) that primarily concentrated on realising the concept of the Semantic Web [38] in newsrooms, and the more recent JKPs (post-2010) that combine semantic technologies [37] with approaches in machine and deep learning.

The earlier JKPs used semantic technologies and ontologies to automate the metadata annotation process, integrate knowledge bases and formalise and standardise media standards. They employed ontologies within NLP pipelines together with LOD to automatically annotate news archives and feeds with rich metadata regarding a diverse array of topics, keywords, categories and other news-relevant information (e.g., persons, places, organisations, sentiments and relationships). For instance, the pioneering project known as *PlanetOnto* [39] focused on delivering an integrated and personalised knowledge management system to enable semantic retrieval and search capabilities for news archives. *Neptuno* [40] developed tools to create, maintain and explore news archives. *AnnoTerra* [41] integrated earth science data sources to augment the news feeds from NASA Earth Observatory[2] using knowledge bases. *SemNews* [42] focused on the automation of metadata annotation to facilitate semantic search and monitoring of RSS feeds. *Hermes* [43] proposed a framework for searching and classifying news stories to bolster decision-making capabilities. The BBC[3] employed knowledge graphs and LOD to create connections across news articles, enhance their content management system and deliver personalised news recommendations [44], [45]. *NEWS* [46] automated metadata annotation of news and images and provided intelligent information retrieval services using semantic technologies for the *Agencia EFE*[4] and *Agencia ANSA*[5].

Recent developments in JKPs have been focused on the identification and analysis of events and advancing machine and deep learning techniques to support journalism. A

[2] NASA Earth Observatory: an online publishing agency of the NASA (National Aeronautics and Space Administration) of the United Stated of America on discoveries about the environment, Earth systems and climate.

[3] BBC (British Broadcasting Corporation): the national broadcaster of the United Kingdom and one of the world leading news agencies.

[4] Agencia EFE: the largest Spanish news agency and global leader in Castilian language news.

[5] Agencia ANSA (Agenzia Nazionale Stampa Associata): the largest Italian news agency.

common thread among these efforts and some earlier examples is their handling of big data. For instance, *EventRegistry* [47] developed a tool that can collect nearly 200.000 news articles every day from 75.000 multilingual sources, identify and extract news-relevant information about events, and summarise and visualise them. *NewsReader* [48] presented a platform that utilises NLP pipelines to extract news-relevant information about events from multilingual news streams. This platform can represent events temporally using knowledge graphs that can reveal, for example, networks of actors and their implications over time. The platform was tested with almost 2.5 million news articles and extracted over 1.1 billion triples from these articles. Reuters[6] developed a real-time platform capable of analysing around 12 million tweets per day from Twitter [49], [50]. Reuters' platform identified and verified newsworthy events before these were reported by other news agencies. *SUMMA* [51], in collaboration with LETA[7], BBC Monitoring[8] and Deutsche Welle[9], developed a multilingual and multimedia platform that employs NLP techniques to monitor internal and external live media, including TV and radio broadcasts, and provides services to data journalists. *INJECT* [52] developed a tool, in collaboration with Adresseavisen[10], AFP[11] and The Globe and Mail[12], that supports journalists by providing creative angles on news stories. *ASRAEL* [53], in collaboration with AFP, presented a system for aggregating news articles utilising the Wikidata to describe and cluster events from a corpus of over 2 million articles.

All these aspects make JKPs a particularly intricate and multifaceted breed of big-data, knowledge-centric and intelligent systems. News organisations have come to rely on multiple independent and task-specific production systems. Depending on multiple systems leads to a greater resource footprint and increased costs for coordinating developer teams and providers, as well as maintaining and updating the systems. This poses a threat to lose control over data and knowledge and potential vendor lock-in situations. Often provided as Software as a Service (SaaS) by third parties, these systems lack common data repositories and representations, and thus, deprive organisations of potential opportunities for exploiting news-relevant information [1]. This situation can be averted with a well-designed system architecture that facilitates the integration and expansion of JKPs in a coherent manner over time. A software reference architecture (SRA) "is a generic architecture for a class of systems that is used as a foundation for the design of concrete architectures from this class" [54]. An SRA defines the fundamental software elements and data flows required for developing the functionalities of a complex system and it encapsulates the best practices for its design and implementation. To news

[6] Reuters: one of the largest news agencies in the world.

[7] LETA: the main news agency in Latvia.

[8] BBC Monitoring: a division of the BBC for monitoring and reporting on world-wide media.

[9] Deutsche Welle: a German international broadcasting agency.

[10] Adresseavisen: a daily newspaper in Trondheim and one of the oldest in Norway

[11] AFP (Agence France-Presse): a French international news agency.

[12] The Globe and Mail: a Canadian newspaper.

organisations, an SRA would provide a blueprint and associated advice for evolving its many current systems towards a cohesive, comprehensive and integrated JKP. However, to date, no reference architecture for such software systems has so far been proposed.

This PhD Thesis is part of the News Angler project [10] which aimed to go beyond the previous generations of JKPs by exploring a third generation of big-data and AI-ready JKPs. The initial objective of the project was to provide journalists with the necessary tools to discover novel and unforeseen connections and angles in unfolding news events. The project focused on identifying and formalising news angles using semantic technologies and ontologies [14], [15], [36]. Aligned with the initial aim, in this PhD Thesis, I focused on designing an SRA for JKPs and addressed the research question of: *What is a good software reference architecture for Journalistic Knowledge Platforms?* To address this question, I followed a design science research approach and developed a prototype of a JKP (Prototype). Part of the work of this thesis consisted of redesigning previous prototypes of JKPs [55] into a new platform as proof of concept to evolve and explore the proposed SRA.

Therefore, this thesis contributes to the field in multiple connected ways. Firstly, it establishes a literature on the concept of Journalistic Knowledge Platform and provides the first-ever definition of this term [1], [4] (Manuscript I). This literature can inform research and practice on designing JKPs and deriving requirements to inform concrete architectural decisions. Secondly, it proposes a software reference architecture to facilitate the adoption and expansion of JKPs [2], [7] (Manuscript II). The proposed architecture addresses the requirements identified in the literature and emphasises on the third generation of JKPs. Finally, it provides an example of how the SRA for JKPs can be realised into a concrete platform, explores in detail how to adapt the NLP pipelines to satisfy the future requirements of JKPs [3] (Manuscript III), and investigates other applications of JKPs [6], [8], [9] (Publication IV, Publication V and Publication VI).

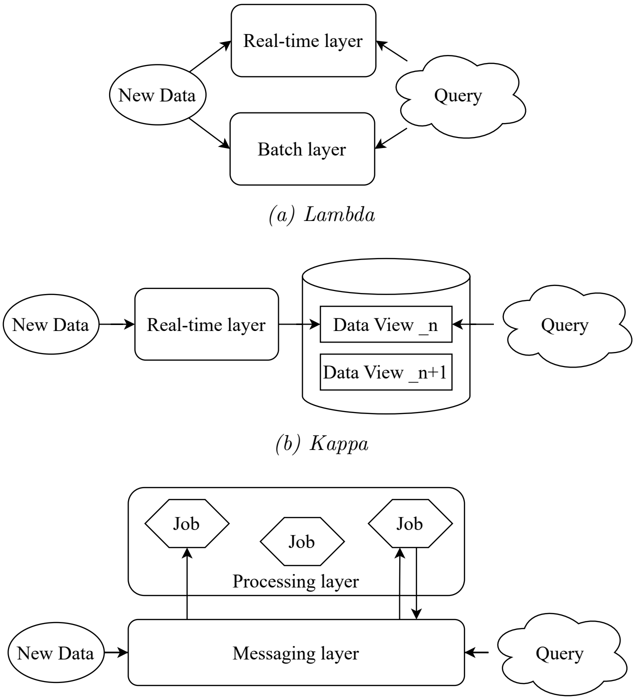
These contributions will inform journalists and newsrooms as they are among the early adopters of integrated knowledge platforms. This thesis may also inform forthcoming big-data AI systems that unite deep learning and knowledge graphs and support evolving knowledge in other knowledge-intensive domains.

Background

This chapter provides the reader with a comprehensive grasp of the theoretical basis that underpins this research. It covers diverse areas such as big-data architectures, microservices, the blackboard model, machine learning (ML), natural language processing (NLP), and knowledge graphs.

Big-Data Reference Architectures

In the field of big-data processing and analysis, there are a few well-known architecture patterns, such as the Lambda, Kappa, and Liquid architectures. The Lambda architecture [56], as illustrated in Figure 2.1a, consists of a batch processing layer and a real-time processing layer. The batch layer stores and processes historical data in batches, while the real-time layer processes current data in real time. The data is duplicated on both layers and must be combined to yield the results. The batch layer periodically reprocesses all the data, including the most recent, to ensure that the results remain updated. The Kappa architecture [57], as shown in Figure 2.1b, offers a simplified version of the Lambda architecture that eliminates the batch layer and processes all data as a stream in real time. It maintains a single view of the data that is continuously updated with the latest results. The view is only fully reprocessed and replaced with a new one when there is a change in the logic of the system. The Liquid architecture [58] was introduced as an alternative to the Lambda and Kappa architectures. It consists of two layers, the messaging layer and the processing layer, as depicted in Figure 2.1c. The processing layer is a hybrid layer that combines the real-time processing of the Kappa architecture with the batch processing of the Lambda architecture. In the processing layer, the data is shared among isolated jobs that process data incrementally and perform different tasks. The massaging layer facilitates data sharing across the jobs.



(a) Lambda

(b) Kappa

(c) Liquid

Figure 2.1: Big-data architectures

According to existing big-data architecture reviews [59]–[64], only four reference architectures for big data [59], [63], [65], [66] have considered semantic technologies. Among them, *LMS* [65] provides a middleware for sensor data and the Internet of Things (IoT). *SOLID* [66] adapted the principles of the Lambda processing architecture to RDF for gathering, storing and serving big data in real time. *Bolster* [59] extends the Lambda architecture with a new semantic layer for representing machine-readable metadata, which is distinct from JKPs that represent the data semantically. *SmartLAK* [63] focuses on supporting learning analytic services and defines components for validation and inference based on ontologies. However, none of these architectures addresses semantic data enrichment, streaming live data or continuously (re-)training machine learning models. Four proposed architectures [67]–[70] have considered the maintenance and updating of ML models and define specific components for storing and training them, but none of these architectures incorporates semantic technologies like knowledge graphs and ontologies. Overall, all these architectures focus on immutable data, and while some architectures integrate semantic technologies, no proposed architecture handles evolving data and the curation of knowledge representations. Hence, none of them provides a suitable starting point for an SRA for JKPs. A new architecture is necessary to address the challenges of JKPs and provide a foundation for handling big data in a semantic and evolving context.

Microservices

Microservices is a software development architectural pattern that breaks down large and complex applications into smaller independent services that typically communicate via well-defined APIs, as opposed to creating monolithic applications that are composed of only one single executable unit [71]. Every functionality of the system is deployed as a service and often independent from the others. Services in a microservice system are self-contained, loosely coupled, reusable and specialised, allowing them to be developed, deployed and scaled independently. Furthermore, services are technology neutral, meaning that each component can be written in different programming languages; use distinct technologies; and still communicate with the other components, as long as they adhere to the same protocols and standards. These characteristics facilitate components replacement, integration, scaling and distribution. This pattern provides interoperability and modularity by design. Components designed following the microservice architecture principles can be easily deployed, integrated and updated because they have clear functional boundaries and are technologically independent; be dynamically replicated to meet specific processing loads; and be utilised independently or in collaboration with other components to fulfil business functions.

With microservices, development teams can concurrently work on different services, without interfering with each other's work. This approach allows for greater agility and flexibility in software development and improves fault isolation and resilience. Solutions like Docker[1] containers can be used to improve the availability, scalability, replaceability and deployment of microservices. However, implementing microservices requires careful planning and coordination to ensure that the individual services work seamlessly together and that the overall system remains secure and maintainable.

Blackboard Model

The blackboard model is a problem-solving paradigm for complex problems [72]. This approach offers an alternative to sequential problem-solving models by providing a parallel and concurrent method. It is particularly useful for solving problems that demand a combination of knowledge from different domains and have no straightforward solution. The blackboard model, as described in [73] and depicted in Figure 2.2, comprises two types of components: the blackboard and the knowledge sources. The blackboard serves as a data structure, such as a database or common repository, where the problem is presented to the knowledge sources. The knowledge sources are agents that possess the necessary knowledge or have access to it and are specialised in solving a specific part of the problem. Each knowledge source contributes to solving the problem by providing a

[1] Docker: `www.docker.com`

partial solution. These contributions can be either sequentially dependent or completely independent, meaning that some knowledge sources can build on the previous contributions while others can directly build from the problem. The partial solutions are shared on the blackboard for other knowledge sources to use. Once all knowledge sources have contributed to the problem, the partial solutions are combined to form the final solution.

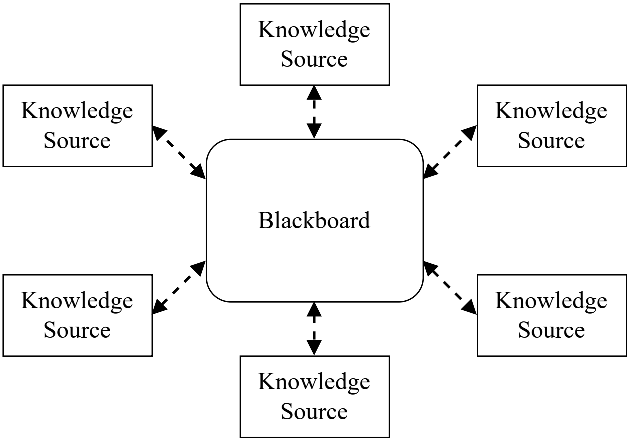


Figure 2.2: The Blackboard Model

The advantages of the blackboard model are diverse. When compared to sequential problem-solving models where tasks are performed one after another and each task waits for the previous one to finish, the blackboard model allows for parallel and concurrent execution of tasks, while still accounting for sequential output-input dependencies. Additionally, it minimises data duplication as the problem and the partial solutions are shared on the blackboard and do not need to be transmitted from one knowledge source to another. This can reduce bottlenecks and, because the knowledge sources are independent, they can be effortlessly scaled, replicated and replaced. However, the blackboard model requires a coordination mechanism to activate the knowledge sources [74], which can become challenging to implement as the amount of knowledge sources grows. Additionally, the outputs of knowledge sources must be represented in a format that facilitates their integration and understanding.

The blackboard model has influenced software system architectures in what is known as the blackboard architecture [73]. This architecture shares the same principles as the blackboard model, but with the blackboard realised as a centralised database and the knowledge source represented by the users or components of the system that contribute and consume from the database. In the blackboard architecture, communication can originate from either the knowledge source or the blackboard, and the blackboard plays an active role by notifying knowledge sources of any changes.

Machine Learning

Machine learning [75] is a subfield of artificial intelligence that enables computers to learn from data, without being explicitly programmed. It involves the development of algorithms and statistical models that can recognise patterns and relationships in data and use this information to make predictions or decisions about the data. Machine learning algorithms can be categorised into three main types: supervised, unsupervised, and semi-supervised, depending on the degree of human intervention required. Supervised learning involves training a model on a dataset that is already labelled with the correct output. In unsupervised learning, the model is trained on an unlabelled dataset and it must identify patterns and relationships without any prior knowledge. Semi-supervised learning is a combination of both. The applications of machine learning are vast, including computer vision, speech recognition and natural language processing. With the availability of big data and powerful computing resources, machine learning has become an increasingly important field in both academia and industry.

One of the key aspects of machine learning is the representation of data in a way that can be effectively processed by algorithms and models. Vector representations, also known as embeddings, are a popular technique for representing data. Embedding techniques are used to represent concepts from data as vectors in a high-dimensional space. These embeddings can be derived from a wide variety of data types from text, images, and audio to sequences and molecular structures. These resulting vector spaces provide sub-symbolic representations through mathematical models, enabling concepts to be organised in the space based on similarity or other relevant relationships. Embedding techniques have a stochastic component and require larger amounts of data to produce meaningful results. The results are hard to explain to humans, but even large vector spaces can be efficiently managed by computers.

Well-known techniques for word and text embedding are word2vec [30] and transformers [31] like BERT [28] and GPT-3 [29]. These models are especially relevant for AI applications that rely on semantic and contextual similarity like natural language processing, chatbots, text summarisation and recommendation systems.

Vector databases are an emerging technology that allows for the efficient storage, indexing and retrieval of large vector models. These databases are optimised for handling vectors and provide functionalities for similarity search using algorithms such as HSNW [76] and FAISS [77]. Currently there are a few available vector databases such as Milvus[2], Weaviate[3] and Vald[4].

[2] Milvus: `milvus.io`

[3] Weaviate: `weaviate.io`

[4] Vald: `vald.vdaas.org`

Natural Language Processing

Natural Language Processing (NLP) “is the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content” [27]. It involves a variety of tasks that deal with natural language such as part-of-speech tagging, named entity recognition, sentiment analysis, relation extraction, machine translation, topic classification, and text generation. Among these, in this work, I primarily utilised named entity recognition and relation extraction to extract relevant information from news articles.

Named entity recognition (NER) [11] focuses on identifying and classifying named entities such as people, organisations and locations from the text. It plays a crucial role in applications like information extraction, text summarisation and question-answering systems by identifying the relevant entities. Named entity linking (NEL) [13] is a related task that involves linking the identified entities to a specific knowledge base such as Wikipedia, Freebase, DBpedia and Wikidata. NEL provides links or identifiers corresponding to the entities from the knowledge base. However, disambiguating polysemous entities remains a significant challenge in NEL. Relation extraction (RE) is another task that involves identifying and classifying the relationships between named entities, which can either be predefined from a target set of relations or belong to a knowledge base.

An array of techniques can be employed to carry out NER, NEL and RE, including rule-based systems, machine learning algorithms and deep learning models. These techniques rely on diverse linguistic features, like part-of-speech tags, syntactic dependencies and semantic information, to accurately identify and categorise entities and relationships.

Knowledge Graphs

According to [32], knowledge graphs capture and abstract knowledge using graph-based data models wherein entities of interest are represented as nodes and the relations between them as the edges of the graph. Knowledge graphs provide symbolic representations through concepts, relations and logic rules. Ontologies and rules define the semantics and terms of the graph, facilitating data integration and reasoning. Knowledge graphs are especially relevant for integrating and extracting value from heterogeneous and dynamic data. They provide precise symbolic representations that do not require large amounts of data to become meaningful. Although knowledge graphs are easy for humans to understand, efficiently managing large graphs can be hard for computers. Compared to relational and NoSQL models, knowledge graphs facilitate semantic integration, flexible data and schema evolution, and graph query languages for exploring complex relations through arbitrary-length paths.

In particular, in my thesis, I focused on knowledge graphs that employ semantic technologies [37]. These technologies are based on the idea of the Semantic Web [38] and enable the representation, integration, linking and processing of data and knowledge in a machine-readable way. Semantic technologies facilitate the creation of structured data that can be easily shared, enriched with semantic meaning and understood, thereby making it possible to derive insights and take informed decisions from complex data. They encompass a variety of technologies and standards, like RDF (Resource Description Framework) [78] and SPARQL (SPARQL Protocol and RDF Query Language) [79]. These technologies also facilitate the creation of Linked Open Data (LOD) [35], which is machine-readable interlinked data that is freely available on the Web and can be integrated and reused across different domains and applications.

Knowledge graphs are particularly relevant for news because they provide means to connect and make sense of the vast amount of information that is generated daily. By representing news articles in a graph, relationships between entities and concepts can be established, leading to more advanced querying and reasoning. This in turn can lead to more accurate and comprehensive results when searching for specific topics or events. Furthermore, knowledge graphs can help to identify patterns and trends in the news, such as the emergence of new stories and connections between events. Knowledge graphs have the potential to greatly enhance our understanding of the world and events. In the context of journalism, knowledge graphs can help journalists to stay informed and updated on the latest developments in their respective fields, while also providing a means for generating new ideas and angles for stories [14].

Methods

In this PhD Thesis, I studied the social and technical factors contributing to the complexity of JKPs, the implications of the recent advances in AI, and the ethical aspects of using JKPs. I aimed to address these factors by exploring the theoretical and practical aspects of JKPs and designing a software reference architecture (SRA) to facilitate the adoption of JKPs. At the same time, I explored the theoretical implications in practice through the instantiation of prototypes as proof of concept to provide an empirical foundation. The Design Science Research [80] was the most appropriate methodology for this work because it involves designing and developing artefacts and theories that solve practical problems and contribute to scientific knowledge.

The Design Science Research methodology "supports a pragmatic research paradigm that calls for the creation of innovative artefacts to solve real-world problems" [81]. In the information systems field, design-science researchers typically undertake an iterative process consisting of three distinct cycles [82] (see Figure 3.1), which are focused on understanding the application context or environment, studying and improving the theoretical framework, and evaluating the artefacts [83]. The *relevance cycle* studies the phenomena of interest, the challenges and opportunities, and assists the researcher in framing the problem. It includes the environment where the phenomenon is perceived and the problem is defined by the researcher, often including the opinions and experiences from stakeholders. The *rigour cycle* provides the researcher with the raw materials, foundations, methodologies, and previous results that ensure the research contributions are original. It also encompasses the study and generation of scientific theories. The *design cycle* is where the development, evaluation and refinement of the artefacts and theories take place, with the inputs from the other two cycles.

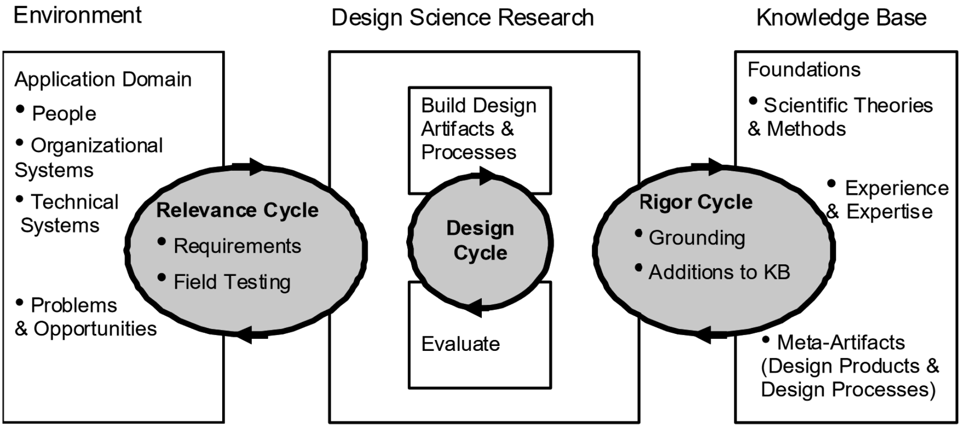


Figure 3.1: Design Science cycle (excerpt from [82])

In the rigour cycle, to identify and analyse the scientific theories on JKPs, I conducted a systematic literature review process [84] and qualitative meta-analysis approach [85]. In the relevance cycle, I met with different experts and representatives from the Norwegian media industry to ensure the relevance of the contributions. The theories on JKPs and the experiences from the environment informed the design cycle. In the design cycle, I employed a method for empirically-grounded software architectures to design the SRA [86]. To evaluate the SRA, I developed a prototype of a big-data and AI-ready JKP called News Hunter [2] as a design-science artefact.

Literature Review

To identify and analyse the literature on JKPs, I followed both the systematic literature review [84] and qualitative meta-analysis methods [85]. The systematic literature review provides the rigour to identify all relevant literature in the field, while the qualitative meta-analysis supplies the scientific method to extract and synthesise the information.

The systematic literature review process is a methodical approach to identifying, assessing and synthesising all relevant research studies associated with a specific research question or topic. The objective of a systematic literature review is to furnish a comprehensive and impartial summary of the existing evidence on a particular research area. To conduct a systematic literature review, researchers must first formulate a precise research question and search strategy to locate all related articles from various scientific databases and journals. Then, researchers must scrutinise and evaluate each article to determine its quality and relevance. Finally, the outcomes are synthesised and presented.

The qualitative meta-analysis involves a process of data extraction, coding, and conducting thematic analysis to identify and synthesise common themes and patterns across studies. In the coding part of the process, researchers independently codify the identified studies and their content using keywords. The resulting codes are compared and used to reach an agreement to derive the themes and conclusions of the analysis. The goal is to provide a comprehensive and insightful understanding of a particular phenomenon or issue based on the available qualitative evidence [87].

To extract themes for an in-depth analysis of the literature on JKPs, I defined a strategy to code claims (i.e., the authors' statements on JKPs) from the identified relevant studies within different categories. Meaning that researchers could assign multiple keywords in each category and leave categories empty to express that a claim is not related, as in Table 3.1. To evaluate the coding agreement between researchers, I used a modified version of Gwet's AC$_1$ [88] inter-rater reliability coefficient with nominal ratings. Gwet's AC$_1$ is an alternative to Cohen's Kappa rate agreement [89] to address its well-documented statistical problems [90]. They both measure how similar the overall codification of the researchers is. Gwet's AC$_1$ inter-rater reliability coefficient with nominal ratings accounts for one keyword per category, mismatching keywords and missing codifications. However, my coding strategy conflicts with Gwet's AC$_1$ as it neither accounts for partial matches in case of multiple keywords in the same category nor considers that a claim cannot be related to one category. Gwet's AC$_1$ would interpret Table 3.1 as one "missing" match for Category 1 and one mismatch for Category 2. Therefore, I had to modify the coding by adding a new keyword `undefined` for the claims that are not related to a specific category and filling in with missing keywords the partial matches, as in Table 3.2.

Table 3.1: Example of a codification where the researcher `B` express that claim `X` does not belong to Category 1 and both researcher `A` and `B` coded Category 2 with multiple values.

Researcher	**Claim**	**Category 1**	**Category 2**
A	X	NLP	semantic web, LOD, ontology
B	X		semantic web, ontology, reasoning

Table 3.2: Reinterpretation of Gwet's AC$_1$ where the code from `B` in Category 1 is set to `undefined` and the Category 2 is interpreted as two positive matches (i.e., one for `semantic web` and one for `ontology`) and two "missing" matches (i.e., one for `LOD` and one for `reasoning`).

Researcher	**Claim**	**Category 1**	**Category 2**			
A	X	NLP	semantic web	ontology	LOD	
B	X	`undefined`	semantic web	ontology		reasoning

Empirically-Grounded Software Architecture

To design and evaluate the SRA, I followed a method for designing empirically-grounded reference architectures [86]. This method provides a systematic process to ensure that the resulting designs are informed by the scientific literature, proven principles, empiric experiences and stakeholder interests. It comprises six steps as illustrated in figure 3.2, where the initial five steps provide the "empirical foundation" and the sixth step provides the "empirical validity".

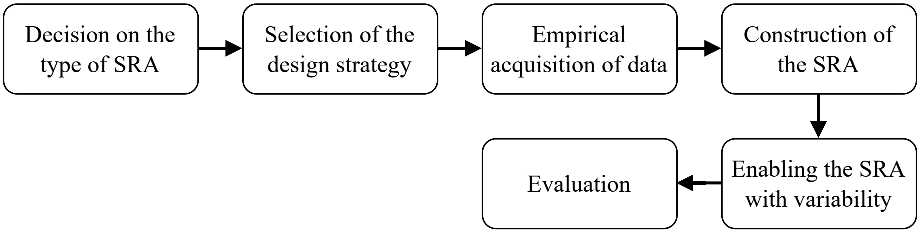

Decision on the type of SRA
Selection of the design strategy
Empirical acquisition of data
Construction of the SRA
Enabling the SRA with variability
Evaluation

Figure 3.2: Method for designing empirically-grounded software architecture

The foremost step in this method is the decision of the architectural model, based on a predetermined set of categories outlined in [54]. The choice informs the level of abstraction or detail in the architecture, as well as the interested parties or stakeholders who must partake in the process. The second stage centres on the selection of a design strategy, which considers the driving factors behind the design. The third stage involves the empirical acquisition of data, including scientific theories, proven architectural principles and experiences. The fourth stage is the construction of the SRA, founded on the information previously gathered. The fifth step entails equipping SRA with variability, which might not always be pursued but is critical for enabling the SRA to adapt to different scenarios. The final stage involves the evaluation of the SRA.

To evaluate the design of the SRA, I used two approaches: a mapping study and the development and testing of a prototype of JKP as proof of concept. The mapping study provided a qualitative evaluation where I mapped the requirements for the SRA to the architectural components and principles, as well as to the identified JKPs in the literature. The mapping ensured that the SRA is grounded in the literature and accounts for the requirements. The development of the JKP prototype following the SRA facilitated the assessment and refinement of the feasibility of the design of the SRA. The prototype also allowed me to use quantitative metrics to evaluate the SRA, considering measures like the total amount of ingested data and the time performance to transform the gathered data into graphs.

Results

JKPs are both a shift and a great opportunity in the news industry [4], [12]. But with every opportunity comes a challenge, and in this case, it is the lack of established knowledge on the complexity of the social and technical aspects of JKPs to facilitate their realisation. To make the transition to JKPs safely and effectively, it is essential to study the phenomena of interest in-depth and provide a blueprint to support news organisations in evolving their infrastructure to JKPs. Therefore, in this PhD thesis, I asked: *What is a good software reference architecture for Journalistic Knowledge Platforms?*

Following a design science research approach, I studied and explored the theoretical and practical aspects of JKPs and contributed to address the problem in several ways. Firstly, I provided a comprehensive analysis of the current state, challenges, opportunities and future directions of JKPs, which can guide further research and development. Secondly, I proposed a software reference architecture for JKPs that captures best practices, provides indications to design concrete solutions, and considers the future AI and big-data challenges, which can serve as a blueprint for implementing concrete JKPs. Finally, I explored how to enhance the current state of JKPs and NLP pipelines, as well as I investigated the ethical and other related aspects of JKPs, which can inform future realisations of JKPs. I believe my contributions will be beneficial for both researchers and practitioners and inform other domains with similar needs.

The contributions of my thesis can be divided into three blocks, each represented by one of the three manuscripts included in this thesis (Manuscripts) and other publications. In the first manuscript (Manuscript I), I explored and defined the concept of JKPs. In the second manuscript (Manuscript II), I designed the SRA for JKP based on the findings from the first manuscript. In the last one, the third manuscript (Manuscript III), I focused on improving one of the internal components of the JKPs.

First Manuscript

Synopsis. The first manuscript (Manuscript I) provides an extensive review and analysis of the phenomenon of JKPs [1]. The analysis is based on 14 platforms that serve as the foundation of this contribution. This publication represents the first in-depth definition of the Journalistic Knowledge Platform concept. It explores the current state, future directions, challenges and opportunities of JKPs, and offers detailed descriptions of their stakeholders, functionalities, techniques, components and other relevant aspects.

The analysis reveals that JKPs encompass a diverse array of stakeholders, ranging from general users to organisations and technical agents. For example, the analysed JKP from the SUMMA project offered services to the journalists of the BBC, the government of the United Kingdom, political scientists, and the newsrooms managers of the Deutsche Welle [51]. The review also examines the different types of stakeholders, highlighting their involvement in various aspects of JKPs, from information gathering and news creation to distribution and knowledge exploitation.

The analysis also identifies the most common functionalities and techniques that JKPs employ. These include automating time-consuming tasks in newsrooms that relieve journalists from the burden of monitoring news, social media and TV and radio broadcasts, and providing insights to facilitate news creation and decision-making for improved news coverage and production decisions. One JKP analysed by the review is the one developed by Reuters, which processes tweets in real-time and identifies, clusters, and summarises emerging newsworthy events before competitors publish them [49]. The review delves into the AI techniques that JKPs utilise, including machine/deep learning, NLP and knowledge graphs. For example, the review analysed the JKP developed in the Annoterra project that used knowledge graphs, linked open data and NLP techniques such as named entity recognition to annotate news feeds from NASA Earth Observatory and enhance information search [41]. In addition, the review provides a description and classification of the common types of software components in JKPs and their relation to the functionalities and techniques.

The review highlights that while JKPs offer numerous opportunities, such as supporting journalists with news creation and generating timely and newsworthy information, there are also significant challenges. These challenges include processing large amounts of heterogeneous data, dealing with multimedia and multilingual news information, and maintaining complex software systems. For example, the review examines the JKP developed in the Newsreader project that managed an archive consisting of billions of articles, biographies and reports [48]. The challenges are not only limited to the technical aspects; as stated in the review, JKPs also face social challenges like ensuring copyrights, authorship and media standards, and dealing with a diversity of customers' needs.

The review concludes by shedding light on the future directions of JKPs in research and practice. For instance, it suggests the development of solutions to combat misinformation and propaganda and discusses the potential of neural-symbolic AI [34] to enhance JKPs. It also emphasises the importance of establishing guidelines on the utilisation of AI in news creation processes and the need to integrate explainable AI.

Additional Contributions. Alongside this manuscript, I investigated potential privacy threads that JKPs could pose and proposed a detailed framework for classifying various scenarios of privacy violation [5] (Publication IV). The framework describes and analyses the types of data and scenarios that may infringe privacy policies, especially if data is collected, derived, stored and merged from diverse sources, including news articles, social media platforms and encyclopaedic knowledge bases.

Contribution to the Research Question. This contribution defines and studies the concept of JKPs from different perspectives, providing a detailed description of the phenomena of interest that can guide further research and development. It contributes towards the research question by presenting a thorough analysis of existing scientific literature that can be used to identify and derive requirements and qualities to define software architectures for JKPs. Moreover, as this contribution describes JKPs in-depth, it can assist newsrooms in understanding and asserting the advantages of adopting JKPs.

Second Manuscript

Synopsis. The second manuscript (Manuscript II) presents a software reference architecture for JKPs [2]. This contribution builds on the results of the first manuscript to derive and define the functional and non-functional requirements for the architecture. The proposed architecture is based on existing literature and the authors' experience developing a series of prototypes in collaboration with industry partners [7], [10].

The manuscript argues that existing software reference architectures for managing big data do not meet the specific requirements of JKPs. Unlike existing architectures that primarily focus on immutable data and analytics, JKPs integrate knowledge bases and AI and focus on exploring and understanding knowledge that evolves over time. To address this gap, this publication proposes a software reference architecture designed to be technology independent, open-ended, and long-lasting, with components and services that can be replaced and integrated with other systems. The proposed architecture is based on several established principles, including microservices [71], semantic technologies [37], [38], liquid architecture [58] and blackboard architecture [73]. Moreover, this contribution introduces two new types of software components: one for updating machine/deep-learning models and schemas, and the other for curating knowledge representations. These new components are part of the five core types of components that form the architecture.

This manuscript also describes a proof-of-concept prototype that I developed following the architectural descriptions and used to test and validate the feasibility of the proposed software reference architecture for JKPs. The prototype gathers news and social media in real time from a variety of sources, including RSS, NewsAPI[1], GDELT[2] and Twitter. It extracts relevant information from text and links it to external knowledge bases, such as DBpedia and Wikidata, using deep learning models. The extracted information is represented following an ontology [91] in a large semantic knowledge graph. The prototype can be used, for example, to provide background information and suggest stories relevant to a location. The overall prototype is built on top of big-data-ready technologies such as Apache Kafka[3], Apache Cassandra[4] and Blazegraph[5], and deployed on a cloud-based platform consisting of 114 containerised services distributed across 33 instances. These instances use a total of 99 vCPU, 324GB RAM and 20TB disk. To experiment with the scalability of the architecture, the prototype was run for a period of 6 months, generating more than 4B ($6 \cdot 10^9$) triples in total, which is the same as 11M ($11 \cdot 10^6$) triples daily from news and tweets and 11M ($11 \cdot 10^6$) from GDELT events.

Additional Contributions. During the realisation of the prototype, I explored some specific aspects of JKPs to inform my decisions. I proposed an approach to integrate end-to-end machine learning models with traditional NLP pipelines for text annotation [6] (Publication V). I examined what techniques that combine semantic technologies and machine learning approaches could be used to identify and cluster emerging events from news [8]. I outlined a framework for transforming structured news events from GDELT into graphs [9] (Publication VI). These works provided me with technical knowledge and expertise on specific aspects of JKP, which in turn helped me better identify the needs and requirements for the architecture.

Contribution to the Research Question. This contribution is the main answer to the research question. It identifies the required functional and non-functional qualities for a good architecture of JKP and proposes a software reference architecture that addresses them. The proposed architecture also considers the challenges posed by AI and evolving knowledge representations on JKPs and AI systems. The manuscript provides a blueprint to support news organisation in evolving their existing infrastructure to JKPs

[1] NewsAPI: a service that provides news from over 80.000 sources. Available at: `newsapi.org`

[2] GDELT: A knowledge base that provides semi-structured information about political-crisis-related events from worldwide news. Available at: `www.gdeltproject.org`

[3] Apache Kafka: A real-time event-streaming platform. Available at: `kafka.apache.org`

[4] Apache Cassandra: A highly-scalable NoSQL database. Available at: `cassandra.apache.org`

[5] Blazegraph: a high-performance graph database for RDF. Available at: `blazegraph.com`

and designing concrete architectures. Moreover, the outputs of this contribution may inform research and practice on architectures in other domains with similar needs.

Third Manuscript

Synopsis. The third manuscript (Manuscript III) delves into a specific part of the JKP by tackling the challenge of creating rich semantic text annotations using multiple natural-language annotation techniques [3]. This manuscript argues that existing NLP pipelines are inefficient and impractical for integrating NLP techniques because they perform annotations sequentially, which creates unnecessary dependencies.

To address this problem, this contribution proposes an approach that combines partially sequential and partially parallel annotations using the blackboard model [72], [73]. This model defines two types of components: the blackboard and the knowledge sources. The blackboard is a shared resource where the problem and partial solutions can be accessed and combined. The knowledge sources are decoupled agents which have the required knowledge for solving different parts of a problem. Each knowledge source contributes to solving the problem by providing a partial solution that combined form the final solution. They collaborate concurrently towards the solution using a coordination mechanism. The proposed approach employs big-data-ready technologies like Apache Kafka, which operates as a blackboard for sharing information and coordinating knowledge sources. It also uses RDF [78] and semantic vocabularies such as NIF [92] to integrate the annotations of different NLP modules. This contribution also proposes extensions to the NIF vocabulary to include additional types of NLP annotations because some types of annotations are not covered. In addition, the manuscript presents a proof-of-concept prototype to evaluate the performance of the proposed solution and discusses different application use cases. The proposed approach significantly reduces the complexity and dependencies of sequential annotation pipelines by using decoupled components, facilitating the extension and scaling of annotation systems.

Contribution to the Research Question. This manuscript presents an efficient and scalable solution for creating rich semantic text annotations for JKPs. It contributes towards the research question by providing a detailed description on the design of a specific part of the JKPs that facilitates the integration of different natural-language annotation techniques and services. The proposed approach supports the realisation of JKPs and can enhance the speed and quality of NLP pipelines, thereby improving the performance of JKPs. Additionally, the proposed solution is not only limited to the problem of creating rich semantic annotations and can be extended to other domains, making it a versatile and adaptable approach.

Discussion

In this PhD Thesis, I tackled the problem of integrating knowledge-based solutions that utilise big data and AI in today's newsrooms. The field lacked an in-depth analysis of the platforms that integrate these solutions, which made it difficult for newsrooms and research to scientifically back decision-making and elucidate the challenges and opportunities. Moreover, a clear and generic description of the software architecture of the platforms was necessary to facilitate their adoption. Thus, in this thesis, I sought to address both the theoretical and applied aspects of the problem.

To address these aspects, I defined and investigated the concept of Journalistic Knowledge Platform [1], designed a software reference architecture for JKPs [2], and delved into techniques for enhancing and optimising the performance of JKPs [3]. These findings can aid in realising JKPs, improving their functionality and efficacy, laying the groundwork for widespread adoption in the journalism industry, leading better-informed decision-making, and guiding future research. However, there are still some aspects of this thesis and in general of JKPs that require further exploration and discussion. In the following sections, I introduce some of these aspects and explore their implications.

On the Contribution[1]

In this thesis, I focused on JKPs that utilise semantic knowledge graphs. This choice was motivated by the widespread use and advantages of semantic technologies in knowledge representation [14], [32]. Semantic technologies provide a structured and machine-readable format to represent the meaning of concepts and their relationships that enables the interpretation and reasoning about knowledge. The result is a more comprehensive and interconnected view of the data that also supports complex query answering. However, JKPs are not limited to semantic knowledge graphs, as other types of technologies can be employed to represent knowledge, including non-semantic knowledge graphs. Al-

[1] An in-depth discussion of each contribution can be found in the manuscripts (Manuscripts).

though these alternative solutions do not offer the same advantages as semantic technologies, they may offer other benefits like high scalability and performance. To account for these alternatives, I designed the software reference architecture open-ended and modular to facilitate the integration and replacement of different technologies.

The disruption of generative models and tools, such as ChatGPT[2], DALL·E[3] and Stable Diffusion [93], presents a unique opportunity for JKPs. However, I did not explore nor build on them, because these solutions were proposed at the end of my thesis and there is a huge gap in the literature to be able to fully understand their opportunities, implications and risks [94]. These AI-powered tools can perform a range of tasks, including summarising text, generating stories, and creating images from textual descriptions. Integrating them into JKPs has the potential to speed up content creation for journalists, enabling them to focus on in-depth reporting and investigation. In addition, they can offer personalised news experiences for readers by crafting customised articles tailored to their interests. These AI tools bring new possibilities for creating solutions and interactive ways that have yet to be seen. They have the potential to change the way JKPs are designed and the interaction between people and information.

The proposed software reference architecture and the developed prototype proof of concept can be employed to further explore the intersection of AI and journalism. This work can serve as a testbed for integrating future solutions that include non-semantic knowledge graphs and generative AI.

On the Challenges

During the realisation of this thesis, I encountered some unforeseen challenges. Initially, there was great enthusiasm from the semantic web community for Blazegraph, a high-performance graph database that promised scalability and efficiency in managing large knowledge graphs. Blazegraph was even used in various applications, including Wikidata, and has gained popularity for its ability to handle complex queries and offer fast query response times. Thus, it was a natural choice to include Blazegraph as a key component of this work. However, Blazegraph was discontinued after the development team was hired by Amazon Neptune and its capability for horizontal scaling was never fully developed nor supported. This limited the realisation of the prototype and evidenced the current limits on open-source triple-stores for RDF and SPARQL. At the same time, it highlighted the need for providing an architecture that does not need to rely on a single type of technology and is flexible enough to integrate different databases.

[2] ChatGPT: a chatbot that performs a wide range of tasks. Available at: `openai.com/blog/chatgpt`

[3] DALL·E: a generative model that creates images from text. Available at: `openai.com/research/dall-e`

Another pitfall I encountered was the expectation that Relation Extraction techniques would provide better results for extracting relations in news. This natural language processing technique is utilised to identify relationships between entities mentioned in a text. The extracted relations can improve the comprehensiveness and completeness of knowledge graphs, especially in the news domain. However, the available supervised models were not good enough to be widely applied in this context due to challenges related to the quality and completeness of the text corpus it is trained on, as well as the complexity and variability of relationships between entities. Currently, the open relation extraction supervised models are trained for a specific reduced set of relations which is not enough for representing news. For these techniques to be reliable for news, they would need to detect a wide range of relations similar to the number of properties in Wikidata or DBpedia. Nevertheless, ongoing research on foundation models [94] such as BERT [28] and GPT-3 [29] provide hope that relation extraction techniques will become more effective in the years to come. These large deep-learning models, despite not being trained for relation extraction tasks, may perform well on open relation extraction.

On the Methodology

To address the complex challenges of Journalistic Knowledge Platforms, a multidisciplinary approach was necessary. This approach required expertise from a range of fields, including information systems and journalism, as well as a thorough understanding of the rapidly evolving landscape of AI and the news industry. To ensure a rigorous and cross-disciplinary collaborative approach, I employed the Design Science Research methodology. This methodology can bring together experts from diverse disciplines to collaborate towards a common goal, fostering interdisciplinary thinking and continuous validation of results. This approach enables the development of comprehensive solutions that consider multiple perspectives and address a wide range of issues. Throughout this thesis, conversations with software developers for the newsroom industry and journalists provided insights into technical, journalistic and ethical requirements. However, the literature on Journalistic Knowledge Platforms is primarily focused on the European and North American perspectives, and further research is needed to explore alternative approaches that may be more relevant to other regions of the world.

To further explore and understand the context of the problem, the design science research approach can be complemented with qualitative research methods to study people's experiences, opinions, and beliefs. These methods could include interviews with the technical teams and journalists of newsrooms working on JKPs or exploring AI tools. These interviews could be focused on gathering experiences and further evaluating and discussing the software reference architecture I proposed.

On Outsourcing

JKPs tailored to newsrooms can provide greater control over data, ensuring data ownership and avoiding vendor lock-in scenarios, among other benefits discussed in [1]. This autonomy is a valuable asset in a world that increasingly values data and information. However, newsroom organisations often lack the resources and expertise to develop and maintain JKPs. In contrast, third-party solutions, often provided Software as a Service (SaaS), offer easy-to-use and accessible solutions that save newsrooms time and resources. This convenience comes at the cost of dependence, as newsrooms become reliant on third-party providers, and may not cater to the unique requirements of the newsroom and provide unoriginal journalistic results.

The balance between autonomy and dependence is a recurring dilemma that extends beyond newsrooms. It considers the potential risks and benefits of relying on third-party solutions for outsourcing crucial resources. Although the current trend may lean towards the usage of third-party solutions for the initial stages of digital exploration, efforts in JKPs by media giants such as BBC and Reuters, as revealed in [1], indicate that the focus is on centralising knowledge within the newsroom for the most critical systems that provide a competitive edge.

Ultimately, the decision to adopt a JKP or third-party solutions must be based on a careful evaluation of the trade-offs between autonomy and dependence, as well as the specific needs and resources of the newsroom. Therefore, the results of this thesis shed light on a better understanding of JKPs and facilitate the assessment of the trade-offs. The software reference architecture [2] I proposed can also be beneficial in determining which components of a JKP are critical systems and which can be outsourced by third-party solutions. These results also support fostering a broader debate about the optimal strategy for achieving the desired equilibrium between autonomy and dependence, thereby allowing for hybrid solutions.

On the Corruption of Knowledge

As the knowledge base of a JKP grows with new information, it can become corrupted, propagating misinformation, and generating information bubbles. The corruption of knowledge revolves around the issues of epistemology, the nature of truth and the role of technology in shaping our understanding of the world. The reliability of a knowledge base involves technical aspects to safeguard against such a scenario and ensure that the information remains trustworthy and transparent.

To safeguard against the corruption and bias of the knowledge base, a rigorous process of fact-checking and verification must be upheld, while also promoting transparency and openness in the creation and dissemination of information. One way to achieve this is by utilising open-access external knowledge bases that promote the free exchange of ideas and information such as DBpedia, Wikipedia, Wikidata and fact-checked repositories. By continuously monitoring and contrasting the information with these trustworthy sources, it is possible to guarantee the consistency of the single source of truth as a centralised knowledge base. This process can be automated to detect inconsistencies and add missing pieces of information, but as the knowledge is constantly evolving, the journalists should be kept in the loop to assert the quality of the updates. Every update must be tracked to provide a transparent process and provenance that facilitates journalists asserting its quality and reliability.

However, this does not mitigate the risk of an information bubble. To prevent such a scenario, the knowledge base must grow in different directions and encompass a variety of information. This can be achieved by incorporating factual information from external trustworthy sources and other sources like social media or curated knowledge bases about news and events like GDELT. By incorporating new information, whether it is related to the existing knowledge or not, the knowledge base will expand its scope and avoid becoming an information bubble. This approach ensures that the information remains diverse, instead of being limited to a narrow perspective.

Ensuring the reliability of the knowledge base is a complex problem, with various dimensions and implications. The software reference architecture [2] I proposed defines a set of components that handle the updating and expansion of knowledge repressions and corroborating information with external knowledge bases. Moreover, the usage of semantic technologies and linked open data facilities the integration of internal and external knowledge and enhances its comprehension. These technical solutions are suggested to prevent the corruption of the knowledge base and the creation of information bubbles, as well as to provide transparent and trustworthy information.

On the Affordability

Implementing JKPs requires significant infrastructure, which may be unaffordable for small and medium-sized newsrooms. This situation can jeopardise the business of these newsrooms, as it enlarges the gap between big and small players in journalism, exacerbating the race for broadcasting the latest news. This delves into the ethical considerations of levelling the playing field for all newsrooms, regardless of size, to guarantee that journalistic integrity and quality are not exclusively reserved for large news outlets with greater financial resources.

Access to JKPs and advance technologies can provide smaller newsrooms with the necessary resources to generate high-quality journalism. This can contribute to a more diverse and representative media landscape, which is essential for a healthy democracy. However, large news outlets with greater financial resources may have an advantage in adopting JKPs, whereas smaller newsrooms may struggle to afford the required infrastructure. This can lead to a scenario where a limited number of dominant players have access to the most advanced technologies, leaving smaller newsrooms left behind.

In addressing the ethical considerations surrounding access to Journalistic Knowledge Platforms, it is crucial to explore strategies that promote equitable access, particularly for smaller newsrooms. Open-source software and shared infrastructure development can be an effective approach to reducing financial barriers to implementation. Through freely accessible software that can be shared among newsrooms, the cost of implementation is minimised. By sharing infrastructure and resources, smaller newsrooms can benefit from economies of scale and access to advanced technologies they might not be able to afford on their own. Additionally, hybrid solutions can offer an optimal strategy where the most demanding parts of the JKPs are provided by third parties that adhere to open standards and principles. These hybrid solutions could combine tools that address certain parts of the JKP, while a shared knowledge base is maintained by newsrooms to ensure ownership of information and avoid vendor lock-in scenarios. The SRA I proposed can facilitate collaboration between smaller actors, as it can provide advice on deciding which parts of the JKPs can be shared or outsourced, but also help open-source projects in organising their efforts.

Ensuring equitable access to JKPs and other advanced technologies is a crucial ethical consideration for the future of journalism. By promoting access and collaboration, it is possible to create a more diverse and representative media landscape and ensure that journalistic integrity and quality are exclusive to large news outlets with great financial resources. In this thesis, I showed how semantic technologies and open standards can be employed to integrate data from multiple sources [1], [2] and how systems can benefit from these standards to intercommunicate [3]. These contributions can serve as a starting point to explore how JKPs can be built as a hybrid-solutions and how open-source solutions can be integrated into current newsroom processes to build JKPs.

On the Data Governance

The collection, storage, and use of vast amounts of data in JKPs is not exempt from concerns over privacy, provenance and licensing. JKPs access data from different sources which can include or reveal personal data [5]. The identification of such potential privacy violations is crucial to ensure the ethical use of JKPs and compliance with privacy regulations. Furthermore, the provenance of the data used in JKPs, as well as the origin and accuracy of the data can be challenging to determine. The lack of transparency in data provenance can lead to the spread of misinformation and biased reporting, which can significantly impact public trust in journalism. Finally, licensing implications are a concern as copyrighted material used in JKPs can potentially violate intellectual property laws. Proper licensing must be obtained for all materials used in JKPs to avoid any legal repercussions. The ethical and legal use of JKPs is critical to the credibility and integrity of the journalism industry. Both the framework for classifying privacy threads [5] and the SRA I proposed can aid in building JKPs that consider these aspects from their design. However, further research is needed to efficiently identify and address privacy, provenance and licensing issues.

The materials used to train the AI models for JKPs can also have implications for privacy, provenance and licensing. The use of publicly available and private data sets can raise concerns about the use of personal information, particularly if the data was not obtained with the consent of the individuals involved. The origin and accuracy of the data used to train the models is also a concern, as inaccurate or biased data can lead to flawed results. Additionally, the use of copyrighted materials in training data sets can raise licensing issues. Therefore, careful consideration must be given to the selection and sourcing of training data sets to ensure that ethical and legal standards are upheld in the development of JKPs.

On Foundation Models

Advances in foundation models [94], such as BERT [28] and GPT-3 [29], can have significant implications for JKPs. Trained on large and broad amounts of data, these models have shown remarkable performance in a wide range of natural language processing tasks, even without explicit training for the task at hand. While the limitations and risks of these models are not yet fully understood [94], they present a unique opportunity to enhance the functionalities and efficacy of JKPs.

Semantic representation and knowledge graphs are a popular choice for inferring new data insights and providing background information. However, this new generation of large deep-learning models could potentially provide even better results. This could lead to more accurate and relevant information being delivered to journalists in real time, improving the quality and speed of news writing. These models may impact JKPs in various ways, from fully replacing the current NLP models and semantic representations to hybrid solutions that incorporate both classic and new techniques.

Foundation models could potentially replace the information extraction process of JKPs, which is currently done by diverse NLP models. Rather than relying on multiple NLP models, a single foundation model could interpret natural language inputs, identify relevant information, and generate accurate outputs. These models could also replace semantic representations and knowledge graphs, acting as a knowledge base to infer relations, identify connections between stories and provide background information.

However, these models have been criticised for perpetuating biases, hallucinating, the lack of explanations, the unclear ownership of both training and output data, and generating misleading, toxic and harmful information. Additionally, they are expensive and time-consuming to train and update, which may affect their ability to keep up with the daily news cycle and produce outdated outputs. To mitigate these risks, a hybrid solution that combines current NLP models and knowledge representations with foundation models may be a more reliable option. In this way, foundation models could feed from and complement the current NLP models and knowledge representations, resulting in more accurate outputs and improving the functionality and efficacy of JKPs. By combining the different approaches, the foundation models can use more accurate and updated information, and at the same time, they can produce diverse and richer outputs for journalists. This will open the possibility for newer functionalities and interactions. Further research is needed to fully understand the potential of these models and to incorporate them into the design of JKPs in a responsible and ethical manner.

On Large Knowledge Graphs

A significant but often overlooked challenge faced by JKPs is the scale and complexity of large knowledge graphs that represent worldwide news. Although knowledge graphs have many benefits, such as their ability to capture complex entity relationships and provide rich contextual information, large knowledge graphs can be challenging to manage effectively. Reasoning over a large graph can be computationally expensive, and the scale and complexity of the graph can make it hard for computers to navigate efficiently, potentially leading to performance issues. Some existing approaches to address this problem include partitioning graphs by themes or temporal aspects and implementing the knowledge graph on top of highly-scalable databases, but these have limitations. Partitioning the graph can exacerbate the problem in the long run and create difficulties in exploring the graph as the information is dispersed and some relations may be lost across databases. Using big-data technologies to implement the database also comes at the cost of reducing query expression and reasoning since the underlying technologies are not specifically designed for knowledge graphs.

To maintain the efficiency and effectiveness of knowledge graphs, two alternative approaches can be considered: simplifying the graph and using vector spaces to reduce the graph search space. Simplifying the graph creates a more easily searchable structure at the cost of reducing expressiveness. The hybrid approach combines vector spaces with the knowledge graph to create a more efficient search and retrieval system. These approaches have the potential to improve the performance of knowledge graphs and overcome the challenges associated with their scale and complexity.

Simplifying the knowledge graph involves removing intermediate nodes and edges to reduce the space required to store the graph at the cost of reducing semantic expressiveness. This can make the graph more manageable and efficient to search through. This technique can reduce the search space and query complexity by simplifying the knowledge representations to the minimal form, sometimes removing even the semantics. For instance, a simplified knowledge graph will only contain direct edges between the news items, entities and concepts, and remove blank and intermediate nodes that shape the structure of the knowledge or provide metadata. This technique can improve performance by reducing the computational load of reasoning over the graph and making it easier for computers to traverse it. However, this technique may sacrifice some of the richness and complexity of the graph, resulting in the loss of information and difficulty in extracting meaningful insights from the data. Therefore, this technique should be used in conjunction with different stages of information to resolve the query and types of databases. First, an initial search is performed over the simplified knowledge graph to reduce the number of search candidates. Next, the potential candidates for solving the query are retrieved in their raw form and expanded into a full graph that contains the previously simplified semantic expressions. Then, a final query is performed over the expanded graph, resulting in a more accurate and informative result.

The hybrid approach combines vector spaces with knowledge graphs to create a more effective search and retrieval system. The knowledge representations are both represented in a knowledge graph and a vector space. Embedding techniques such as BERT [28], TransE [95] and RDF2vec [96] can be used to generate vectors from either text or knowledge representations. These vectors may be stored in a vector database which facilitates indexing and provides different similarity search functions. The vector database is first employed to conduct a search for related or similar concepts, which is computationally efficient for this type of search, even for large collections of vectors. Then, the retrieved results are used to reduce the search space over the knowledge graph. As knowledge graphs provide a structured representation of the relationships between entities and concepts, they can be used to conduct a more complex search over the reduced space and refine the query results. This approach takes advantage of the strengths of both vector spaces and knowledge graphs. Vector representations improve the efficiency of information retrieval, while knowledge graphs improve the results and make them easier to understand and explain.

On the Usage Scenarios

The proof-of-concept prototype of a JKP I developed (Prototype) can be easily expanded to integrate and test new functionalities and support further research. Different usage scenarios can be explored within the prototype such as detecting newsworthy news, clustering events and identifying key actors' roles in a story.

Journalists are overwhelmed with massive amounts of news-related information from social media, broadcasts and news feeds. Detecting breaking news in real-time has become challenging for journalists, as they must sift through a large volume of information to identify the newsworthy stories. The prototype I developed can be expanded to explore the detection of newsworthy stories by utilising foundation models like GPT-3. These large models may perform well on complex tasks without being explicitly trained for them and extract sophisticated patterns and relations from data. By integrating these models into the prototype and adapting them to classify the newsworthiness of the news stories, it would be feasible to explore their performance on this type of task. To facilitate journalists in assessing and evaluating the newsworthiness, these models can also be used to summarise the stories while the prototype can provide related background information and facts from the knowledge graph.

Detecting and following the development of events is a crucial task for journalists as it provides a deeper understanding of the context and significance of stories. With real-time event detection, journalists can stay up to date with the latest developments of a story and provide more accurate and comprehensive coverage. To accomplish this task, models like BERT can be deployed to represent newsworthy news items as vectors and cluster them into events. Models such as GPT-3 and LLaMA [97] could also be explored to evaluate the feasibility of instruction-following models to cluster news items into events as an instructed task. The resulting events can be enriched and connected to other events through the knowledge graph, enabling journalists to identify patterns and trends that would be difficult to detect manually. Furthermore, these large models could be employed to interactively explain the events by, for example, using chatbots like ChatGPT. This will create and facilitate research on new opportunities for human interaction between stories and journalists, where the events and graphs are fed to chatbots which can be used to explore the information through interactive storytelling techniques.

These large models can be utilised to identify the roles of the different characters of a news story [98]. This can be beneficial for evaluating potential biases and creative tasks such as writing and framing news stories. By determining the main characters and their roles, various techniques can be applied, such as conducting a sentiment analysis of each character to assess the neutrality of the story and measuring the weight of each character to suggest new story angles with different weights. This can help journalists to provide more balanced coverage.

Concluding Remarks

Despite the potential of Journalistic Knowledge Platforms to revolutionise the field of journalism, their adoption has been slow. This can be attributed to the complexity of their social and technical aspects, as well as the lack of a generic software architecture for JKPs to facilitate their realisation. To address these challenges and facilitate the adoption of JKPs, I followed a design science research approach to study both the theoretical and practical aspects of JKPs.

This thesis provides a comprehensive study of JKPs from the perspectives of information systems and software engineering. The first manuscript represents the first-ever detailed exploration and definition of the JKP concept. It establishes the theoretical foundation for JKPs and analyses their current stage, opportunities, challenges and future directions. The second manuscript is the main answer to the research question and proposes software reference architecture for JKPs. It defines the requirements of JKPs and presents a prototype of JKP as a proof of concept. The third manuscript introduces a new framework for parallel and flexible text annotation that combines semantic and big-data technologies. It provides a detailed description for integrating different NLP models. In addition, in this thesis, I discuss the implications of my research contributions and how cutting-edge research on AI can impact my results. I also reflect on diverse ethical considerations for JKPs and propose further research.

The results and discussion of this thesis can assist newsrooms in understanding and assessing the advantages of JKPs, support news organisations in evolving their existing infrastructure to JKPs and designing concrete architectures, improving the performance of JKPs, and enhancing the speed and quality of NLP pipelines. The findings of my research can inform both scholars and practitioners interested in this domain as well as other domains with similar needs.

In summary, this thesis provides insights into the theoretical and practical aspects of JKPs and offers concrete solutions to the challenges of their adoption. The contributions of my research can provide a foundation for further research in this domain. I hope this thesis will contribute to the development and adoption of JKPs, improving the quality and efficiency of journalism and its role in society.

AI systems can be powerful tools for good,
only if we build them right.

Bibliography

[1] M. Gallofré Ocaña and A. L. Opdahl, 'Supporting Newsrooms with Journalistic Knowledge Graph Platforms: Current State and Future Directions,' *Technologies*, vol. 10, no. 3, p. 68, May 2022. DOI: `10.3390/technologies10030068`.

[2] M. Gallofré Ocaña and A. L. Opdahl, 'A Software Reference Architecture for Journalistic Knowledge Platforms,' *Knowledge-Based Systems*, vol. 276, p. 110 750, 2023, ISSN: 0950-7051. DOI: `10.1016/j.knosys.2023.110750`.

[3] M. Gallofré Ocaña and A. L. Opdahl, 'A Blackboard Model for Parallel and Flexible Text Annotation,' Submitted.

[4] M. Gallofré Ocaña and A. L. Opdahl, 'Challenges and opportunities for journalistic knowledge platforms,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020.

[5] M. Gallofré Ocaña, T. Al-Moslmi and A. L. Opdahl, 'Data privacy in journalistic knowledge platforms,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020.

[6] T. Al-Moslmi and M. Gallofré Ocaña, 'Lifting news into a journalistic knowledge platform,' in *Proceedings of the CIKM 2020 Workshops*, 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), CEUR Workshop Proceedings, 2020.

[7] M. Gallofré Ocaña and A. L. Opdahl, 'Developing a software reference architecture for journalistic knowledge platforms,' in *Companion Proceedings of the 15th European Conference on Software Architecture (ECSA 2021)*, CEUR Workshop Proceedings, 2021.

[8] M. Gallofré Ocaña, 'Identifying events from streams of rdf-graphs representing news and social media messages,' in *The Semantic Web: ESWC 2021 Satellite Events*, Cham: Springer International Publishing, 2021, pp. 186–194.

[9] M. Gallofré Ocaña, T. Al-Moslmi and A. L. Opdahl, 'Knowledge graph semantic annotation and population with real-time events data from gdelt,' in *2022 IEEE 24th Conference on Business Informatics (CBI)*, vol. 02, 2022, pp. 65–72. DOI: 10.1109/CBI54897.2022.10050.

[10] M. Gallofré Ocaña, L. Nyre, A. L. Opdahl, B. Tessem, C. Trattner and C. Veres, 'Towards a big data platform for news angles,' in *Norwegian Big Data Symposium*, CEUR Workshop Proceedings, 2018.

[11] M. Albared, M. Gallofré Ocaña, A. Ghareb and T. Al-Moslmi, 'Recent progress of named entity recognition over the most popular datasets,' in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, 2019, pp. 1–9. DOI: 10.1109/ICOICE48418.2019.9035170.

[12] T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl and B. Tessem, 'Detecting newsworthy events in a journalistic platform,' in *The 3rd European Data and Computational Journalism Conference*, 2019, pp. 3–5.

[13] T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl and C. Veres, 'Named entity extraction for knowledge graphs: A literature overview,' *IEEE Access*, vol. 8, pp. 32 862–32 881, 2020. DOI: 10.1109/ACCESS.2020.2973928.

[14] A. L. Opdahl, T. Al-Moslmi, D.-T. Dang-Nguyen, M. Gallofré Ocaña, B. Tessem and C. Veres, 'Semantic knowledge graphs for the news: A review,' *ACM Comput. Surv.*, vol. 55, no. 7, Dec. 2022. DOI: 10.1145/3543508.

[15] B. Tessem, M. Gallofré Ocaña and A. L. Opdahl, 'Construction of a relevance knowledge graph with application to the LOCAL news angle,' in *Nordic Artificial Intelligence Research and Development*, 2023.

[16] N. Newman, R. Fletcher, A. Schulz, S. Andi, C. T. Robertson and R. K. Nielsen, 'Reuters institute digital news report 2021,' Reuters Institute for the Study of Journalism, Tech. Rep., 2021. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (visited on 11/01/2023).

[17] F. M. Simon and L. Graves, 'Pay models for online news in the us and europe: 2019 update,' Reuters Institute for the Study of Journalism, Tech. Rep., 2019. [Online]. Available: https://www.digitalnewsreport.org/publications/2019/pay-models-2019-update (visited on 11/01/2023).

[18] R. Fletcher and R. K. Nielsen, 'Paying for online news,' *Digital Journalism*, vol. 5, no. 9, 2017. DOI: 10.1080/21670811.2016.1246373.

[19] N. Newman, R. Fletcher, A. Kalogeropoulos and R. K. Nielsen, 'Reuters institute digital news report 2019,' Reuters Institute for the Study of Journalism, Tech. Rep., 2019. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/our-research/digital-news-report-2019 (visited on 11/01/2023).

[20] N. Newman, R. Fletcher, A. Schulz, S. Andi and R. K. Nielsen, 'Reuters institute digital news report 2020,' Reuters Institute for the Study of Journalism, Tech. Rep., 2020. [Online]. Available: https://www.digitalnewsreport.org/survey/2020 (visited on 11/01/2023).

[21] B. J. Toff, S. Badrinathan, C. Mont'Alverne, A. R. Arguedas, R. Fletcher and R. K. Nielsen, 'Overcoming indifference: What attitudes towards news tell us about building trust,' Reuters Institute for the Study of Journalism, Tech. Rep., 2021. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/overcoming-indifference-what-attitudes-towards-news-tell-us-about-building-trust (visited on 11/01/2023).

[22] N. Newman and R. Fletcher, 'Bias, bullshit and lies: Audience perspectives on low trust in the media,' Available at SSRN, Tech. Rep., 2017. DOI: 10.2139/ssrn.3173579.

[23] J. Vázquez Herrero, S. Direito-Rebollal, A. S. Rodríguez and X. García, *Journalistic Metamorphosis: Media Transformation in the Digital Age*. Springer Cham, 2020. DOI: 10.1007/978-3-030-36315-4.

[24] C. Beckett, 'New powers, new responsibilities: A global survey of journalism and artificial intelligence,' Polis, London School of Economics and Political Science, Tech. Rep., 2019. [Online]. Available: https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities (visited on 11/01/2023).

[25] S. C. Lewis and O. Westlund, 'Big data and journalism,' *Digital Journalism*, vol. 3, no. 3, 2015. DOI: 10.1080/21670811.2014.976418.

[26] J. Keefe, Y. Zhou and J. B. Merrill. 'The present and potential of AI in journalism.' (2021), [Online]. Available: https://knightfoundation.org/articles/the-present-and-potential-of-ai-in-journalism (visited on 10/10/2021).

[27] J. Hirschberg and C. D. Manning, 'Advances in natural language processing,' *Science*, vol. 349, no. 6245, pp. 261–266, 2015. DOI: 10.1126/science.aaa8685.

[28] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding,' in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[29] T. Brown *et al.*, 'Language models are few-shot learners,' *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] T. Mikolov, K. Chen, G. Corrado and J. Dean, *Efficient estimation of word representations in vector space*, 2013. DOI: 10.48550/ARXIV.1301.3781.

[31] A. Vaswani *et al.*, 'Attention is all you need,' in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[32] Aidan Hogan, et al., 'Knowledge graphs,' *ACM Comput. Surv.*, vol. 54, no. 4, Jul. 2021. DOI: `10.1145/3447772`.

[33] S. Martínez-Fernández *et al.*, 'Software engineering for ai-based systems: A survey,' *ACM Trans. Softw. Eng. Methodol.*, vol. 31, no. 2, Apr. 2022. DOI: `10.1145/3487043`.

[34] A. d. Garcez and L. C. Lamb, 'Neurosymbolic ai: The 3rd wave,' *Artificial Intelligence Review*, 2023. DOI: `10.1007/s10462-023-10448-w`.

[35] C. Bizer, T. Heath and T. Berners-Lee, 'Linked data: The story so far,' in *Semantic services, interoperability and web applications: emerging concepts*, IGI global, 2011, pp. 205–227.

[36] E. Motta, E. Daga, A. L. Opdahl and B. Tessem, 'Analysis and design of computational news angles,' *IEEE Access*, 2020.

[37] N. Shadbolt, T. Berners-Lee and W. Hall, 'The semantic web revisited,' *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006. DOI: `10.1109/MIS.2006.62`.

[38] T. Berners-Lee, J. Hendler and O. Lassila, 'The semantic web,' *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001. [Online]. Available: `http://www.jstor.org/stable/26059207` (visited on 11/01/2023).

[39] J. Domingue and E. Motta, 'Planetonto: From news publishing to integrated knowledge management support,' *IEEE Intelligent Systems and their Applications*, vol. 15, no. 3, pp. 26–32, 2000. DOI: `10.1109/5254.846282`.

[40] P. Castells, F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras and J. Lorés, 'Neptuno: Semantic web technologies for a digital newspaper archive,' in *The Semantic Web: Research and Applications. ESWS 2004.*, 2004. DOI: `10.1007/978-3-540-25956-5_31`.

[41] D. B. Ramagem, B. Margerin and J. Kendall, 'Annoterra: Building an integrated earth science resource using semantic web technologies,' *IEEE Intelligent Systems*, vol. 19, no. 3, 2004. DOI: `10.1109/MIS.2004.3`.

[42] A. Java, T. Finin and S. Nirenburg, 'Semnews: A semantic news framework,' in *The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, 2006. [Online]. Available: `https://www.aaai.org/Papers/AAAI/2006/AAAI06-316.pdf` (visited on 11/01/2023).

[43] F. Frasincar, J. Borsje and L. Levering, 'A semantic web-based approach for building personalized news services,' *International Journal of E-Business Research (IJEBR)*, vol. 5, no. 3, pp. 35–53, 2009.

[44] G. Kobilarov *et al.*, 'Media meets semantic web – how the BBC uses DBpedia and linked data to make connections,' in *The Semantic Web: Research and Applications. ESWC 2009.*, vol. 5554, Berlin, Heidelberg, 2009. DOI: 10.1007/978-3-642-02121-3_53.

[45] Y. Raimond, T. Scott, S. Oliver, P. Sinclair and M. Smethurst, 'Use of semantic web technologies on the BBC web sites,' in *Linking Enterprise Data*. Boston, MA, 2010. DOI: 10.1007/978-1-4419-7665-9_13.

[46] N. Fernández, D. Fuentes, L. Sánchez and J. A. Fisteus, 'The NEWS ontology: Design and applications,' *Expert Systems with Applications*, vol. 37, no. 12, 2010. DOI: 10.1016/j.eswa.2010.06.055.

[47] G. Leban, B. Fortuna, J. Brank and M. Grobelnik, 'Event registry: Learning about world events from news,' in *Proceedings of the 23rd International Conference on World Wide Web*, 2014. DOI: 10.1145/2567948.2577024.

[48] P. Vossen *et al.*, '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*, vol. 110, 2016. DOI: 10.1016/j.knosys.2016.07.013.

[49] X. Liu *et al.*, 'Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter,' in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '16, Indianapolis, Indiana, USA: Association for Computing Machinery, 2016, pp. 207–216. DOI: 10.1145/2983323.2983363.

[50] X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin and J. Duprey, 'Reuters tracer: Toward automated news production using large scale social media data,' in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, 2017, pp. 1483–1493.

[51] U. Germann, R. Liepins, G. Barzdins, D. Gosko, S. Miranda and D. Nogueira, 'The SUMMA platform: A scalable infrastructure for multi-lingual multi-media monitoring,' in *Proceedings of ACL 2018, System Demonstrations*, 2018. DOI: 10.18653/v1/P18-4017.

[52] N. Maiden *et al.*, 'Making the news: Digital creativity support for journalists,' in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–11. DOI: 10.1145/3173574.3174049.

[53] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy and X. Tannier, 'Searching news articles using an event knowledge graph leveraged by wikidata,' in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019. DOI: 10.1145/3308560.3316761.

[54] S. Angelov, P. Grefen and D. Greefhorst, 'A framework for analysis and design of software reference architectures,' *Information and Software Technology*, vol. 54, no. 4, 2012. DOI: 10.1016/j.infsof.2011.11.009.

[55] A. Berven, O. A. Christensen, S. Moldeklev, A. L. Opdahl and K. J. Villanger, 'A knowledge-graph platform for newsrooms,' *Computers in Industry*, vol. 123, 2020. DOI: 10.1016/j.compind.2020.103321.

[56] N. Marz. 'How to beat the cap theorem,' Thoughts from the Red Planet. (2011), [Online]. Available: http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html (visited on 08/03/2023).

[57] J. Kreps. 'Questioning the lambda architecture,' O'Reilly.com. (Jul. 2014), [Online]. Available: https://www.oreilly.com/radar/questioning-the-lambda-architecture (visited on 12/02/2023).

[58] R. C. Fernandez *et al.*, 'Liquid: Unifying nearline and offline big data integration,' in *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2015.

[59] S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren and D. Valerio, 'A software reference architecture for semantic-aware big data systems,' *Information and Software Technology*, vol. 90, 2017. DOI: 10.1016/j.infsof.2017.06.001.

[60] B. Sena, A. P. Allian and E. Y. Nakagawa, 'Characterizing big data software architectures: A systematic mapping study,' in *Proceedings of the 11th Brazilian Symposium on Software Components, Architectures, and Reuse*, New York, NY, USA: Association for Computing Machinery, 2017. DOI: 10.1145/3132498.3132510.

[61] B. Sena, L. Garcés, A. P. Allian and E. Y. Nakagawa, 'Investigating the applicability of architectural patterns in big data systems,' in *Proceedings of the 25th Conference on Pattern Languages of Programs*, ser. PLoP '18, Portland, Oregon: The Hillside Group, 2020. DOI: 10.5555/3373669.3373677.

[62] C. Avci, B. Tekinerdogan and I. N. Athanasiadis, 'Software architectures for big data: A systematic literature review,' *Big Data Analytics*, vol. 5, no. 1, pp. 1–53, 2020.

[63] P. Ataei and A. T. Litchfield, 'Big data reference architectures, a systematic literature review,' in *ACIS 2020 Proceedings.*, 2020. [Online]. Available: https://aisel.aisnet.org/acis2020/30 (visited on 11/01/2023).

[64] T. V. R. da Costa, E. Cavalcante and T. Batista, 'Big data software architectures: An updated review,' in *Computational Science and Its Applications – ICCSA 2022*, Cham: Springer International Publishing, 2022, pp. 477–493. DOI: 10.1007/978-3-031-10522-7_33.

[65] D. Le-Phuoc, H. Q. Nguyen-Mau, J. X. Parreira and M. Hauswirth, 'A middleware framework for scalable management of linked streams,' *Journal of Web Semantics*, vol. 16, 2012, The Semantic Web Challenge 2011. DOI: 10.1016/j.websem.2012.06.003.

[66] M. A. Martínez-Prieto, C. E. Cuesta, M. Arias and J. D. Fernández, 'The SOLID architecture for real-time management of big semantic data,' *Future Generation Computer Systems*, vol. 47, 2015, Special Section: Advanced Architectures for the Future Generation of Software-Intensive Systems, ISSN: 0167-739X. DOI: 10.1016/j.future.2014.10.016.

[67] P. Pääkkönen and D. Pakkala, 'Reference architecture and classification of technologies, products and services for big data systems,' *Big Data Research*, vol. 2, no. 4, 2015, ISSN: 2214-5796. DOI: 10.1016/j.bdr.2015.01.001.

[68] D. Xu, D. Wu, X. Xu, L. Zhu and L. Bass, 'Making real time data analytics available as a service,' in *Proceedings of the 11th International ACM SIGSOFT Conference on Quality of Software Architectures*, ser. QoSA '15, Montréal, QC, Canada: Association for Computing Machinery, 2015, pp. 73–82. DOI: 10.1145/2737182.2737186.

[69] G. M. Sang, L. Xu and P. de Vrieze, 'A reference architecture for big data systems,' in *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, 2016, pp. 370–375. DOI: 10.1109/SKIMA.2016.7916249.

[70] L. Heilig and S. Voß, 'Managing cloud-based big data platforms: A reference architecture and cost perspective,' in *Big Data Management.* Cham: Springer International Publishing, 2017, pp. 29–45. DOI: 10.1007/978-3-319-45498-6_2.

[71] N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin and L. Safina, 'Microservices: Yesterday, today, and tomorrow,' in *Present and Ulterior Software Engineering.* Springer International publishing, 2017. DOI: 10.1007/978-3-319-67425-4_12.

[72] L. D. Erman, F. Hayes-Roth, V. R. Lesser and D. R. Reddy, 'The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty,' *ACM Comput. Surv.*, vol. 12, no. 2, pp. 213–253, Jun. 1980. DOI: 10.1145/356810.356816.

[73] H. P. Nii, 'The blackboard model of problem solving and the evolution of blackboard architectures,' *AI Magazine*, vol. 7, no. 2, p. 38, Jun. 1986. DOI: 10.1609/aimag.v7i2.537.

[74] H. P. Nii, 'Blackboard application systems, blackboard systems and a knowledge engineering perspective,' *AI Magazine*, vol. 7, no. 3, p. 82, Jul. 1986. DOI: 10.1609/aimag.v7i3.550.

[75] T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997, ISBN: 0070428077.

[76] Y. A. Malkov and D. A. Yashunin, 'Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, Apr. 2020. DOI: 10.1109/TPAMI.2018.2889473.

[77] J. Johnson, M. Douze and H. Jégou, 'Billion-scale similarity search with GPUs,' *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[78] World Wide Web Consortium (W3C). 'RDF 1.1 Concepts and Abstract Syntax.' R. Cyganiak, D. Wood and M. Lanthaler, Eds. (Feb. 2014), [Online]. Available: http://www.w3.org/TR/rdf11-concepts (visited on 11/01/2023).

[79] World Wide Web Consortium (W3C). 'SPARQL 1.1 Query Language.' S. Harris and A. Seaborne, Eds. (Mar. 2013), [Online]. Available: https://www.w3.org/TR/sparql11-overview (visited on 11/01/2023).

[80] H. A. Simon, *The sciences of the artificial.* MIT press, 1996.

[81] A. Hevner and S. Chatterjee, 'Design science research in information systems,' in *Design Research in Information Systems: Theory and Practice.* Boston, MA: Springer US, 2010. DOI: 10.1007/978-1-4419-5653-8_2.

[82] A. R. Hevner, 'A three cycle view of design science research,' *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4, 2007.

[83] A. R. Hevner, S. T. March, J. Park and S. Ram, 'Design science in information systems research,' *MIS Quarterly*, vol. 28, no. 1, 2004. [Online]. Available: http://www.jstor.org/stable/25148625 (visited on 11/01/2023).

[84] B. Kitchenham, 'Procedures for performing systematic reviews,' Software Engineering Group, Department of Computer Science, Keele University, Empirical Software Engineering (TR/SE-0401) and National ICT Australia Ltd (0400011T.1), Tech. Rep., 2004.

[85] K. M. Eisenhardt, 'Building theories from case study research,' *The Academy of Management Review*, vol. 14, no. 4, pp. 532–550, 1989. [Online]. Available: http://www.jstor.org/stable/258557 (visited on 11/01/2023).

[86] M. Galster and P. Avgeriou, 'Empirically-grounded reference architectures: A proposal,' in *Proceedings of the Joint ACM SIGSOFT Conference – QoSA and ACM SIGSOFT Symposium – ISARCS on Quality of Software Architectures – QoSA and Architecting Critical Systems – ISARCS*, Association for Computing Machinery, 2011. DOI: 10.1145/2000259.2000285.

[87] C. Hoon, 'Meta-synthesis of qualitative case studies: An approach to theory building,' *Organizational Research Methods*, vol. 16, no. 4, pp. 522–556, 2013. DOI: 10.1177/1094428113484969.

[88] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC, 2014.

[89] J. Cohen, 'A coefficient of agreement for nominal scales,' *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, ISSN: 1552-3888. DOI: 10.1177/001316446002000104.

[90] N. Wongpakaran, T. Wongpakaran, D. Wedding and K. L. Gwet, 'A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples,' *BMC medical research methodology*, vol. 13, pp. 1–7, 2013.

[91] A. L. Opdahl and B. Tessem, 'Ontologies for finding journalistic angles,' *Software and Systems Modeling*, vol. 20, pp. 71–87, 2021.

[92] S. Hellmann, J. Lehmann, S. Auer and M. Brümmer, 'Integrating nlp using linked data,' in *The Semantic Web – ISWC 2013*, H. Alani *et al.*, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 98–113, ISBN: 978-3-642-41338-4.

[93] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, 'High-resolution image synthesis with latent diffusion models,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 10 684–10 695. DOI: 10.1109/CVPR52688.2022.01042.

[94] R. Bommasani *et al.*, *On the opportunities and risks of foundation models*, 2022. arXiv: 2108.07258 [cs.LG].

[95] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, 'Translating embeddings for modeling multi-relational data,' in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf (visited on 11/01/2023).

[96] P. Ristoski and H. Paulheim, 'RDF2Vec: RDF graph embeddings for data mining,' in *The Semantic Web – ISWC 2016*, Cham: Springer International Publishing, 2016, pp. 498–514. DOI: 10.1007/978-3-319-46523-4_30.

[97] H. Touvron *et al.*, *Llama: Open and efficient foundation language models*, 2023. arXiv: `2302.13971 [cs.CL]`.

[98] D. Stammbach, M. Antoniak and E. Ash, 'Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data,' in *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 47–56. DOI: `10.18653/v1/2022.wnu-1.6`.

[99] M. Gallofré Ocaña, A. L. Opdahl and D.-T. Dang-Nguyen, 'Emerging news task: Detecting emerging events from social media and news feeds,' MediaEval, 2021.

Prototype

As part of this thesis, I developed a proof-of-concept prototype of a JKP. To implement the prototype, I followed an iterative process to continuously test and improve my decisions. The prototype design (Figure 1) is based on the proposed SRA for JKPs.

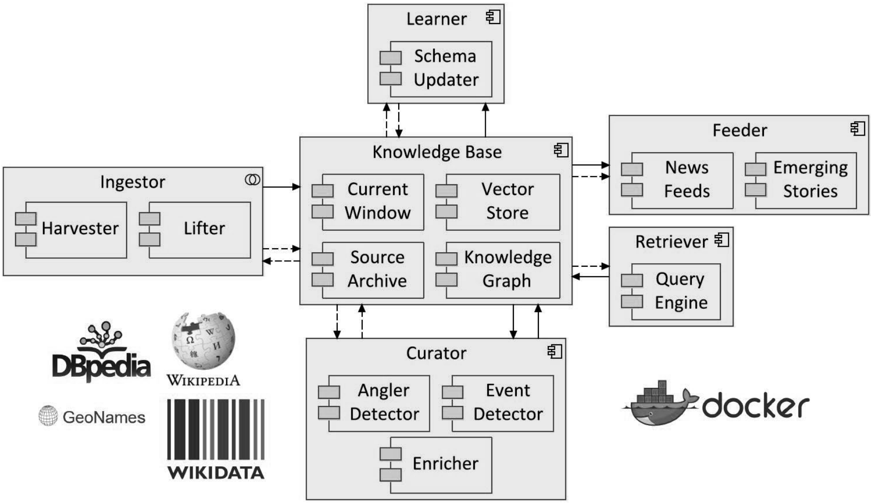


Figure 1: The instantiated architecture for the JKP prototype.

The prototype implements various components that enable the ingestion of news-related data from different sources. This includes crawlers for harvesting websites, RSS feeds, Twitter accounts, NewsAPI, and GDELT. The Twitter API provides real-time tweet streams that can be filtered by specific accounts, geographical areas and topics. NewsAPI aggregates and provides streams of news articles from more than 80.000 sources and blogs. GDELT provides semi-structured information about conflict events collected from news worldwide that are automatically translated into English from 65 different languages. By incorporating these sources, the prototype can simulate a real-world situation where large volumes of diverse news-related data are ingested in real-time.

The harvested news items and events are transformed into semantic knowledge representations in real time according to an event-description ontology [91]. This is achieved through a lifting process that combines out-of-the-box NLP systems such as DBpedia Spotlight[4] and SpaCy[5] with different end-to-end deep learning models to semantically annotate the potentially news-relevant textual items. The lifting process follows the blackboard model described in the Manuscript III and employs named entity linking, relation extraction, coreference resolution, and natural language inference models to extract and annotate named entities, relations, sentiments and topics. The extracted annotations are linked to external knowledge bases such as Wikidata[6], DBpedia[7] and Geonames[8]. As a result, the annotation system generates RDF graphs that represent the news items and events.

The news items and their graph representations are stored in a knowledge base that consists of several components. The raw text of news items is archived in Apache Cassandra, while the graph representations are stored in the knowledge graph implemented with Blazegraph. A current view of the last incoming news items and graph representation is provided by Apache Kafka and ksqlDB. The infrastructure is illustrated in Figure 2. The Apache Cassandra database is deployed as a cluster of three nodes, and Blazegraph is distributed over four instances. One instance is dedicated to storing news-relevant items from news articles and Twitter messages, while the other three store events from GDELT. These four instances ingest approximately 11M ($11 \cdot 10^6$) triples daily from news and tweets, and another 11M ($11 \cdot 10^6$) from GDELT events. In six months, the Blazgraph cluster can ingest more than 4B ($6 \cdot 10^9$) triples in total. Apache Kafka and ksqldBD are distributed over three instances each and can be configured with, for example, a retention policy of one week to provide a current view with a window of a week. In addition, I deployed a mongoDB instance to store metadata related to the management and configuration of the prototype. All instances and components are deployed as containerised applications using Docker.

To enhance the generated graph representations, I developed an Enricher, an Event detector and an Angle detector. The Enricher expands the annotated items by incorporating geographical information extracted from DBpedia, Wikidata and LOD sources. The Event Detector provides aggregated real-time events pulled from GDELT streams. The Angle Detector analyses the representations of the news items and identifies relevant location angles for a set of locations of interest [15]. In addition, I developed a Schema updater that scans incoming GDELT events for new themes and updates an internal hierarchy in the knowledge graph that semantically represents these themes.

[4] DBpediaSpotlight: `www.dbpedia-spotlight.org`

[5] Spacy: `spacy.io`

[6] Wikidata: `www.wikidata.org`

[7] DBpedia: `dbpedia.org`

[8] Geonames: `www.geonames.org`

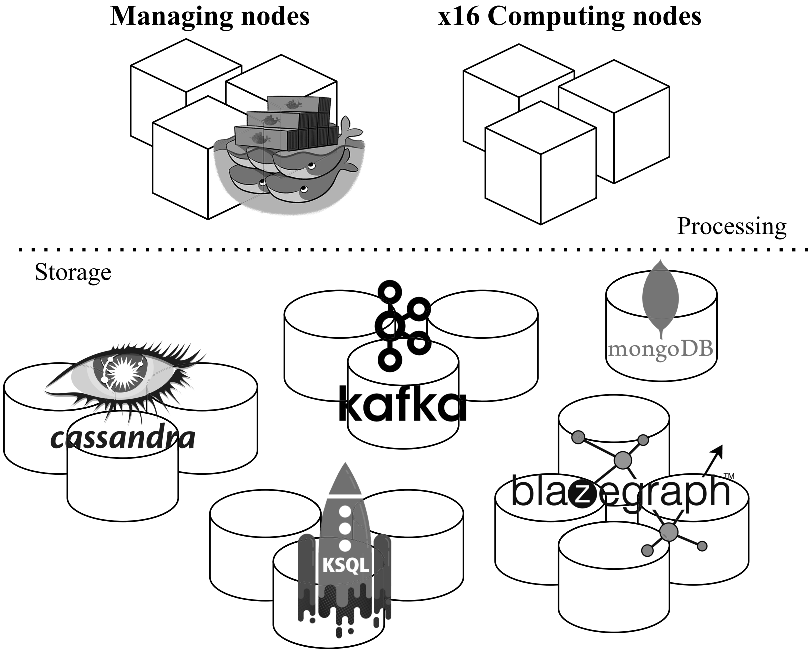


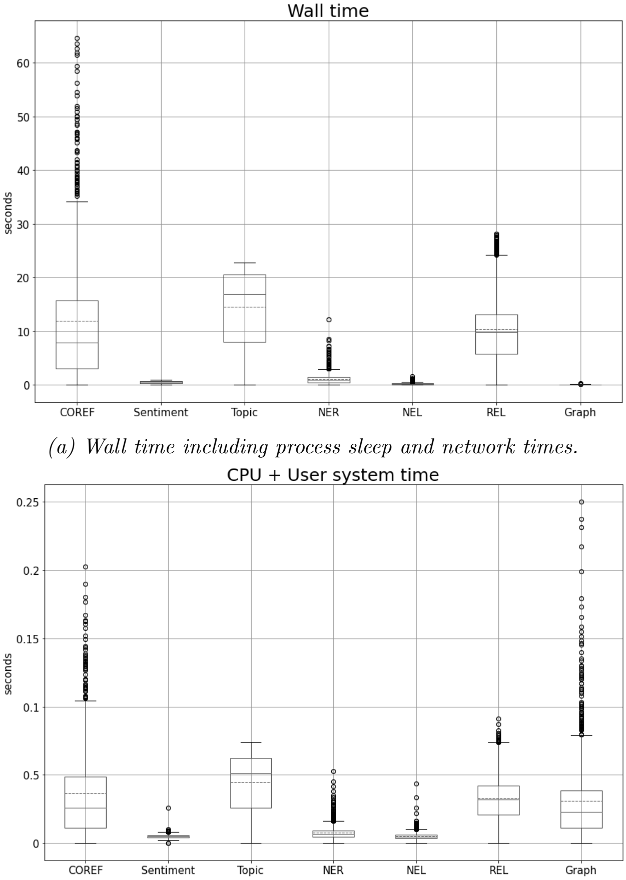
Figure 2: Infrastructure resources

The prototype provides an API that offers an annotated news feed from the knowledge base and enables external users to engage with the system. To explore co-development with external contributors, I organised a research challenge called EmergingNews[9] where participants were invited to submit solutions using live feeds directly from the knowledge base [99]. Moreover, I collaborated in another research challenge[10] where participants had access to a sample of text and images from my prototype to explore the connection between news and images. The knowledge base also provides several APIs to access the different data stores and other services of the system. Additionally, the prototype also offers an editing interface designed to suggest relevant news, background information and local angles related to the story the journalist is typing in.

The communication between components in the prototype is facilitated by Apache Kafka, which serves as a message broker. Services can communicate with one another by producing and consuming messages from a message stream. To structure the messages in order to facilitate understanding and integration, I used *JSON-LD*[11] and semantic vocabularies. JSON-LD is a widely used extension of JSON for serialising linked data.

[9] EmergingNews: `multimediaeval.github.io/editions/2021/tasks/emergingnews`
[10] NewsImages: `multimediaeval.github.io/editions/2022/tasks/newsimages`
[11] JSON-LD: `www.w3.org/TR/json-ld11`

The prototype runs on 33 cloud instances with a total of 99 vCPU, 324GB RAM and 20TB disk space. To automate the setup process, I used Ansible and Terraform. Additionally, I employed Docker Swarm to orchestrate a total of 114 microservices running as containerised services. Of these services, 70 are associated with the JKP, while the remainder are responsible for monitoring the services and cloud instances. The services are available via APIs, facilitating their replacement with newer versions without impacting the performance of the platform. I decoupled the ML and DL models from the services by exposing them via APIs, enabling changing and updating the models at any time. The prototype gathers news items and transforms them into graphs in an average time of 21.112 seconds per news item, this includes sleep and network waiting times. When considering the system and user CPU time alone, the average time drops to 0.246 seconds per item. Figure 3 shows the times for each step of the annotation process. The generated graphs contain an average of 712 triples.



(a) Wall time including process sleep and network times.

(b) CPU and User system process time.

Figure 3: Times for each NLP step for annotating news items using the blackboard approach.

Manuscripts

Manuscript I

Supporting Newsrooms with Journalistic Knowledge Graph Platforms: Current State and Future Directions

Marc Gallofré Ocaña and Andreas L. Opdahl

Technologies, **10**, 68. (2022)

technologies

Review

Supporting Newsrooms with Journalistic Knowledge Graph Platforms: Current State and Future Directions [†]

Marc Gallofré Ocaña * and **Andreas L. Opdahl**

Department of Information Science and Media Studies, University of Bergen, 5020 Bergen, Norway; andreas.opdahl@uib.no

* Correspondence: marc.gallofre@uib.no

[†] This paper is an extended version of our paper published in Proceedings of the CIKM 2020 Workshops.

Abstract: Increasing competition and loss of revenues force newsrooms to explore new digital solutions. The new solutions employ artificial intelligence and big data techniques such as machine learning and knowledge graphs to manage and support the knowledge work needed in all stages of news production. The result is an emerging type of intelligent information system we have called the Journalistic Knowledge Platform (JKP). In this paper, we analyse for the first time knowledge graph-based JKPs in research and practice. We focus on their current state, challenges, opportunities and future directions. Our analysis is based on 14 platforms reported in research carried out in collaboration with news organisations and industry partners and our experiences with developing knowledge graph-based JKPs along with an industry partner. We found that: (a) the most central contribution of JKPs so far is to automate metadata annotation and monitoring tasks; (b) they also increasingly contribute to improving background information and content analysis, speeding-up newsroom workflows and providing newsworthy insights; (c) future JKPs need better mechanisms to extract information from textual and multimedia news items; (d) JKPs can provide a digitalisation path towards reduced production costs and improved information quality while adapting the current workflows of newsrooms to new forms of journalism and readers' demands.

Keywords: journalistic knowledge platform; artificial intelligence; knowledge graph; intelligent information system; newsrooms; journalism

Citation: Gallofré Ocaña, M.; Opdahl, A.L. Supporting Newsrooms with Journalistic Knowledge Graph Platforms: Current State and Future Directions. *Technologies* **2022**, *10*, 68. https://doi.org/10.3390/technologies10030068

Academic Editor: Manoj Gupta

Received: 10 April 2022
Accepted: 26 May 2022
Published: 31 May 2022

1. Introduction

News agencies and news organisations are under pressure from the loss of advertisement and revenues [1,2], and facing an audience that is less likely willing to pay for digital content [3,4]. Despite an increase in digital consumption, information is no longer consumed from a limited number of TV stations and news outlets. Instead, readers have access to and can contrast fresh and first-hand information from free-available sources on the internet and social media at any time. As a consequence of their freedom of choice, readers demand high-quality journalism [5] and trusted sources [4,6,7].

In response, news agencies and news organisations are constantly adapting their business models to digital media innovations in order to improve information quality, competitiveness and growth [8]. Innovation and digitalisation of newsrooms are needed to increase the quality and lower the cost of news production, changing how journalists and readers interact with news content and background information [9]. Newsrooms are therefore embracing big data and artificial intelligence (AI) techniques such as knowledge graphs and machine learning (ML) for journalistic purposes [10,11] such as identifying and contextualising newsworthy events in investigative journalism; facilitating data visualisation in digital journalism; analysing information in data journalism; automating news writing in robot journalism; providing real-time fact-checking tools for political journalism. The result is an emerging type of intelligent information system that we call the *Journalistic*

Knowledge Platform (JKP) which is currently gaining interest in research and practice. In this paper, we define JKPs as platforms that apply AI and big data to journalism in order to manage and support the knowledge work needed in all stages of news production.

JKPs can be described from a functional, an organisational and a technical perspective. From a functional point of view JKPs automate the process of annotating metadata and support daily workflows like news production [12,13], archiving [14,15], management [16,17] and distribution [18–21]. JKPs harvest and analyse news and social media information over the net in real time [22], leverage encyclopaedic sources [23], and provide journalists with both meaningful background knowledge [24] and newsworthy information [25]. From an organisational viewpoint: JKPs are deployed in newsrooms to manage the knowledge needed to support journalists with creativity and discovery tasks. These are tailored to the particular digital strategies and editorial lines to improve news broadcast. JKPs also follow media standards to facilitate communication with customers and providers, and are subject to legal regulations such as data privacy. From a technical perspective JKPs implement state-of-the-art AI technologies such as machine learning, natural language processing (NLP) and knowledge representation and reasoning. News-relevant information is represented in knowledge bases which are exploited with data analysis, reasoning and information retrieval techniques to help journalists and readers dive more deeply into information, events and storylines. Today, knowledge graphs [26] are a topical technique for knowledge representation that continues to grow in importance, therefore, we centre our analysis on JKPs building on knowledge graphs.

According to the authors of Hogan et al. [26], knowledge graphs capture and abstract knowledge using graph-based data models. They are particularly relevant for scenarios that integrate and extract value from diverse and dynamic data. Wherein entities of interest are represented as nodes and the relations between them as edges of the graph. Ontologies and rules are used to define the semantics and terms of the graph and reason about it, but also to ease data integration from, for example, Linked Open Data (LOD) [27] and existing large-scale knowledge graphs like Wikidata and DBpedia. Compared to relational and NoSQL models, knowledge graphs facilitate semantic integration, flexible data and schema evolution and graph query languages with mechanisms to explore complex relations through arbitrary-length paths.

In this article, we explore the current state and suggest future research directions for knowledge graph-based JKPs. We ask: “What challenges and opportunities for newsrooms have motivated the knowledge graph-based JKPs?” (RQ1), “How does the research on knowledge graph-based JKPs address these challenges and opportunities?” (RQ2) and “What are the most important open areas for research on knowledge graph-based JKPs?” (RQ3). To answer these three questions we have performed a detailed analysis of 14 JKPs reported in the literature that apply AI and big data to journalism in order to manage and support the knowledge work needed in all stages of news production. A broader literature on related technologies exists. Our analysis does not ignore other solutions applying artificial intelligence to journalism, but our focus is on providing a comprehensive analysis of the main concepts of those JKPs that build on knowledge graphs rather than specific techniques, optimisations, tools and systems. The JKPs were selected in context of a broader systematic literature review on how knowledge graphs can support news in a wide sense [28]. Compared to this study, Opdahl et al. [28] was not restricted to JKPs and did not analyse the challenges and opportunities nor the current and future directions of JKPs. We conducted a qualitative meta-analysis (see Appendix A for a detailed description of the meta-analysis method), and we examined the existing JKPs in light of our experiences with developing JKPs along with an industry partner for the international newsroom market [29].

The present article extends Gallofré Ocaña and Opdahl [30], which analyses challenges and opportunities for developing JKPs along six axes: stakeholders, information, functionalities, techniques, components and concerns. This article extends the analysis by considering more JKPs. It investigates how well the challenges and opportunities are covered in the research literature and suggests future research directions. The rest of the paper is organised as follows: we summarise the identified JKPs in Section 2; analyse the current challenges and opportunities for newsrooms that motivated JKPs in Section 3; present the state of research on JKPs in terms of their stakeholders, information, functionalities, techniques, components and concerns in Section 4; discuss the future directions for research on JKPs in Section 5.

2. Analysed Platforms

We identified 14 platforms that fit under our definition of JKPs, which we list in Table 1. The identified JKPs cover a total of 28 papers carried out by distinct research groups located in 11 different countries and in collaboration with a variety of news agencies, news organisations and industry partners.

Table 1. Selected platforms. N: news media partner and T: technology partner. The identified countries represent the news media partners' countries.

Platform	Industry partners	Countries	References
PlanetOnto	-	UK	[18,31]
Neptuno	Diari SEGRE^{N} and iSOCO^{T}	Spain	[14]
AnnoTerra	NASA's Earth Observatory^{N}	USA	[24]
SemNews *	-	USA	[19]
Hermes *	-	The Netherlands	[20,32,33]
BBC CMS	BBC^{N}	UK	[16,34]
NEWS	Agencia EFE^{N}, Agencia ANSA^{N} and Ontology Ldt.^{T}	Spain and Italy	[12,35]
EventRegistry *	-	Slovenia	[21]
NewsReader *	LexisNexis^{T}, The Sensible Code Company (before ScraperWiki)^{T} and Synerscope^{T}	Netherlands, Spain and Italy	[15,36,37]
Reuters Tracer	Reuters^{N}	USA	[22,38,39]
SUMMA	LETA^{N}, BBC Monitoring^{N}, Deutsche Welle^{N} and Priberam Labs^{T}	Latvia, UK, Germany	[17,40–42]
INJECT	Adresseavisen^{N}, AFP^{N}, The Globe and Mail^{N}, Stibo^{T}	Norway, France, Canada	[13]
ASRAEL	AFP^{N}	France	[23]
News Hunter ‡	Wolftech^{T}	Norway	[29,43,44]

‡ News Hunter is the JKP in which the authors are involved. * Related systems that can be used as JKPs—either directly or with some adaptations—but have not been published in the context of newsrooms.

The JKPs from 2000 to early 2010 implemented the Semantic Web idea [45] in newsrooms. These JKPs used semantic web technologies [46] to automate the metadata annotation process [16], combine different knowledge bases [24], and formalise media standards [14]. They used ontologies in NLP pipelines together with Linked Open Data (LOD) [27] resources from external knowledge bases (i.e., Wikipedia, DBpedia) to automatically annotate news archives and feeds with metadata about topics, keywords, categories and other relevant information (e.g., persons, places, organisations, sentiments and relations) [14,19]. The annotated information was stored in knowledge bases, facilitating the interlinking of news across different archives, online catalogues and external LOD repositories [16,24]. For instance, Neptuno was the first project to publish a journalistic ontology and adapt the IPTC topics [47] as RDF [14,48], and Troncy [49] converted the IPTC NewsCodes [50] into SKOS [51] thesaurus and defined an OWL [52] ontology for the IPTC News Architecture [53]. The resulting systems provided services for supporting news creation [14], personalising news retrieval [18,20], facilitating semantic search [14,18,19,24,33], visualising ontologies [14], managing content [16], aggregating information [24,34] and recommending news [16].

The JKPs from early 2010s until today focused on identifying and analysing events and advancing AI/ML for supporting journalism. In addition, some of them focused on scaling over large volumes of live streams of multimedia news [36], social media [39] and TV/radio broadcasts [17]. Similar to the previous JKPs, news items were annotated using either media standards [35] and LOD resources [37] or both [23] and stored in knowledge bases to facilitate cross-lingual information retrieval services through semantic technologies and ontologies [21,23]. These JKPs continuously monitored and curated the annotated items using AI/ML and LOD to provide relevant insights for journalists and identify current, past and future events. For example, the annotated news items were used to identify networks of actors [15], suggest news angles [13,54], automate news creation [22] and facilitate fact-checking [17], and the events were analysed using different AI/ML techniques for grouping events and news items [21,23], reasoning over events, and reconstructing the evolution of the events along time [15].

3. Challenges and Opportunities Facing Newsrooms

In current newsroom workflows, metadata annotation like tagging and categorisation is often performed manually by journalists. This is a time-consuming process that is error-prone, imprecise and restricts future usability [12]. The added metadata is reduced to a few general categories that are limited to authorship, dates, content language and news management information. This metadata is used to address newsworthiness and filter events according to news customers' and audiences' interests. However, due to the lack of fine-grained annotations, newsrooms have difficulties implementing high-quality information retrieval and filtering services [14,16,20]. Hence, they return irrelevant, incomplete and even biased results to customers [21].

Journalists spend a lot of their time monitoring and filtering large volumes of news feeds like TV broadcasts, radio shows, social media and published news to keep them up-to-date, time that otherwise would have been invested in producing news [42]. Today's worldwide daily news volumes scale over 100,000 articles making it unfeasible for journalists to manually handle tasks like fact-checking and searching for related articles. Germann et al. [41] (p. 1) claim that "each of [BBC] ca. 300 monitoring journalists usually keeps track up to 4 live sources in parallel (typically TV channels received via satellite), plus a number of other sources of information such as social media feeds". This is an undesired situation for a business sector where time is a critical factor, delays can lower the value of information and imply economic losses [35].

This massive volume of textual and multimedia data is often organised in different catalogues or databases and managed by external services [24,35]. Because these catalogues are not integrated nor share a common schema and lack fine-grained annotations, they limit the possibilities for newsrooms to extract valuable insights and knowledge. Structuring the information and integrating the data from a variety of sources bring newsrooms with better ways to exploit data and facilitate the adoption of AI. For example, it can ease the implementation of information retrieval services and recommender systems and the automation of news creation processes and the detection of fake news and newsworthy events.

To help with these processes, newsrooms currently use a mix of proprietary systems, external services, tools and in-house taxonomies or categorisation schemas that are challenging to integrate and operate together [35,55]. It is a complex ecosystem of applications that hinders the expansion and evolution of digitally integrated newsrooms. It makes it difficult for managers to get an overview of what is happening in news rooms [41]. It limits the interaction with customers [35]. Additionally, it can lead to vendor binding or dependence situations due to the difficulties of maintaining multiple and diverse proprietary solutions. All together and with the urge of reducing cost, increasing high-quality journalism and adapting current newsrooms to digital advances, journalists and newsrooms are becoming interested in the services that JKPs can offer [9].

4. State of Research on JKPs

We describe the state of research on JKPs by investigating the stakeholders, information, functionalities, techniques, components and concerns dealt with in the identified JKPs. These six analysis axes are based on a qualitative analysis reported in an earlier paper [30].

4.1. Stakeholders

JKPs provide services to and interact with a large variety of stakeholders. Figure 1 shows the identified stakeholders and their three top-level categories: general user, organisation and technical agent.

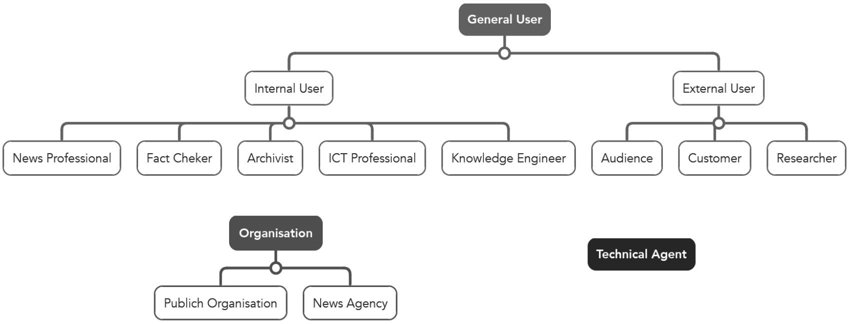


Figure 1. Stakeholder categories.

The general users 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 [35,39]; fact-checkers who conduct an essential task in combating with fake news and misinformation [17]; archivists who maintain up-to-date the schemas and news archives [14]; ICT professionals and knowledge engineers who develop and maintain JKPs [12]. Whereas, the external users are the audience [21]; the customers to whom new agencies offer services and researchers who investigate JKPs or use JKP to analyse data, as in the SUMMA project where "[political scientists want] to perform data analyses based on large amounts of news reports" [42] (p. 2).

JKPs support organisations in different ways: The most direct is in news agencies and news organisations where JKPs are deployed and adapted to particular digital strategies and purposes, but also to other news organisations that consume services from external JKPs. Moreover, JKPs provide services to both private and public organisations like governmental agencies that interact with or consume services from newsrooms, for example, the SUMMA project "provides media monitoring and analysis services to [. . .] the British government" [42] (p. 1). JKPs also interact indirectly with the organisations responsible for controlling news media standards, vocabulary and ontologies (e.g., the IPTC organisation). This impacts how JKPs are designed because the work of many news agencies depends on those standards, and JKPs often need to build on and comply with them. However, the media standards may not cover or fit the use cases of newsrooms, as in the NEWS project where "most of the NewsCodes defined by IPTC do not have alternative versions in different languages, only in English" [35] (p. 9). Hence, JKPs need to adapt or expand the media standards according to their needs.

Last but not least, the technical agent represents the JKPs and any system or technical infrastructure in newsrooms that support or interact with JKPs. A sub-type of the technical agent is the external system that communicates with newsroom services, like the customers' information systems [35].

4.2. Information

JKPs cover the whole news production pipeline from gathering information and news creation to knowledge exploitation and distribution. Table 2 lists the identified categories of information.

Table 2. The most common types of information managed by JKPs.

Information	Explanation
News content	The reported story or event.
Textual data	Textual information.
Multimedia data	Images, videos and audio information.
Data format	The format in which the data is stored or structured.
Metadata	Data about or that describe the news content.
Linked Open Data (LOD)	Structured and open available data on the Internet (e.g., data from Wikidata and DBpedia) [27]
Events	Newsworthy happenings.
Information needs	Different information types and categories of interest.

JKPs deal with textual and multimedia news content produced by news agencies, news organisations and external sources that are managed and distributed to customers and audience [12,14,15]. As textual data we consider the raw text from any source like news articles, social media feeds, web pages, blogs, PDF files, biographies, reports, historical data and geopolitical data. Whereas, as multimedia we consider live broadcasts, photographs, audio files and video files. Moreover, news agencies produce and distribute content in different formats like plain text, Information Interchange Model (IIM), News Industry Text Format (NITF), NewsML and RDF [16,35].

News content is annotated and enriched with metadata using LOD, semantic vocabularies and ontologies, for example, the ASRAEL project "leverage[s] the Wikidata knowledge base to produce semantic annotations of news articles" [23] (p. 1). Metadata can describe different types of basic information like the authorship, language, creation time, ownership, media type, priority, status, version, keywords and categories; as well as inferred information like provenance, tone and sentiment, and the relevant persons, stories, locations, organisations and events [14,34,37].

Journalists and customers of newsrooms are highly interested in current events and their related information [12]. In addition, JKPs are designed to support additional information needs: General users want to have access to details about the stories (i.e., who, what, why, where and when), identify networks of actors and implications, search the events based on their type or place, obtain facts, and retrieve evidences [15,16,24]. News professionals need access to news archives and knowledge bases for documentation purposes, finding connections from past events, following histories and identifying emerging topics [14,35,36,42]. Additionally, customers have different information needs depending on their business or interests, for example, "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" [35] (p. 1).

4.3. Functionalities

JKPs provide different functionalities to their users. Table 3 lists the identified main functionalities.

Table 3. Most common type of functionalities and services provided in JKPs.

Functionality	**Explanation**
News creation	The process to create a news story.
Verification	The process of checking the facts and claims.
Source selection	The ability to select the information sources of interest.
Monitoring	The ability to continuously distil information from source.
Knowledge discovery	Functionalities for exploring relevant information.
Trends	The current newsworthy developments.
Alert	A notification.
Summarisation	Extracting and representing the key information from a larger text or group of text.
Clustering	Grouping similar stories or events.
Business support	Functionalities to support management workflows.
Content management	Functionalities oriented to store, organise and distribute information.
Personalisation	Providing information according to the user's interests.

News professionals use JKPs for news creation. This creative process involves different tasks such as discovering, collecting, organising, contextualising and publishing [56,57]. JKPs guide news professionals in writing up their stories [29], support them with contextual background knowledge [12,13,29], provide the means for comparing current events with other events [23] and facilitate access to previous work for creating similar content for a different audience, region or language [42]. JKPs also support news professionals with verification [58] tasks like fact-checking [19,59], provenance [15], rights and authorship management [35]. These are typically time-consuming tasks for journalists and fact-checkers that JKPs automate [17].

Source selection and monitoring functionalities are common across the studied JKPs that harvest and store content from internal and external sources and monitor them in real-time [19,21,36,42]. These functionalities allow journalists to automatically follow and distil news and social media of interest and relieve them from these time-consuming tasks.

Knowledge discovery [60] is one of the most attractive functionalities of JKPs. It allows users to obtain news insights, analysis and relevant information. For instance, in NewsReader 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" [15] (p. 1). Other interesting functionalities among the studied JKPs are the trends identification used to discover emerging topics, long-term developments and changes in events over time [21,37]; alerts to keep users up-to-date with the last incoming items [19,31,41]; summarisation [61] of news histories and events to provide additional insights [21]; clustering of story lines and events [23,42].

JKPs can be used as business support systems to manage and monitor internal newsrooms production, news coverage and broadcast decisions [31,42]. This helps managers and editors in allocating resources, avoiding duplicate work and detecting news that can be relevant to different audiences. JKPs are also used for content management that allows newsrooms to store, organise and distribute the daily produced content and metadata [14,16,35].

Most of these functionalities should be personalised and tailored to the stakeholders' needs. Hence, JKPs allow the personalisation of their functionalities according to users' preferences and profiles [12,18,33].

4.4. Techniques

JKPs implement and combine different IT techniques to fulfil their functionalities. Table 4 lists the IT techniques that we have identified.

Table 4. The most common IT techniques used in JKPs.

Technique	Explanation
Semantic technologies	Set of technologies designed to work with LOD and semantic data [46].
Fact extraction	The techniques used to identify factual claims.
Conceptual model	A representations of the world or a part of.
Reasoning	The techniques used to infer knowledge.
Network analysis	The techniques used to analyse networks of things.
Event analysis	The techniques used to analyse events.
Natural Language Processing (NLP)	A set of techniques intended to work and process language.
AI training	The process of creating and tuning an AI model to perform on a given dataset or scenario.

Semantic technologies [46] and similar semantic representation techniques are widely utilised in all the studied JKPs. They use semantic technologies for automating annotation, disambiguating, enriching and leveraging news items with information from external knowledge bases [12,14,19,37]. The semantic representations provide neutral language, explicit relations and facilitate structural matching and lingual independence. They are used for clustering news items and events [23] and detecting trends and story lines [15]. These semantic representations together with fact extraction techniques are used to obtain factual claims from news items and link them to their sources and facts in external knowledge bases (e.g., Wikidata, Wikipedia) [15,19,42].

Conceptual models provide vocabularies, schemas and ontologies. These are often implemented using semantic technologies and represent news stories, events and related information. In addition, conceptual models can define users' interests and preferences [18,20,35], and provide shared resources and formats to facilitate content management and semantic interoperability [14,16,24,37].

Conceptual models and semantic technologies are also used for reasoning, network analysis and event analysis. Reasoning techniques abstract and infer new knowledge from news items, events and temporal aspects [37]. Network analysis is used to find networks of actors, organisations and their implications [15]. Event analysis is applied to detect, identify, cluster and annotate the events described in the news [21,23,35].

The aforementioned techniques are supported by NLP tasks such as named entity recognition, relation extraction and temporal expression normalisation [19–21,37,40]. These NLP tasks, among others, are used in many of the components and functionalities of JKPs. In order to obtain optimal results from the NLP tasks, near-continuous training on extensive news corpora [23] is needed to always keep the machine learning models up-to-date.

4.5. Components

JKPs rely on different components to fulfil their functionalities and support users. We split these components into four groups: processing, storage, interaction and distribution (see Figure 2). The processing components deal with harvesting data from different sources and processing them. The storage components store and manage data. The interaction components allow users to interact with the information from the system and the distribution components distribute information to users.

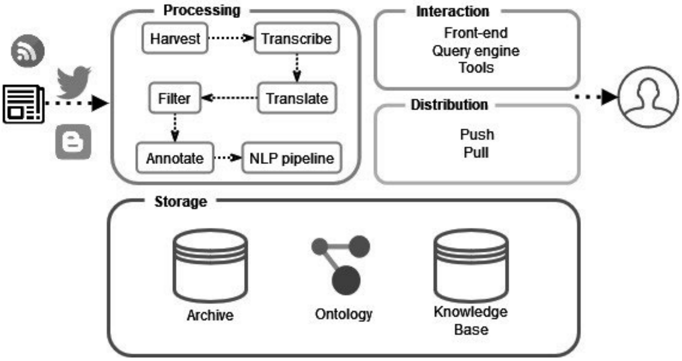


Figure 2. JKP components.

The processing components cover tasks from data gathering to transforming input sources into knowledge representations. The textual and multimedia sources are continuously harvested. However, not all contents receive the same interest from news professionals, like in SUMMA where "entertainment programming such as movies and sitcoms, commercial breaks, and repetitions of content (e.g., on 24/7 news channels) [. . .] [are] of limited interest to monitoring operations" [42] (p. 1). Thus, the harvested content is also translated [42] and filtered according with the different stakeholders' interests and needs. In the studied JKPs, spoken content is transcribed [42] and images are textually described [12] to be able further process them.

The harvested content is automatically annotated with metadata (e.g., authorship, categories and topics) to support functionalities like business support, content management and personalisation [14,31,33,35]. The annotated content is often processed by a NLP pipeline using state-of-the-art NLP and natural language understanding modules to perform linguistic tasks such as co-reference resolution, named entity recognition, relation extraction and sentiment analysis [15,19,62]. Both the results of the NLP pipeline and the annotated content are represented semantically following a predefined schema or ontology. These representations link the annotations to a knowledge base (e.g., an RDF-based knowledge graph) [20,37] and enrich the news items with facts from external knowledge bases (e.g., the LOD cloud, DBpedia and Wikidata) [15,23].

The storage infrastructure of a JKP can be composed of an archive, an ontology and a knowledge base. The archive can store millions of historical news articles, biographies, reports [14,37] and other relevant textual and multimedia items. The knowledge base is where the annotated semantic representations of news items are stored and enriched with external information [14,15,24]. The ontology is used to represent the structure of the news items, leveraged information, metadata and vocabulary [14,24,31,35]. Most recent JKPs also include dedicated storage for real-time news-related feeds [42].

Stakeholders interact with the previous components and have access to the functionalities of JKPs mainly by using three types of interaction components: front-ends that implement specific functionalities, for example, news editors with automatic annotation for creating news articles, statistical and visual analysis features for generating reports [19,21] and enhanced insights for discovering new stories [18]; tools that provide useful resources for creating news like currency converters and dictionaries [35]; query engines that can be accessed through APIs and user interfaces. These allow journalists and customers to query, explore, analyse and visualise the archives and knowledge bases [16,19,20,31,42].

News agencies and news organisations use the push and pull components for delivering and distributing content to their users. Push components offer interfaces where information consumers can select and subscribe to feeds of news [12,16,19,31,41]. Whereas the pull components are used to access and browse the repositories of JKPs [14,16,21,31,35].

4.6. Concerns

Stakeholders, information, functionalities, techniques and components are influenced or affected by additional concerns of various types. Table 5 lists the identified concerns.

Table 5. Concerns related to JKPs.

Aspect	**Explanation**
Customers heterogeneity	The diversity of newsroom customers.
Standards	Standards like IPTC topics or RDF.
Ownership	Copyrights, authorship and licensing information.
Multilingual content	Content produced in various languages.
Timeliness	The temporal aspect of news, when they are published and when the stories happen.
Human factors	Human-related aspects that affect newsroom and JKPs.
Quality	The information and data quality.
Big data	Aspects related to the large volume of data, variety of data and velocity in which data is produced.
Performance	The ability to provide results with the expected quality and on time.
Legacy	Old systems or repositories.
Software architecture	The structure and components of a software system [63].
Maintenance	The ability to reuse, fix and update existing systems.

The customers of JKPs are heterogeneous. They cover diverse sectors and industries, from other newsrooms to companies and institutions, and use different systems to interact with JKPs [35,42]. To improve the interoperability between news agencies and stakeholders, JKPs utilise standards like the IPTC news codes, media topics, semantic vocabularies and RDF [14,35], and keep track of information related to ownership, such as authorship, copyrights, privacy and sources [12,64]. JKPs can also use the ownership information to control the information provenance and reliability [15] by, for example, tracking back the information to its original source and identifying trustworthy providers.

Customers and audiences prefer different languages [21,23,35,37,42]. Hence, JKPs deal with and produce multilingual news items (e.g., Norwegian, Italian, Spanish, English) that are translated, transcribed and delivered in the preferred languages. In addition, these news items have an intrinsic timeliness aspect that defines their value either as a fresh event or as part of a past or present storyline or historic development that can be reconstructed [12,15,20,42].

JKPs attempt to address different human factors in newsrooms. JKPs automate error-prone and time-consuming processes that were performed manually like news tagging, source monitoring, information filtering, verification, fact-checking and finding related articles and relevant information [14,17,19,21,35]. Hence, JKPs free journalists from these tedious tasks and improve their results. As a result, JKPs facilitate high-quality information to meet the standards of their stakeholders [12].

On the technical side, JKPs deal with big data requirements like volume, velocity, variety. ASRAEL estimates that "the number of collected articles ranges between 100,000 and 200,000 articles per day [. . .] from around 75,000 news sources" [21] (p. 1). NewsReader uses an archive that "contains billions of articles, biographies, and reports" [37] (p. 1). SUMMA platform "[is] able to ingest 400 TV streams simultaneously" [42] (p. 6). Hence, the components of JKPs are designed considering their performance to minimise the processing and distribution times [12,15]. JKPs also integrate legacy components and facilitate interoperability with other systems and external services [16,24,35,37,41]. All these factors make the software architecture of JKPs complex and difficult to maintain without guidance.

5. Future Directions for Research on JKPs

5.1. Implications for Research

5.1.1. Stakeholders

Studies on understanding how journalists embrace digital tools can aid in better adapting JKPs to the way journalists work. Such studies should consider the journalists' perceptions on using intelligent systems for creating news, how journalists process and use background information and the journalists' experiences working with AI, etc. Along these lines, related studies have been proposed, but not limited to, the journalists' usage of social media for gathering and verifying information [65,66] and the relation of the journalism practices and AI [67,68]. Similar user-oriented studies should be conducted on readers and younger and future generations of news consumers to identify what new forms of interaction and consumption are more appealing to them. These studies could consider, for example, the readers' perceptions of automated journalism [69,70] and young people's engagement with news recommendations [71].

5.1.2. Information

To date, the knowledge extraction and recognition of entities from images and videos remain limited. Due to that, JKPs are not able to capture enough information from multimedia news. Promising directions for extracting knowledge from multimedia sources are multimodal machine learning approaches [72] that combine different types of data such as visual and text representations [73,74] and spoken language understanding tasks that analyse and detect audio speech [75]. Another limitation for knowledge extraction is the dark entities (i.e., those entities that do not exist yet in the knowledge base) [76,77]. Fresh stories about newer facts are the most attractive news, therefore, the chances of finding entity representations for those newer facts in knowledge bases are low. Therefore, research on knowledge extraction from multimedia news and dark entities can improve news representation in JKPs.

5.1.3. Functionalities

Non-technical users find it difficult to perform complex searches in knowledge bases, archives and background information due to their lack of expertise. The usage of chatbots can aid user interaction using natural language [42,78]. Additional solutions that can support journalists' interaction with knowledge and information, and automate news production are text summarisation [61], automated reporting or story generation [79,80] and automatic data visualisation [81]. Augmented reality may also bring new possibilities for assisting the exploration of information using knowledge representations and LOD [82].

5.1.4. Techniques

Due to the increase in misinformation and propaganda, it is crucial for journalists and readers to detect and distinguish trustworthy information from fake and biased news. Hence, research on JKPs should include automating the detection of fake news, political bias and rumours across social media platforms and news sources [58,83]. Techniques for such purposes can benefit from research on automating fact-checking [17,59], detecting derived or copied works [21], and media and audio forensics to identify manipulated or tempered multimedia files [84,85]. In addition, identifying misinformation items before they are stored in the knowledge base can improve the data quality of JKPs. Another promising direction is the inclusion of neural-symbolic AI [86] techniques as part of the different components of JKPs. Neural-symbolic AI combines neural networks with reasoning and logic. This can facilitate the inference and deductive reasoning over the data in the JKPs and reduce the computational cost of reasoning over knowledge graphs [87].

5.1.5. Components

In addition to automatic techniques for verification and fact-checking, promising collaborative tools for news and social media verification that involve journalists and readers [88] should be considered, for example, the tools developed in the ReVeaL (https://revealproject.eu accessed on 15 May 2022), InVID (https://www.invid-project.eu accessed on 15 May 2022) [89] and WeVerify (https://weverify.eu accessed on 15 May 2022) [90] projects. Some of these tools such as WeVerify employ blockchain and knowledge graphs services for recording debunked claims and news. These collaborative repositories could be considered as additional information sources from which JKPs can obtain checked claims and provenance information but also contribute with verified information. Apart from this, the current JKPs are focused on in-house platforms that are typically accessed through a computer and oriented to print journalism. However, there is limited research on components that can facilitate access to the services offered by JKPs for mobile journalism [91] (i.e., journalism edited and published through smartphones and oriented towards audio-visual storytelling).

5.1.6. Concerns

There are no gold standards or methodologies to evaluate JKPs. Accordingly, research needs to include the design and study of evaluation methods for JKPs. Moreover, readers and journalists may perceive results from JKPs as less transparent and difficult to understand [92] as they are driven by AI. To improve their perception of trustworthiness and transparency, research on JKPs should consider explainable AI methods [93].

5.2. Implications for Practice

5.2.1. Stakeholders

To date, there have not been any studies on the implementation of JKPs in newsrooms. Such studies should evaluate the effectiveness, adoption and demand of JKPs. The experiences in implementing JKPs can help to draw a digitalisation path for newsrooms by providing best practices and identifying the main obstacles and solutions. This can support newsrooms with the definition of their roadmaps towards the adoption of JKPs, as it facilitates the identification of the most relevant aspects of JKPs and particular needs according to their current stage. Related studies have considered and provided guidelines for the utilisation of AI in news creation processes in a broader sense [55].

5.2.2. Information

The literature is unclear on how JKPs should best represent events and there is no general agreement on what constitutes an event [21]. Events can range from fine-grained actions like a shot, injury or a handshake between two actors [15] to bigger and broader events like the Spanish Civil War and the COVID-19 pandemic [23] or events in between like a trial process. Therefore, research on JKPs needs to define and discuss how different types of events at different granularity can co-exist in a JKP and what conceptualisations of the event are useful for specific use cases.

5.2.3. Functionalities

A better understanding of how to represent events and news items can bring new possibilities for JKPs, for example, on data analysis like measuring the popularity of people and companies [15], finding cause and effect relations [21], and identifying newsworthy events for specific audiences and particular user′ interests [18,33,94].

5.2.4. Techniques

One of the main limitations of the studied JKPs is the extraction of enough and precise information from text and multimedia to represent news stories in high detail [19,31]. For the knowledge graph-based JKPs we have considered in this paper, this means representing the content of text and multimedia as knowledge graphs. JKPs use relation extraction models to extract the textual relations between the entities in news text [15,62]. However, these models are in an early research stage and the extracted relations are basic and limited for representing news [95]. Therefore, the functionalities that are based on these models must be considered for the longer term.

5.2.5. Components

Current open-source large triple-stores are not scalable and their reasoning services are time-consuming and use too many computing resources. This limits the possibilities for JKPs to exploit reasoning capabilities and analyse large knowledge graphs. Hence, scalable triple-stores and mechanisms for better reasoning over large knowledge graphs can ease the incorporation of such solutions and bring new possibilities for JKPs. A promising approach is the inclusion of entity spaces [96]. These are vector spaces that represent the different entities of a knowledge graph and also capture their semantic information. They can be used to speed up processes that require complex graph explorations like inferring and disambiguating knowledge for unseen entities. Another promising approach for integrating and managing information from different types of databases is the usage of virtual knowledge graph [97]. Virtual knowledge graphs represent the schema of the different databases and provide mechanisms for querying the databases using SPARQL, hence, it integrates databases on the schema level and reduces data replication.

5.2.6. Concerns

Only the most recent projects proposed systems to deal with big data [37,39,42]. Their architectures must also keep the machine learning models up-to-date and replace them for future best-of-breed, facilitate the schema evolution of knowledge bases and ease the expansion, distribution and independence of services [44]. Research on software reference architectures [98] for JKPs can assist in better designing and implementing them, as well as establishing a vocabulary and a framework to compare JKPs.

6. Limitations

This study only covers the English-language literature and is based on JKPs developed in Europe, Canada and USA. We have not identified any relevant JKPs in other geographical regions, but of course such JKPs may have been reported in languages other than English. The study is also influenced by the authors' involvement in the development of News Hunter. To reduce bias, we have not included our JKP during the meta-analysis process and we limited the News Hunter contribution to supporting and extending the findings. Additionally, the purpose of our analysis is to review the current state and future directions of the field, and not to evaluate the quality of the proposals.

7. Conclusions

This study has addressed which challenges and opportunities have motivated knowledge graph-based JKPs (RQ1), how knowledge graph-based JKPs are addressing these challenges and opportunities (RQ2), and the future directions of research on knowledge graph-based JKPs (RQ3). To our knowledge, no previous studies have identified and analysed JKPs as an emerging type of intelligent information system in this way. Although there are examples of such systems in the literature, to date, ours is the first clear definition and broad analysis of JKPs and their context.

In current newsroom workflows, metadata annotation is a manual, time-consuming and error-prone process. Newsrooms face difficulties to implement high-quality information systems. Journalists spend a lot of their time monitoring and filtering vast volumes of news, time that otherwise could be invested in creative tasks. These vast volumes of data often lack fine-grained annotation and are split into different repositories with different schemas. This limits the capacity of newsrooms to analyse and exploit their information resources, and share data with news consumers. To help with these processes, newsrooms use a large variety of services that are challenging to integrate and operate together, hindering their evolution towards digitally-integrated newsrooms. JKPs are a new type of intelligent information system that offer many opportunities for high-quality journalism in newsrooms by combining AI, knowledge bases, LOD, NLP, ML and deep learning techniques. JKPs automate the metadata annotation and content enrichment with background information from external sources; monitor internal and word-wide news media output; facilitate event detection; support news creation and verification. They also facilitate the ingestion of vast amounts of data, and its storage, organisation and distribution. JKPs can provide newsrooms with a digitalisation path to reduce production costs and improve information quality while adapting the current workflows of newsrooms to new forms of journalism and readers' demands. We expect the next generation of JKPs to focus on enhancing journalism and providing unexpected news insights for journalists.

Many JKPs are big-data-oriented systems [15,21,22,35,42,44] that need a significant investment effort from newsrooms, making the adoption of JKPs challenging for small or local newsrooms. The adoption of JKP can yield many benefits, but newsrooms may perceive JKPs as an investment risk and look for alternative services. Thus, the formalisation of JKPs and the usage of open-source and out-of-the-box solutions, together with the popularisation of knowledge graphs will lower the adoption risk and increase the benefits. For small and local newsrooms, sharing JKPs can reduce the entrance barriers—a practice that is becoming more popular among digital-born news organisations and freelancers.

Section 5 has already proposed several paths for further research on JKPs. As an immediate continuation of this study, we are designing the software reference architecture for JKPs and developing tools to further study and enhance JKPs [44]. Through this work, we plan to define a reference model for JKPs that will allow their comparison and validation.

Author Contributions: M.G.O.: conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, visualization, writing—original draft. A.L.O.: conceptualization, resources, data curation, writing—review and editing, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Norwegian Research Council IKTPLUSS project 275872.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The study is influenced by the authors' involvement in the development of the News Hunter platform. To reduce bias, we have not included our JKP during the meta-analysis process and we limited the News Hunter contribution to supporting and extending the findings. Additionally, the purpose of our analysis is to review the current state and future directions of the field, and not to evaluate the quality of the proposals. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

JKP	Journalistic Knowledge Platform
AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
LOD	Linked Open Data
RDF	Resource Description Framework

Appendix A. Analysis Method

To synthesise data from the literature on the platforms that fit under our definition of JKP, we have used a qualitative meta-analysis approach [99,100]. We have searched the research literature and identified 28 papers describing 14 JKPs carried by distinct research groups located in different countries and in collaboration with a variety of news organisations and industry partners (Table 1 presents an overview of the selected JKPs and papers). According to Maxwell [101], our sample represents an adequate variation in the phenomenon of interest. During the meta-analysis process, we focused on the last 10 years of advances and excluded our JKP from the process of extracting and coding data. After we synthesised the first conclusions, the excluded JKPs were added and analysed to support and expand our findings. This decision was taken to focus on the most recent advances and minimise the bias in the meta-analysis process inducted from our point of view.

From the selected literature we manually extracted 322 claims about the JKPs, i.e., statements that described the current state or expressed potential challenges or opportunities. Two independent expert coders (viz., the authors) conducted a purposive sampling [102,103] using the extracted claims that became marked up with 406 codes. We cleaned the generated codes with the support of NLP and natural language understanding techniques (implemented in python with support of Scikit-learn [104], NLTK [105], SpaCy [106] and other libraries) (i.e., Damerau-Levenshtein distance [107], word2vec [108] and Wordnet [109]). After cleaning and tidying up the initial codes, we interatively classified the resulting codes into six top-level categories and 64 sub-categories (Figures 1 and 2 and Tables 2–5 shown the final top-level and sub-categories).

We used the top-level and sub-categories to re-code the 322 claims and we measured the final agreement using Gwet's AC_1 [110] inter-rater reliability coefficient with nominal ratings. The AC_1 coefficient for each category was 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; with an average AC_1 of 0.69 and a standard deviation of 0.063. Following the recommendations of Gwet [110] about Landis–Koch and Altman's benchmark scales, our AC_1 express an acceptable agreement among coders. Finally, we agreed on the final codes for each claim and initiated an abductive process to understand and derive the current state and future directions of the research of JKPs.

References

1. Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; Robertson, C.T.; Nielsen, R.K. *Reuters Institute Digital News Report 2021*; Technical Report; Reuters Institute for the Study of Journalism: Oxford, UK, 2021.
2. Simon, F.M.; Graves, L. *Pay Models for Online News in the US and Europe: 2019 Update*; Technical Report; Reuters Institute for the Study of Journalism: Oxford, UK, 2019.
3. Fletcher, R.; Nielsen, R.K. Paying for Online News. *Digit. J.* **2017**, *5*, 1173–1191. [CrossRef]
4. Newman, N.; Fletcher, R.; Kalogeropoulos, A.; Nielsen, R.K. *Reuters Institute Digital News Report 2019*; Technical Report; Reuters Institute for the Study of Journalism: Oxford, UK, 2019.
5. Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; Nielsen, R.K. *Reuters Institute Digital News Report 2020*; Technical Report; Reuters Institute for the Study of Journalism: Oxford, UK, 2020.
6. Toff, B.J.; Badrinathan, S.; Mont'Alverne, C.; Arguedas, A.R.; Fletcher, R.; Nielsen, R.K. *Overcoming Indifference: What Attitudes Towards News Tell Us about Building Trust*; Technical Report; Reuters Institute for the Study of Journalism: Oxford, UK, 2021.
7. Newman, N.; Fletcher, R. *Bias, Bullshit and Lies: Audience Perspectives on Low Trust in the Media*; Technical Report; SSRN: Rochester, NY, USA, 2017. [CrossRef]
8. Vázquez Herrero, J.; Direito-Rebollal, S.; Rodríguez, A.S.; García, X. *Journalistic Metamorphosis: Media Transformation in the Digital Age*; Springer: Cham, Swizerland, 2020. [CrossRef]
9. Beckett, C. *New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence*; Technical Report; Polis, London School of Economics and Political Science: London, UK, 2019.
10. Lewis, S.C.; Westlund, O. Big Data and Journalism. *Digit. J.* **2015**, *3*, 447–466. [CrossRef]
11. Keefe, J.; Zhou, Y.; Merrill, J.B. The Present and Potential of AI in Journalism. Knight Foundation. 2021. Available online: https://knightfoundation.org/articles/the-present-and-potential-of-ai-in-journalism/ (accessed on 22 February 2022).
12. Fernández, N.; Blázquez, J.M.; Fisteus, J.A.; Sánchez, L.; Sintek, M.; Bernardi, A.; Fuentes, M.; Marrara, A.; Ben-Asher, Z. NEWS: Bringing Semantic Web Technologies into News Agencies. In Proceedings of the Semantic Web—ISWC 2006, Athens, GA, USA, 5–9 November 2006; pp. 778–791. [CrossRef]
13. Maiden, N.; Zachos, K.; Brown, A.; Brock, G.; Nyre, L.; Nygård Tonheim, A.; Apsotolou, D.; Evans, J. Making the News: Digital Creativity Support for Journalists. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–11. [CrossRef]
14. Castells, P.; Perdrix, F.; Pulido, E.; Rico, M.; Benjamins, R.; Contreras, J.; Lorés, J. Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In Proceedings of the Semantic Web: Research and Applications, ESWS 2004, Heraklion, Crete, Greece, 10–12 May 2004. [CrossRef]

15. Rospocher, M.; van Erp, M.; Vossen, P.; Fokkens, A.; Aldabe, I.; Rigau, G.; Soroa, A.; Ploeger, T.; Bogaard, T. Building Event-Centric Knowledge Graphs from News. *J. Web Semant.* **2016**, *37–38*, 132–151. [CrossRef]
16. Raimond, Y.; Scott, T.; Oliver, S.; Sinclair, P.; Smethurst, M. Use of Semantic Web technologies on the BBC Web Sites. In *Linking Enterprise Data*; Springer: New York, NY, USA, 2010. [CrossRef]
17. Miranda, S.A.; Nogueira, D.; Mendes, A.; Vlachos, A.; Secker, A.; Garrett, R.; Mitchel, J.; Marinho, Z. Automated Fact Checking in the News Room. In Proceedings of the World Wide Web Conference, WWW '19, San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA; pp. 3579–3583. [CrossRef]
18. Kalfoglou, Y.; Domingue, J.; Motta, E.; Vargas-Vera, M.; Buckingham Shum, S. myPlanet: An ontology driven Web based personalised news service. In Proceedings of the International Joint Conference on Artificial Intelligence, Washington, DC, USA, 4–10 August 2001; Volume 2001, pp. 44–52.
19. Java, A.; Finin, T.; Nirenburg, S. SemNews: A Semantic News Framework. In Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, Boston, MA, USA, 16–20 July 2006.
20. Borsje, J.; Levering, L.; Frasincar, F. Hermes: A Semantic Web-Based News Decision Support System. In Proceedings of the 2008 ACM Symposium on Applied Computing, SAC '08, Fortaleza, Brazil, 16–20 March 2008; Association for Computing Machinery: New York, NY, USA, 2008; pp. 2415–2420. [CrossRef]
21. Leban, G.; Fortuna, B.; Brank, J.; Grobelnik, M. Event Registry: Learning about World Events from News. In Proceedings of the 23rd International Conference on World Wide Web, WWW'14 Companion, Seoul, Korea, 7–11 April 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 107–110. [CrossRef]
22. Liu, X.; Nourbakhsh, A.; Li, Q.; Shah, S.; Martin, R.; Duprey, J. Reuters tracer: Toward automated news production using large scale social media data. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 1483–1493.
23. Rudnik, C.; Ehrhart, T.; Ferret, O.; Teyssou, D.; Troncy, R.; Tannier, X. Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In Proceedings of the Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1232–1239. [CrossRef]
24. Ramagem, D.B.; Margerin, B.; Kendall, J. AnnoTerra: Building an integrated earth science resource using semantic Web technologies. *IEEE Intell. Syst.* **2004**, *19*, 48–57. [CrossRef]
25. Al-Moslmi, T.; Gallofré Ocaña, M.; Opdahl, A.L.; Tessem, B. Detecting Newsworthy Events in a Journalistic Platform. In Proceedings of the 3rd European Data and Computational Journalism Conference, Malaga, Spain, 1–2 July 2019; pp. 3–5.
26. Hogan, A.; Blomqvist, E.; Cochez, M.; D'amato, C.; Melo, G.D.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S.; et al. Knowledge Graphs. *ACM Comput. Surv.* **2021**, *54*, 1–257. [CrossRef]
27. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*; IGI Global: Hershey, PA, USA, 2011; pp. 205–227.
28. Opdahl, A.L.; Al-Moslmi, T.; Dang-Nguyen, D.T.; Gallofré Ocaña, M.; Tessem, B.; Veres, C. Semantic Knowledge Graphs for the News: A Review. *Comput. Surv.* 2022, *to appear*.
29. Berven, A.; Christensen, O.A.; Moldeklev, S.; Opdahl, A.L.; Villanger, K.J. A knowledge-graph platform for newsrooms. *Comput. Ind.* **2020**, *123*, 103321. [CrossRef]
30. Gallofré Ocaña, M.; Opdahl, A.L. Challenges and Opportunities for Journalistic Knowledge Platforms. In Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 19–23 October 2020.
31. Domingue, J.; Motta, E. PlanetOnto: From news publishing to integrated knowledge management support. *IEEE Intell. Syst. Their Appl.* **2000**, *15*, 26–32. [CrossRef]
32. Frasincar, F.; Borsje, J.; Levering, L. A semantic web-based approach for building personalized news services. *Int. J. E-Bus. Res. (IJEBR)* **2009**, *5*, 35–53. [CrossRef]
33. Schouten, K.; Ruijgrok, P.; Borsje, J.; Frasincar, F.; Levering, L.; Hogenboom, F. A semantic web-based approach for personalizing news. In Proceedings of the 2010 ACM Symposium on Applied Computing—SAC '10, Sierre, Switzerland, 22–26 March 2010; ACM Press: Sierre, Switzerland, 2010; p. 854. [CrossRef]
34. Kobilarov, G.; Scott, T.; Raimond, Y.; Oliver, S.; Sizemore, C.; Smethurst, M.; Bizer, C.; Lee, R. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *The Semantic Web: Research and Applications*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5554. [CrossRef]
35. Fernández, N.; Fuentes, D.; Sánchez, L.; Fisteus, J.A. The NEWS ontology: Design and applications. *Expert Syst. Appl.* **2010**, *37*, 8694–8704. [CrossRef]
36. Kattenberg, M.; Beloki, Z.; Soroa, A.; Artola, X.; Fokkens, A.; Huygen, P.; Verstoep, K. Two architectures for parallel processing for huge amounts of text. In Proceedings of the Language Resources and Evaluation Conference (LREC). European Language Resources Association (ELRA), Portorož, Slovenia, 23–28 May 2016; pp. 4513–4519.
37. Vossen, P.; Agerri, R.; Aldabe, I.; Cybulska, A.; van Erp, M.; Fokkens, A.; Laparra, E.; Minard, A.L.; Aprosio, A.P.; Rigau, G.; et al. NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Spec. Issue Knowl.-Based Syst. Elsevier* **2016**, *110*, 60–85. [CrossRef]

38. Li, Q.; Shah, S.; Liu, X.; Nourbakhsh, A.; Fang, R. TweetSift: Tweet Topic Classification Based on Entity Knowledge Base and Topic Enhanced Word Embedding. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Indianapolis, IN, USA, 24–28 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2429–2432. [CrossRef]
39. Liu, X.; Li, Q.; Nourbakhsh, A.; Fang, R.; Thomas, M.; Anderson, K.; Kociuba, R.; Vedder, M.; Pomerville, S.; Wudali, R.; et al. Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Indianapolis, IN, USA, 24–28 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 207–216. [CrossRef]
40. Paikens, P.; Barzdins, G.; Mendes, A.; Ferreira, D.C.; Broscheit, S.; Almeida, M.S.; Miranda, S.; Nogueira, D.; Balage, P.; Martins, A.F. SUMMA at TAC Knowledge Base Population Task 2016. In Proceedings of the Ninth Text Analysis Conference (TAC), Gaithersburg, MA, USA, 14–15 November 2016.
41. Germann, U.; Liepins, R.; Gosko, D.; Barzdins, G. Integrating Multiple NLP Technologies into an Open-source Platform for Multilingual Media Monitoring. In Proceedings of the Workshop for NLP Open Source Software (NLP-OSS): Melbourne, Australia, 19–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 47–51. [CrossRef]
42. Germann, U.; Liepins, R.; Barzdins, G.; Gosko, D.; Miranda, S.; Nogueira, D. The SUMMA Platform: A Scalable Infrastructure for Multi-lingual Multi-media Monitoring. In Proceedings of the ACL 2018, System Demonstrations: Melbourne, Australia, 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 15–20 July 2018; pp. 99–104. [CrossRef]
43. Gallofré Ocaña, M.; Nyre, L.; Opdahl, A.L.; Tessem, B.; Trattner, C.; Veres, C. Towards a Big Data Platform for News Angles. In Proceedings of the 4th Norwegian Big Data Symposium (NOBIDS) 2018, Trondheim, Norway, 14 November 2018; pp. 17–29.
44. Gallofré Ocaña, M.; Opdahl, A.L. Developing a Software Reference Architecture forJournalistic Knowledge Platforms. In Proceedings of the ECSA2021 Companion Volume, Växjö, Sweden, 13–17 September 2021.
45. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
46. Shadbolt, N.; Berners-Lee, T.; Hall, W. The Semantic Web Revisited. *IEEE Intell. Syst.* **2006**, *21*, 96–101. [CrossRef]
47. International Press Telecommunications Council. IPTC: Media Topics. 2022. Available online: https://iptc.org/standards/media-topics/ (accessed on 22 February 2022).
48. Cyganiak, R.; Wood, D.; Lanthaler, M. RDF 1.1 Concepts and Abstract Syntax. 2014. Available online: http://www.w3.org/TR/rdf11-concepts/ (accessed on 22 February 2022).
49. Troncy, R. Bringing the IPTC News Architecture into the Semantic Web. In Proceedings of the Semantic Web—ISWC 2008, Karlsruhe, Germany, 26–30 October 2008; Sheth, A., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 483–498.
50. International Press Telecommunications Council. IPTC: NewsCodes. 2022. Available online: https://iptc.org/standards/newscodes/ (accessed on 22 February 2022).
51. Miles, A.; Bechhofer, S. SKOS Simple Knowledge Organization System Namespace Document—HTML Variant. 2009. Available online: http://www.w3.org/2004/02/skos/core.html (accessed on 22 February 2022).
52. W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview (Second Edition). 2012. Available online: https://www.w3.org/TR/owl-overview/ (accessed on 22 February 2022).
53. International Press Telecommunications Council. IPTC: News Architecture. 2022. Available online: https://iptc.org/standards/news-architecture/ (accessed on 22 February 2022).
54. Opdahl, A.L.; Tessem, B. Ontologies for finding journalistic angles. *Softw. Syst. Model.* **2020**, *20*, 71–87. [CrossRef]
55. Lopez, M.G.; Porlezza, C.; Cooper, G.; Makri, S.; MacFarlane, A.; Missaoui, S. A Question of Design: Strategies for Embedding AI-Driven Tools into Journalistic Work Routines. *Digit. J.* **2022**, *10*, 1–20. [CrossRef]
56. Gutierrez Lopez, M.; Makri, S.; MacFarlane, A.; Porlezza, C.; Cooper, G.; Missaoui, S. Making newsworthy news: The integral role of creativity and verification in the human information behavior that drives news story creation. *J. Assoc. Inf. Sci. Technol.* 2022, *online version of record*.
57. Deuze, M. On creativity. *Journalism* **2019**, *20*, 130–134. [CrossRef]
58. Meel, P.; Vishwakarma, D.K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.* **2020**, *153*, 112986. [CrossRef]
59. Guo, Z.; Schlichtkrull, M.; Vlachos, A. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 178–206.
60. Diakopoulos, N. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digit. J.* **2020**, *8*, 945–967. [CrossRef]
61. El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679. [CrossRef]
62. Al-Moslmi, T.; Gallofré Ocaña, M. Lifting News into a Journalistic Knowledge Platform. In Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 19–23 October 2020.
63. Garlan, D. Software Architecture. In *Encyclopedia of Software Engineering*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008. [CrossRef]
64. Gallofré Ocaña, M.; Al-Moslmi, T.; Opdahl, A.L. Data Privacy in Journalistic Knowledge Platforms. In Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 19–23 October 2020.

65. Neuberger, C.; Nuernbergk, C.; Langenohl, S. Journalism as Multichannel Communication. *J. Stud.* **2019**, *20*, 1260–1280. [CrossRef]
66. Zhang, X.; Li, W. From Social Media with News: Journalists' Social Media Use for Sourcing and Verification. *J. Pract.* **2020**, *14*, 1193–1210. [CrossRef]
67. Stray, J. Making Artificial Intelligence Work for Investigative Journalism. *Digit. J.* **2019**, *7*, 1076–1097. [CrossRef]
68. Broussard, M.; Diakopoulos, N.; Guzman, A.L.; Abebe, R.; Dupagne, M.; Chuan, C.H. Artificial Intelligence and Journalism. *J. Mass Commun. Q.* **2019**, *96*, 673–695. [CrossRef]
69. Graefe, A.; Bohlken, N. Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media Commun.* **2020**, *8*, 50–59. [CrossRef]
70. Tandoc, E.C., Jr.; Yao, L.J.; Wu, S. Man vs. Machine? The Impact of Algorithm Authorship on News Credibility. *Digit. J.* **2020**, *8*, 548–562. [CrossRef]
71. Swart, J. Experiencing Algorithms: How Young People Understand, Feel About, and Engage with Algorithmic News Selection on Social Media. *Soc. Media Soc.* **2021**, *7*, 20563051211008828. [CrossRef]
72. Guo, W.; Wang, J.; Wang, S. Deep Multimodal Representation Learning: A Survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
73. Mogadala, A.; Kalimuthu, M.; Klakow, D. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Intell. Res.* **2021**, *71*, 1183–1317. [CrossRef]
74. Chen, S.; Aguilar, G.; Neves, L.; Solorio, T. Can images help recognize entities? A study of the role of images for Multimodal NER. In Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Online, 11 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 87–96. [CrossRef]
75. Shon, S.; Pasad, A.; Wu, F.; Brusco, P.; Artzi, Y.; Livescu, K.; Han, K.J. SLUE: New Benchmark Tasks For Spoken Language Understanding Evaluation on Natural Speech. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7927–7931. [CrossRef]
76. van Erp, M.; Ilievski, F.; Rospocher, M.; Vossen, P. Missing Mr. Brown and buying an Abraham Lincoln—Dark entities and DBpedia. In Proceedings of the Third NLP & DBpedia Workshop, Bethlehem, PA, USA, 11 October 2015; pp. 81–86.
77. Al-Moslmi, T.; Gallofré Ocaña, M.; Opdahl, A.L.; Veres, C. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access* **2020**, *8*, 32862–32881. [CrossRef]
78. Luo, B.; Lau, R.Y.; Li, C.; Si, Y.W. A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Min. Knowl. Discov.* **2022**, *12*, e1434. [CrossRef]
79. Miroshnichenko, A. AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes"). *Information* **2018**, *9*, 183. [CrossRef]
80. Alhussain, A.I.; Azmi, A.M. Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
81. Zhu, S.; Sun, G.; Jiang, Q.; Zha, M.; Liang, R. A survey on automatic infographics and visualization recommendations. *Vis. Inform.* **2020**, *4*, 24–40. [CrossRef]
82. Lampropoulos, G.; Keramopoulos, E.; Diamantaras, K. Enhancing the functionality of augmented reality using deep learning, semantic web and knowledge graphs: A review. *Vis. Inform.* **2020**, *4*, 32–42. [CrossRef]
83. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 1–40. [CrossRef]
84. Pasquini, C.; Amerini, I.; Boato, G. Media forensics on social media platforms: A survey. *EURASIP J. Inf. Secur.* **2021**, *2021*, 1–19. [CrossRef]
85. Bhagtani, K.; Yadav, A.K.S.; Bartusiak, E.R.; Xiang, Z.; Shao, R.; Baireddy, S.; Delp, E.J. An Overview of Recent Work in Media Forensics: Methods and Threats. *arXiv* **2022**, arXiv:2204.12067.
86. Hitzler, P.; Bianchi, F.; Ebrahimi, M.; Sarker, M.K. Neural-symbolic integration and the Semantic Web. *Semant. Web* **2020**, *11*, 3–11. [CrossRef]
87. Hitzler, P.; Krotzsch, M.; Rudolph, S. *Foundations of Semantic Web Technologies*; CRC Press: Boca Raton, FL, USA, 2010. [CrossRef]
88. Thomson, T.; Angus, D.; Dootson, P.; Hurcombe, E.; Smith, A. Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities. *J. Pract.* **2020**, *16*, 1–25. [CrossRef]
89. Collyda, C.; Apostolidis, E.; Pournaras, A.; Markatopoulou, F.; Mezaris, V.; Patras, I. VideoAnalysis4ALL: An On-Line Tool for the Automatic Fragmentation and Concept-Based Annotation, and the Interactive Exploration of Videos. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR '17, Bucharest, Romania, 6–9 June 2017; Association for Computing Machinery: New York, NY, USA, 2017. [CrossRef]
90. Marinova, Z.; Spangenberg, J.; Teyssou, D.; Papadopoulos, S.; Sarris, N.; Alaphilippe, A.; Bontcheva, K. Weverify: Wider and Enhanced Verification for You Project Overview and Tools. In Proceedings of the 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–4. [CrossRef]
91. Salzmann, A.; Guribye, F.; Gynnild, A. "We in the Mojo Community"—Exploring a Global Network of Mobile Journalists. *J. Pract.* **2021**, *15*, 620–637. [CrossRef]
92. Shin, D. Why Does Explainability Matter in News Analytic Systems? Proposing Explainable Analytic Journalism. *J. Stud.* **2021**, *22*, 1047–1065. [CrossRef]
93. Kaur, D.; Uslu, S.; Rittichier, K.J.; Durresi, A. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]

94. Motta, E.; Daga, E.; Opdahl, A.L.; Tessem, B. Analysis and Design of Computational News Angles. *IEEE Access* **2020**, *8*, 120613–120626. [CrossRef]
95. Yan, Y.; Sun, H.; Liu, J. A Review and Outlook for Relation Extraction. In Proceedings of the 5th International Conference on Computer Science and Application Engineering, CSAE 2021, Sanya, China, 19–21 October 2021; Association for Computing Machinery: New York, NY, USA, 2021. [CrossRef]
96. van Erp, M.; Groth, P. Towards Entity Spaces. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 2129–2137.
97. Xiao, G.; Ding, L.; Cogrel, B.; Calvanese, D. Virtual Knowledge Graphs: An Overview of Systems and Use Cases. *Data Intell.* **2019**, *1*, 201–223. [CrossRef]
98. Martínez-Fernández, S.; Ayala, C.P.; Franch, X.; Marques, H.M. Benefits and drawbacks of software reference architectures: A case study. *Inf. Softw. Technol.* **2017**, *88*, 37–52. [CrossRef]
99. Eisenhardt, K.M. Building Theories from Case Study Research. *Acad. Manag. Rev.* **1989**, *14*, 532–550. [CrossRef]
100. Hoon, C. Meta-Synthesis of Qualitative Case Studies: An Approach to Theory Building. *Organ. Res. Methods* **2013**, *16*, 522–556. [CrossRef]
101. Maxwell, J.A. *A Realist Approach for Qualitative Research*; Sage: Newcastle upon Tyne, UK, 2012.
102. Corbin, J.; Strauss, A. Grounded theory research: Procedures, canons, and evaluative criteria. *Qual. Sociol.* **1990**, *13*, 2–21. [CrossRef]
103. Corbin, J.; Strauss, A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*; Sage Publications: Newcastle upon Tyne, UK, 1998.
104. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
105. Bird, S.; Loper, E.; Klein, E. *Natural Language Processing with Python*; O'Reilly Media, Inc.: Newton, MA, USA, 2009.
106. Honnibal, M.; Montani, I.; van Landeghem, S.; Boyd, A. *spaCy: Industrial-Strength Natural Language Processing in Python*; Zenodo: Geneva, Switzerland, 2020. [CrossRef]
107. Damerau, F.J. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* **1964**, *7*, 171–176. [CrossRef]
108. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26, NIPS 2013, Barcelona, Spain, 11–19 December 2013; Neural Information Processing Systems Foundation: San Diego, CA, USA.
109. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
110. Gwet, K.L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*; Advanced Analytics, LLC: Oxford, MS, USA, 2014.

Manuscript II

A Software Reference Architecture for Journalistic Knowledge Platforms

Marc Gallofré Ocaña and Andreas L. Opdahl

Knowledge-Based Systems, **276**, 110750. (2023).

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A Software Reference Architecture for Journalistic Knowledge Platforms[☆]

Marc Gallofré Ocaña [*], Andreas L. Opdahl

University of Bergen, Department of Information Science and Media Studies, Bergen, 5020, Norway

ARTICLE INFO

Article history:
Received 18 October 2022
Received in revised form 5 June 2023
Accepted 24 June 2023
Available online 30 June 2023

Keywords:
Software reference architecture
Newsrooms
Knowledge graphs
Journalism
Artificial intelligence
Big data

ABSTRACT

Newsrooms and journalists today rely on many different artificial-intelligence, big-data and knowledge-based systems to support efficient and high-quality journalism. However, making the different systems work together remains a challenge, calling for new unified journalistic knowledge platforms. A software reference architecture for journalistic knowledge platforms could help news organisations by capturing tried-and-tested best practices and providing a generic blueprint for how their IT infrastructure should evolve. To the best of our knowledge, no suitable architecture has been proposed in the literature. Therefore, this article proposes a software reference architecture for integrating artificial intelligence and knowledge bases to support journalists and newsrooms. The design of the proposed architecture is grounded on the research literature and on our experiences with developing a series of prototypes in collaboration with industry. Our aim is to make it easier for news organisations to evolve their existing independent systems for news production towards integrated knowledge platforms and to direct further research. Because journalists and newsrooms are early adopters of integrated knowledge platforms, our proposal can hopefully also inform architectures in other domains with similar needs.

1. Introduction

News organisations today are forced to constantly adapt their business models to digital media innovations to increase information quality, competitiveness and growth [1,2]. Potentially news-relevant information can come from almost any type of source and in any data format. The daily global production of news exceeds 100.000 articles [3], while social media generate similar volumes within a second. Consequently, news organisations can benefit from using big data and artificial intelligence (AI) solutions to manage information, extract knowledge and create value for more and more journalistic purposes [4] including: identifying and contextualising newsworthy events to find connections along millions of articles in investigative journalism; facilitating data visualisation with the support of storytelling techniques in digital journalism; automating news writing utilising structured data to automatically create and publish reports about markets, sports and weather (a.k.a. robot journalism, algorithmic journalism or automated journalism); and, providing real-time fact-checking tools to identify fake claims using external knowledge bases in political journalism.

Unsurprisingly, both research and industry agree on the relevance and challenges associated with future AI systems across domains [5]. Particularly, future AI systems must be semantically sound and explainable, as well as foster trustworthy AI. To achieve this, these systems must be able to integrate sub-symbolic deep learning, symbolic knowledge representation and logical reasoning [6]. Knowledge graphs are a topical choice for knowledge representation and reasoning [7] alongside neural networks for implementing sub-symbolic AI. As the world is constantly changing, these systems must incorporate continuous-learning techniques to keep deep learning (DL) and machine learning (ML) models up-to-date.

Journalistic knowledge platforms. An emerging type of information system that integrates AI, big data and knowledge bases to support high-quality journalism [8]. In this article, we refer to these systems as *Journalistic Knowledge Platform* (JKPs). JKPs harvest and analyse news and social media information over the net in real time [3] and leverage encyclopedic sources [9,10]. News-relevant information is semantically annotated and represented in knowledge bases using linked open data (LOD) [11] and AI techniques like natural language processing (NLP) [12]. The resulting knowledge bases are exploited with data analysis, reasoning and information retrieval techniques to provide journalists with meaningful background knowledge and newsworthy information [13,14], as well as to help journalists and readers dive more deeply into information, events and story lines [15–17].

[☆] This work was supported by the Norwegian Research Council IKTPLUSS project 275872.

[*] Corresponding author.
E-mail addresses: marc.gallofre@uib.no (M. Gallofré Ocaña), andreas.opdahl@uib.no (A.L. Opdahl).

https://doi.org/10.1016/j.knosys.2023.110750

JKPs typically implement different mechanisms for interacting with the system, for example, to provide live feeds and alerts and to search for information. Because JKPs combine and represent personal data from different sources, they must also consider the privacy policies [18]. To combat the dissemination of fake news and misinformation, JKPs must also manage the provenance of news and its sources, facilitating its identification. All these aspects make JKPs a particularly complex kind of big-data knowledge-centric intelligent system.

Work on JKPs has so far been driven by the research community, albeit often in collaboration with industry. We envision that the field will continue to gain industrial traction in the near term. Because journalists and newsrooms are early adopters of integrated knowledge platforms in general, we also envision that our work in the journalistic domain can also inform other knowledge-intensive domains that rely critically on exploiting high-volume, high-velocity and high-variety information sources.

Software reference architecture. Today, most news organisations rely on many current, independent and task-specific production systems. However, depending on multiple systems entails a higher resource footprint compared to integrated systems that reduce code and data duplication. The utilisation of multiple systems also increases the cost of coordinating developer teams and providers, as well as the cost of maintaining and updating the systems. Organisations may lose control over their data and knowledge because their systems do not share common data repositories nor representations or are provided as Software as a Service (SaaS) by third parties. As a consequence, organisations may miss out on opportunities for exploiting potentially news-relevant information [8]. These concerns can be addressed by having a clear system design and a system architecture that allows organisations to integrate and expand their solutions in a coherent manner over time. Therefore, in this article, we propose a software reference architecture (SRA) for JKPs. The proposed architecture can also serve as an example of a more general high-level architecture for future big-data AI systems that combine deep learning and knowledge graphs and that support evolving knowledge. We are focusing in particular on JKPs that employ semantic knowledge graphs [7] for knowledge representation.

An SRA for JKPs should provide news organisations with a blueprint and associated advice for how to evolve its many current systems towards a cohesive, comprehensive, and integrated JKP [8]. On the organisational level, central challenges are that JKPs (a) are complex systems that must balance many concerns [3] and are thus challenging to adopt without architectural guidance; (b) must interoperate with a wide variety of in-house legacy systems and external services [13]; and, (c) are long-term investments that must be able to evolve to incorporate future best-of-breed components that replace or come in addition to existing ones [3]. On the technical level, JKPs need to support (a) the ingestion of big data from diverse sources, (b) the semantic annotation, representation and enrichment of news-relevant items [15]; (c) the inclusion of diverse mechanisms for serving potential newsworthy events and information [10,16]; (d) the addition of processes for continuously evolving and adapting machine learning models, ontologies and schemas [9]; (e) the integration of explainable sub-symbolic and symbolic AI approaches [8]; and, (f) the control of data privacy and provenance [18]. The research literature on JKPs has focused on the application side and addressed different challenges for news production. However, the authors are not aware of other lines of work that have studied the architecture of JKPs specifically.

SRA for JKP. Researchers have proposed several software architectures that deal with big data in general [19–24], but few of them deal with the central challenges that JKPs face and, to the best of our knowledge, none of them deals with them all. For example, the current big-data architectures are typically designed for data analysis and immutable data, whereas the focus of JKPs is on exploring and understanding knowledge that evolves over time. In addition, few current big-data architectures consider the integration of knowledge bases and AI in detail. Therefore, In this article, we address the question: "What would be a good software reference architecture for journalistic knowledge platforms?" We propose a software reference architecture that addresses the central challenges of journalistic knowledge platforms and integrates artificial intelligence and knowledge bases to support journalists and newsrooms. We also introduce two novel types of components: one for continuously improving and updating AI models and the other for curating knowledge representations. The first component includes services to monitor data and schema changes and update the models respectively. The second component scans knowledge representations to enrich the content and rectify inconsistencies or missing information. This architecture is the first of its kind proposed for the systems described in this work. Unlike existing big-data architectures focused on immutable data and data analysis, the proposed architecture focuses on evolving knowledge and analysing knowledge representations. The design of the architecture is primarily grounded in the research literature but also relies on our practical experience with developing a series of JKP prototypes in collaboration with the industry. We believe the proposed architecture can be adaptable to other domains with characteristics similar to news production.

In a previous publication [25], the authors have outlined a preliminary version of an SRA for JKPs. The present article extends the earlier outline in several ways: it presents the high-level qualities that an SRA for JKPs must satisfy; it explains the architectural principles that guided the design; it describes a generic architecture for big-data knowledge-based AI systems; it further elaborates the description and argumentation of the SRA specific to JKPs; and it compares the coverage of the proposed SRA for JKPs with the research literature.

The remainder of the article is organised as follows: Section 2 defines our terminology and introduces SRAs, knowledge graphs, embeddings and vector databases. Section 3 describes our research method. Section 4 analyses the related literature. Section 5 outlines the high-level required qualities for an SRA. Section 6 presents the SRA for JKPs. Section 7 evaluates the proposed SRA. Finally, Section 8 states our conclusions and plans for further work.

2. Background

2.1. Central terms

By *big-data technology* we mean the recent generation of middleware that accommodate web-scale data processing and storage. For example, the big-data technologies we use in our work include Apache Kafka and Cassandra. By *knowledge bases* we mean data repositories that maintain strong semantic definitions of and links between the data. Our work focuses on knowledge graphs, using techniques such as RDF, OWL and SPARQL, and Blazegraph for storage. By *AI* we mean symbolic and sub-symbolic techniques, including machine and deep learning for tasks like natural-language processing. Examples of AI techniques we use in our work are named entity linking, relation extraction and inference rules. We also refer the reader to our previous work on the usage of knowledge graphs for news [26] and our review on JKPs [8] for further details on big data, knowledge bases and AI techniques. We proceed to discuss a few other central terms in more detail.

2.2. Software reference architecture

A software reference architecture (SRA) "is a generic architecture for a class of systems that is used as a foundation for the design of concrete architectures from this class" [27]. It defines the basic software elements and data flows and captures the best practices for designing and implementing complex systems and their functionalities.

Two types of SRAs can be distinguished: practice-driven and research-driven [28]. Practice-driven SRAs are based on practical experience developing concrete architectures in a domain. They describe the "best practices" and address legacy problems. Research-driven SRAs address areas that are expected to become important in the future but where there are few or no development experiences yet. They are based on theoretical reflections grounded in the research literature.

2.3. Knowledge graphs

Knowledge graphs provide symbolic representations through concepts, relations and logic rules. According to [7], knowledge graphs capture and abstract knowledge using graph-based data models wherein entities of interest are represented as nodes and the relations between them as edges of the graph. Ontologies and rules are employed to define the semantics and terms of the graph, reason about it, and ease data integration. Knowledge graphs are particularly relevant for systems that integrate and extract value from heterogeneous and dynamic data. They are exact symbolic representations that do not require large amounts of data to become meaningful. Their workings are easy to explain to humans, but managing large graphs efficiently can be hard for computers. Compared to relational and NoSQL models, knowledge graphs facilitate semantic integration, flexible data and schema evolution, along with graph query languages for exploring complex relations through arbitrary-length paths.

2.4. Embeddings and vector databases

Embedding techniques are used in machine and deep learning to represent concepts as vectors in a latent space. Concepts can be extracted from a wide variety of data from text, images, and audio to sequences like DNA and molecular structures. Vectors are generated through mathematical models and provide sub-symbolic representations, positioning concepts in the embedding space according to similarity or other relations. These vectors, as sub-symbolic representations, have a stochastic component and require large amounts of data to be meaningful. They are hard to explain to humans, but even large collections of vectors can be efficiently managed by computers.

Well-known techniques for word and text embedding are word2vec [29] and transformers [30] like BERT [31] and, most recently, GPT-3 [32]. These techniques are particularly relevant for systems that exploit the semantic and contextual similarity between data and used in many AI applications like natural language processing, chatbots, image recognition, and recommendation.

Storing large collections of vectors requires specialised databases with optimised storage, access and search. Vector databases are an emerging technology for storing and indexing vectors efficiently and provide functionalities for retrieving vectors using similarity search algorithms like HSNW [33] and FAISS [34]. Examples of vector databases are Milvus[1], Weaviate[2] and Vald[3].

3. Method

To design and validate the SRA, we follow an established method for designing empirically-grounded reference architectures [35]. The method comprises six steps, where the initial five steps provide the "empirical foundation" and the sixth provides the "empirical validity" as illustrated in Fig. 1:

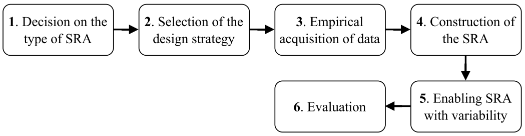

1. Decision on the type of SRA
2. Selection of the design strategy
3. Empirical acquisition of data
4. Construction of the SRA
5. Enabling SRA with variability
6. Evaluation

Fig. 1. Construction process of the SRA for JKPs.

Step 1 – Decision on type of SRA: We chose the preliminary and facilitation type of SRA described in [27]. This is a type of research-driven of SRA that aims to facilitate guidelines for designing systems and concrete architectures that are likely to become important in the future and will be utilised by multiple organisations. The guidelines are designed by researchers in collaboration with interested software organisations and grounded on existing research literature and practical experience.

Step 2 – Selection of design strategy: We chose a research-driven strategy because we expect JKPs to become increasingly important in the future and we have not identified a substantial number of industrial implementations and experiences on JKPs.

Step 3 – Empirical acquisition of data: We carefully selected 13 research projects that matched our definition of JKPs (see Section 4.1). A detailed, qualitative meta-analysis review was presented in [8] to derive the challenges, opportunities, main stakeholders, information, functionalities, techniques, components and concerns in JKPs. We also drew on our practical experiences developing a series of JKP prototypes with Wolftech[4], a software developer for the international newsroom market [36,37], and collaborating with a large cluster of news media industry partners in the MediaFutures research centre[5] [38].

Step 4 – Construction of SRA: Firstly, we identified the related literature on JKPs and software architectures for big data and semantic technologies (see Section 4). Secondly, since we could not find a suitable architecture for JKPs, we decided to propose a new SRA grounded on the specific research literature, our practical experiences and the general literature on architectural principles. From this, we derived the main required qualities for JKPs (see Section 5) and selected suitable architecture principles (see Section 6.1). Thirdly, the required qualities were mapped into architectural elements, such as components, functionalities, data stores and data flows. These identified elements were then further mapped into a high-level architecture view of the SRA (see Section 6), guided by the architecture principles. Finally, we instantiated the SRA as a proof-of-concept prototype of a JKP (see Section 7), which we used to iteratively improve the SRA.

[1] milvus.io

[2] weaviate.io

[3] vald.vdaas.org

[4] wolftech.no

[5] mediafutures.no

Step 5 – Enabling SRA with variability: By iterating over the SRA design and continuously developing and testing the JKP prototype, we improved and refined the preliminary design decisions reported in [25,37] concerning the architecture, components, principles and semantics. To facilitate the adaptation of the SRA in other domains, we drew insights from generic big-data architectures to identify common characteristics shared by our SRA for JKPs and big-data AI architectures. As a result, we propose a high-level view of our architecture for JKPs that can potentially be adapted to big data and AI systems in other domains. We also show how the high-level view is refined into our specific SRA for JKPs and further instantiated into our prototype JKP. As JKPs are an emerging field and there are not enough research results or experience available to empirically back-up architecture variants, a more thorough investigation of variability must be left for future work.

Step 6 – Evaluation of the SRA: Following the established method for empirically-grounded SRA development [35], we have evaluated our SRA proposal in two ways:

Mapping-based evaluations: To ensure fulfilment of the required functional qualities for an SRA for JKP (Section 5.2), we systematically mapped these qualities to the architecture components (Table 3). Similarly, we mapped the required non-functional qualities (Section 5.3) to both architecture principles (Table 2) and architecture component (Table 4). Finally, to ensure that our SRA proposal accounts for the components, functionalities and goals of JKPs reported in the existing research literature (Section 4.1), we systematically mapped them into architecture components (Table 5).

Evaluations by prototype development and testing: We have validated the viability of the high-level architecture view (Section 6.2) by refining it into a concrete SRA for JKPs (Section 6.3). Furthermore, we have validated the feasibility of this concrete SRA for JKPs by instantiating it into a running JKP prototype that has been iteratively developed and tested (Section 7.2).

To construct and evaluate our prototype in Steps 4–6, we have followed a design science approach [39], which "supports a pragmatic research paradigm that calls for the creation of innovative artefacts to solve real-world problems" [40]. Within the field of information systems, design-science researchers often adopt an iterative process comprising three different cycles [41], which involve understanding the application context or environment, studying and improving the theoretical framework, and evaluating the artefacts [42]. Accordingly, we have iteratively designed and validated our SRA for JKPs by developing and refining artefacts informed by the relevant literature, while considering the contextual environment of both developers and users.

4. Related literature

4.1. Journalistic knowledge platforms

Several JKPs have been proposed in the research literature. We summarise the projects we have identified, along with their industry partners, in Table 1. We can broadly categorise them into two groups: the earlier JKPs (until around 2010), which primarily focused on implementing the Semantic Web idea [43] within newsrooms, and the more recent JKPs (after 2010), which combined semantic technologies [44] with machine- and deep-learning approaches.

The earlier JKPs employed semantic technologies and ontologies to automate the metadata annotation process, combine different knowledge bases, and formalise media standards. They used ontologies in NLP pipelines, together with LOD, to automatically annotate news archives and feeds with metadata about topics, keywords, categories and other relevant information (e.g., persons, places, organisations, sentiments and relations). For example, *PlanetOnto* [45] focused on providing a knowledge management system to provide personalised semantic retrieval and search in news archives. *Neptuno* [47] developed tools for creating, maintaining and exploring news archives. *AnnoTerra* [48] proposed a prototype for integrating earth science data sources to enhance news feeds from NASA's Earth Observatory using knowledge bases. *SemNews* [49] focused on automating metadata annotation for semantic search and monitoring of RSS feeds. *Hermes* [59] proposed a framework for searching and classifying news to support decision-makers. The *BBC* used knowledge graphs and LOD to link information across news articles, enrich their Content Management System (CMS) and recommend news [9,51]. *NEWS* [13] automatised the metadata annotation of news and images and provided news intelligent information retrieval services using semantic technologies.

More recent JKPs focused on identifying and analysing events and advancing machine and deep learning for supporting journalism. A common thread among them, and some of the earlier examples, was that they deal with big data. *EventRegistry* [15] developed a tool for collecting news articles from around 75.000 multilingual sources, identifying and extracting information about the events, and summarising and visualising events from close to 200.000 articles daily. *NewsReader* [16] presented a platform for machine reading of multilingual streams of news and extracting information about what, who, where and when for representing events temporally using knowledge graphs, for example allowing users to find networks of actors and their implication over time. The platform was tested on nearly 2.5 million news articles and extracted over 1.1 billion triples from these articles. *Reuters* [55,56] developed a real-time platform to analyse around 12 million tweets per day from Twitter to identify and verify newsworthy events before they are reported by other news agencies and automate news production processes. *SUMMA* [3] developed a multilingual and multimedia platform employing NLP techniques for monitoring internal and external live media, including TV and radio broadcasts, and providing services for data journalists. *INJECT* [58] developed a tool to support journalists by providing creative angles on news stories. *ASRAEL* [10] presented a system for aggregating news articles and utilising the Wikidata knowledge base for describing and clustering events in news from a corpus of over 2 million articles.

Typical problems faced in these projects are: huge volumes of heterogeneous data, some of them arriving in real time [3,15,16]; complex processing pipelines that combine NLP, machine learning and knowledge representation [10,15,16]; and integration of legacy and external systems [9,13]. These are problems that typically call for architectural guidance.

4.2. Software architectures for big data and semantic technologies

According to existing big-data architecture reviews [19–24], only four architectures for big data [19,23,60,61] have considered semantic technologies. *LMS* [60] was designed for providing a middleware for sensor data and the Internet of Things (IoT). *SOLID* [61] adapted the principles of the *Lambda* processing architecture [62] to RDF for gathering, storing and serving big data in real time. *Bolster* [19] extended the Lambda architecture by adding a new semantic layer to represent machine-readable metadata, contrary to JKPs that represent the data semantically.

Table 1
Selected projects. N = news media partner and T = technology partner.

Project	Industry partners	References
PlanetOnto	–	[45,46]
Neptuno	Diari SEGRE^{N} and iSOCO^{T}	[47]
AnnoTerra	NASA's Earth Observatory^{N}	[48]
SemNews	–	[49]
Hermes	–	[50]
BBC CMS	BBC^{N}	[9,51]
NEWS	Agencia EFE^{N}, Agencia ANSA^{N} and Ontology Ldt.^{T}	[13,52]
Event Registry	–	[15]
NewsReader	LexisNexis^{T}, The Sensible Code Company (before ScraperWiki)^{T} and Synerscope^{T}	[16,53]
Reuters Tracer	Reuters^{N}	[54–56]
SUMMA	LETA^{N}, BBC Monitoring^{N}, Deutsche Welle^{N} and Priberam Labs^{T}	[3,57]
INJECT	Adresseavisen^{N}, AFP^{N}, The Globe and Mail^{N}, Stibo^{T}	[58]
ASRAEL	AFP^{N}	[10]
News Angler[a]	Wolftech^{T}	[25,36]

[a] News Angler is the research project in which the authors are involved.

SmartLAK [23] focused on supporting learning analytic services and defines components for validation and inference based on ontologies. However, none of them covers mechanisms for semantic data enrichment, continuously pushing live data streams, or continually (re-)training machine learning models. Four proposed architectures [63–66] have considered maintaining and updating ML models and defined specific components for storing and training them, but none of them considered semantic technologies like knowledge graphs and ontologies. Furthermore, none of the existing architectures considers the curation of knowledge representations. In conclusion, none of them is a suitable starting point for an SRA for JKPs.

5. Required qualities for the SRA

5.1. Approach

To drive the design and evaluation of our SRA, we systematically derived the required high-level qualities from earlier JKP projects reported in the literature and our previous studies [8,18]. We used the most recent JKPs to derive the qualities, while the earlier JKPs provided supplementary insights to support and augment these qualities. We divided the JKP qualities into functional (i.e., specific behaviours that the system must implement) and non-functional (i.e., general properties of the system). In addition, we identified the required general qualities for any SRA from the literature [27,67], i.e., being feasible, representative, essential, easy to grasp, long-lasting and technology independent. The derived qualities are also corroborated by the current literature on big data architectures [21,22,24] and AI systems [5], as well as they align with quality attributes of ISO/IEC 25010 [20].

5.2. Required functional qualities

`Annotating` To better manage, analyse and derive knowledge to support journalists in creating high-quality stories, JKPs must annotate content with relevant information such as people, organisations, places, relations, categories, themes and other metadata [3,15,46–50,55,58]. JKPs use semantic annotations to facilitate the representation of the meaning of concepts and relations, standardise annotations using well-defined schemas and ontologies, and improve relation and concept mining [9,10,13,16].

`Knowledge-representation` JKPs are knowledge-centric systems that provide knowledge representations of news, events and background information and the relations between them [3,9,10,15,55,58]. Most of the JKPs are focused on exploiting the relations and connections between information. Hence, JKPs must employ systems that facilitate working with relations and updating the knowledge representations [13,16,46–50].

`Enriching` JKPs must implement mechanisms to update and expand the extracted information. Because the information can change over time (e.g., a newly elected head of a government) and some other information may not be completed (e.g., an article referencing a country instead of the city where an event took place), journalists need to constantly have access to up-to-date and fine-grained information to produce high-quality journalism [3,9,10,13,15,16,46–50].

`Schema-evolution` As novel developments and themes may appear, JKPs must facilitate schema evolution [13,16,46–50]. To do so, the technologies used to represent and store schemas must provide flexible and easy mechanisms to update them.

`Model-updating` AI models must be constantly updated to follow new information and news development [13,15,46,55]. JKPs must implement mechanisms for continuously evolving ML models to provide state-of-the-art results and adapt them to new events and users' needs.

`Storage` JKPs deal with a variety of data and representation formats [3,9,10,13,15,16,46–50,55,58]. They need to access information in different ways, for example, obtaining most recent feeds in real-time, reading data in bulk, retrieving historical data and finding similar texts. Therefore, JKPs must employ different databases for specific purposes to optimise storage and access.

`Push` JKPs must continuously push potentially newsworthy events to journalists [3,13,46–49,55]. This is achieved by, for example, sending feeds or alerts to journalists according to their preferences or current work.

`Pull` JKPs must provide services for pulling information from the knowledge base [3,9,10,13,15,16,46–50,55,58]. These services typically require direct interaction with the user to search information.

`Data-ownership` Stories generated with the support of JKPs are disseminated to a broad or even worldwide audience and the information resources need to be protected as they may be subject to ownership and usage policies [13,48]. JKPs must keep track of these policies and their implications, especially when information is merged or derived from multiple sources.

`Privacy` Merging and connecting information from different sources and social media can lead to data privacy challenges [18]. Hence, JKPs must implement mechanisms to monitor potential data privacy violations.

`Provenance` Information about the source from where the information has been derived helps journalists to assess the quality of news and find its origins. In addition, metadata about the process and version that gathered, modified or updated the information facilitates the traceability of the process and detection of errors [13,16,48]. Thus, JKPs must facilitate keeping track of the metadata associated with the information sources and processes.

5.3. Required non-functional qualities

`Interoperability` JKPs interoperate with heterogeneous in-house legacy systems, external services and other JKPs [3,9,10,13,16,46–49]. To do so, they need to provide clear meaning and data representations, as well as use standard formats, interfaces and exchange protocols.

`Modularity` JKPs must be able to incorporate future components that replace existing ones [9,13,16,48]. This guarantees the addition of new components to adapt the JKPs to particular users' needs and update them with future best-of-breath solutions. Hence, the implemented components need to be independent, modular and abstract.

`Scalability` To deal with large volumes of news-relevant information and sources, JKPs must employ tools and storage systems designed to increase to, efficiently support and uniformly process big data volumes [3,9,10,13,15,16,47,55].

`Velocity` JKPs support news production where time is a critical factor and delays can lower the value of information. News-relevant information is rapidly and continuously produced and broadcast worldwide [3,16,47,55]. JKPs must obtain this information, process it, analyse it, and make it available as soon as possible to maximise its value.

`Variety` News-relevant information is produced and broadcast as unstructured and structured data [3,9,13,16,47,48]. It comprise diverse modalities of data like audio, video and images, structuration principles like tables and graphs, time cycles like live and historic, and formats like plain text, RDF and JPEG. Therefore, JKPs must be able to ingest, process and store varied data consistently.

`Knowledge-evolution` As the world is constantly evolving, current events and developments become past and are preceded by new ones [13,15,50]. Therefore, JKPs must implement components that can adapt their behaviour in response to emerging entities, events and relations along with new terms and their meaning.

`Sub-/symbolic-AI` JKPs integrate sub-symbolic and symbolic techniques. This integration benefits both approaches: sub-symbolic techniques may be enhanced with logic and reasoning from symbolic AI, and symbolic AI may be sped up with sub-symbolic techniques [6,68]. To support this integration, JKPs must facilitate both symbolic representations like knowledge graphs and ontologies and sub-symbolic data like models and training materials.

`Trustworthy-AI` As JKPs support journalists in creating stories that may effect society, journalists need to trust the system [9]. Hence, following the European guidelines on AI [69], JKPs must allow journalists to take informed decisions, ensure data privacy and integrity and provide transparent, traceable and explained solutions.

5.4. Qualities addressed by the JKP projects

In Appendix we trace from which projects each quality has been derived (see Tables A.6 and A.7). The analysed projects addressed qualities primarily related to the technical and research challenges such as `Annotating`, `Knowledge-representation`, `Enriching`, `Schema-evolution`, `Storage`, `Push`, `Pull`, `Interoperability`, `Scalability` and `Variety`. Only a limited number of projects addressed qualities that support the validation of the newsroom production such as `Data-ownership` and `Provenance`. Despite early examples of `Model-updating` in the earliest projects, advances in machine learning increased its relevance in the newer projects, while `Sub-/symbolic-AI` remains unexplored. The rise of web-scale volumes made `Velocity` relevant. Earlier projects explicitly addressed `Modularity`, which newer projects achieve implicitly through technological decisions. Few projects considered the `Knowledge-evolution` as opposed to static knowledge. `Privacy` has not been addressed by any project, despite its importance for complying with regulations such as the GDPR. `Trustworthy-AI`, relevant for providing safe systems and understanding their outcomes, was only addressed by one project. We particularly emphasise on the underrepresented yet relevant qualities for future systems.

6. Software reference architecture for JKPs

6.1. Architectural principles

We propose a set of architectural principles for the SRA for JKPs. These principles are composed of different architectural patterns and technologies that we consider the most appropriate to fulfil the elicited high-level non-functional qualities as illustrated in Table 2.

Microservice architecture pattern. Microservices is an architectural pattern for applications where every functionality is deployed as its own service and often independent from the others [70]. Components in a microservice system are self-contained, loosely coupled, technology neutral, reusable and specialised. They typically communicate via clear APIs. These characteristics facilitate components replacement, integration, scaling and distribution. This pattern is an ideal candidate for an SRA for JKPs as it provides `Interoperability` and `Modularity` by design. Components designed following the microservice architecture principles can (a) be easily deployed, integrated and updated because they have clear boundaries and minimal technological dependencies on other components; (b) be dynamically replicated to meet specific processing loads; and, (c) be utilised independently or in collaboration with other components to fulfil business functionalities. Solutions like *Docker*[6] containers can be used to improve the availability, `Scalability`, replaceability and deployment of microservices.

Liquid architecture pattern. Liquid architecture [71] is an architecture pattern for integrating nearline and offline big-data processing with two distinguished layers: the processing and the messaging layer. The processing layer executes ETL-like jobs for different back-end systems. The messaging layer follows a topic-based publish/subscribe communication model where streams of incoming messages are identified by topics. Jobs can read from selected topics and output to new ones. Messages can contain metadata annotation such as timestamps that are used to provide stateful and incremental processing. Each job is an isolated resource that may perform several tasks and communicate with other jobs, creating a dataflow processing graph. Compared to

[6] www.docker.com

Table 2
Connection between non-functional qualities and architecture principles.

	Microservices	Liquid	Blackboard	Semantic technologies
`Interoperability`	✓			✓
`Modularity`	✓			✓
`Scalability`	✓	✓		
`Velocity`		✓		
`Variety`				✓
`Knowledge-evolution`			✓	✓
`Sub-/symbolic-AI`			✓	✓
`Trustworthy-AI`				✓

the well-known Lambda [62] and Kappa [72] architectures, Liquid does not duplicate the code, as opposed to Lambda; and, it does not need to reprocess the current data view to run batch jobs, as opposed to the Kappa. Unlike other architecture patterns like Phi [73] that offer similar benefits, Liquid provides resource isolation and incremental data processing, as well as it removes the need for duplicating the data for downstream processing. The Liquid architecture pattern is an excellent candidate for an SRA for JKPs, because it is designed for meeting the `Scalability` and `Velocity` requirements of big data, and reduces the development, maintenance efforts and hardware demands [71]. Event-streaming solutions like *Apache Kafka*[7] can be employed to implement the message layer.

Blackboard model. The blackboard model is a problem-solving approach to solve complex problems where different kinds of domain knowledge and expertise are needed [74]. In a blackboard-based system, independent components cooperate to solve problems using a shared knowledge base (viz., the blackboard) [75]. These components activate when there is a change in the knowledge base or an event that meet a certain condition. They may also modify the knowledge base to contribute towards the solution. The behaviour of these components depends on the current state of the knowledge base and adapts as the knowledge evolves. This increases the response of the components of JKPs to `Knowledge-evolution`. As components cooperate to solve a problem by sharing resources or preliminary solutions, components can use the output of a sub-symbolic method to build on a symbolic one or vice versa. Hence, the blackboard model facilitates the integration of `Sub-/symbolic-AI`.

Semantic technologies. Semantic technologies encompass technologies and standards in the context of the Semantic Web [43] that deal with the meaning of the data rather than its structure [44]. They are designed to represent entities, their relationships and attributes using well-defined ontologies, and optionally, logic rules. Semantic technologies are commonly used to construct knowledge graphs and integrate LOD [7]. Software architectures of JKPs implementing semantic technologies may benefit from (a) language neutrality and clear representations of data and meaning to improve `Interoperability` and `Modularity` between systems and `Variety` of sources and data formats; (b) LOD sources that constantly update their knowledge bases like DBpedia and Wikidata to improve `Knowledge-evolution`; (c) using knowledge graphs to facilitate `Trustworthy-AI` by improving explanations, for example, by exposing connections and relations, providing context and showing semantic similarities [76]; and, (d) the symbolic representations that these technologies provide to enable `Sub-/symbolic-AI` integration.

6.2. High-level view

Fig. 2 proposes a high-level SRA for journalistic knowledge platforms, which implements the architectural principles. The architecture is centred around the knowledge base and includes four types of components. The *Knowledge Base* manages all the information needed to operate the system, including data stores, knowledge representations, schemas, ontologies, metadata, and even AI models which are provided and deployed as a service to facilitate their access and integration in the architecture. The knowledge base utilises unique identifiers, such as IRIs, to ensure consistent representation across services. This eases the integration of knowledge representations and vectors from AI models, as they share the same identifiers. To integrate the knowledge base with the rest of the system, the system is designed following the blackboard model, where the knowledge base is placed in the centre of the system and communicates bi-directionally with the other components. The *Input* components collect and analyse relevant information from outside the platform and store it in the knowledge base. The *Output* components provide services to push relevant information to users and let users pull information from the knowledge base on demand. The *Learner* components employ continuous-learning techniques to keep the ML/DL models, ontologies and schemas up to date. The *Curator* components maintain and improve knowledge base contents. The principal difference between the Learner and the Curator is that the Learner is focused on improvement at the type or meta level, whereas the Curator is focused on the instance or production level. For example, the Learner may include services for automatic re-training of ML/DL models or semi-manual update of the schema and ontology. The Curator services deal with knowledge fusion and enrichment tasks such as addressing knowledge contradictions, updating knowledge representations based on external knowledge bases like DBpedia, merging similar knowledge representations, and controlling potential privacy violations as the ones described in [18].

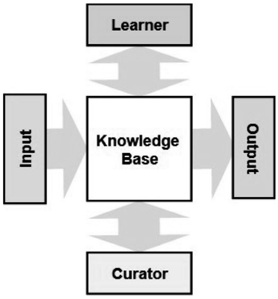


Fig. 2. High-level view of the architecture.

[7] kafka.apache.org

To facilitate machine-readable and understandable data, all components utilise semantic descriptions to represent and describe the content, thereby reducing ambiguities and facilitating integration and communication. Furthermore, to guarantee to ensure traceability of every piece of information back to the generating process, every service of the SRA maintains `Provenance` information. Our proposed architecture goes beyond the existing big-data architectures in the literature, as it explicitly incorporates components like the Curator and Learner, providing clear pathways for their inclusion.

6.3. SRA for JKPs

Fig. 3 illustrates the SRA for JKPs in more detail. While the high-level view in Fig. 2 may apply to other knowledge-based domains, this architecture is specific to news work. To align with existing literature on JKPs and enhance comprehension, we have renamed the components of the SRA accordingly, providing a nomenclature that reflects their most common intended purposes. We also decided to split the Output component into a Feeder and a Retriever to differentiate between the `Pull` and `Push` types of interaction in JKPs. As a result, the SRA for JKPs is composed of six groups of components, namely, the *Ingestor*, *Knowledge Base*, *Curator*, *Learner*, *Feeder* and *Retriever*, each consisting of several microservices. A less refined version of the architecture was presented in [25], which did not consider, among other things, explicit components for training and updating AI models and schemas nor the Current Window and Vector Store of the Knowledge Base.

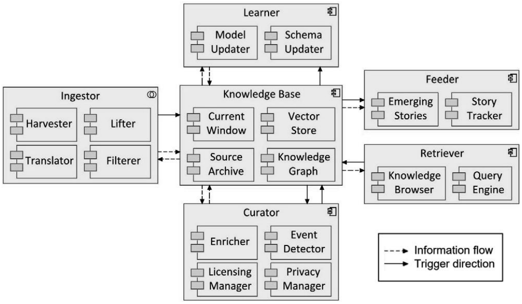


Fig. 3. The SRA for JKPs (represented with ArchiMate 3.1 notation).

6.3.1. Ingestor

The Ingestor collects and `Annotating` potentially news-relevant information items such as news articles and social media messages, multimedia files and structured data from online sources. The most relevant components are the *Harvester* and the *Lifter*. The Harvester continuously downloads and ingests scheduled and real-time news-relevant items from sources like RSS, APIs and websites. To handle data `Variety`, multiple harvesters can be deployed, each of them targeting specific sources or data formats. The Lifter annotates and transforms these news-relevant items into `Knowledge-representations` using semantic technologies and AI techniques before they are uploaded to the Knowledge Base. The resulting knowledge representations can be using RDF and predefined ontologies, such as the Event Description Ontology [77]. Ontologies must be designed general enough to facilitate `Schema-evolution` and `Interoperability` between services. Lifters are composed of different AI modules specialised in different tasks (e.g., named entity recognition and face recognition), which are designed to be replaced or extended (`Modularity`) to follow the state-of-the-art [12]. To combine the results from the different AI modules and improve data `Interoperability`, these can use vocabularies for representing annotations like NLP Interchange Format (NIF) [78] or NLP Annotation Format (NAF) [79]. Each annotation must provide information about its quality (e.g., accuracy and support values), the `Provenance` to trace back to the source and process that generated it, the `Data-ownership` and the terms of use.

Additional services like the *Translator* and *Filterer* can be added to pre-process and clean the collected news-relevant items. For example, the Translator service can be utilised for translating the text into a canonical language, while the Filter can handle tasks like normalising data types, standardising formats, and filtering out advertisements. In some cases, it may be necessary to employ other services that can group micro-texts, such as Twitter messages, into chunks of similar messages, enabling them to be processed collectively. These micro-texts may not be relevant enough on their own, but they may turn relevant when analysed together or when many similar texts occur at the same time or within the same location.

As a result, the Ingestor performs the real-time transformations once and near the source before they are stored in the Knowledge Base, and processed further by the Feeder and Curator. As this provides both the raw data and knowledge representations from the beginning, it facilitates the deployment and integration of `Sub-/symbolic-AI` approaches. By processing the data near the sources, we also avoid software and data duplication, reducing the computational resources needed to deploy the platform. Hence, this can have an impact on the overall power and resource utilisation.

6.3.2. Knowledge base

The Knowledge Base provides persistent `Storage` and is composed of multiple specialised databases for different data, including raw files, metadata, `Knowledge-representation`, schemas, ontologies, vectors and ML/DL models. The use of IRIs facilitates data identification across databases and provides `Provenance`. The usage of specialised databases is encouraged to optimise data storage as they can efficiently manage specific types of data and perform better on certain types of queries and workloads. For example, by only storing the knowledge representations in a knowledge graph and storing the raw data text in a different database, we can reduce resource utilisation associated with knowledge representations and improve performance when exploring relationships between concepts.

The *Source Archive* can be composed of multiple databases for managing a `Variety` of raw files such as text and multimedia files. The *Knowledge Graph* stores the knowledge and schema representations. It represents news-related knowledge and provides a historical data view that can be updated to capture `Knowledge-evolution` and `Schema-evolution`. Semantic technologies like RDF and triple-store databases facilitate both the `Knowledge-evolution` and `Schema-evolution` because they provide flexible data representations and natural integration with linked data. The Knowledge Graph is used as a hub to provide `Interoperability` with the different repositories and integrate legacy archives as it can be used to manage data lakes [80]. The *Current Window* provides a live and dynamic view of the most recent data and knowledge representations coming from the Ingestor, Curator and Retriever. It must provide real-time responses and support streaming operations. Many machine and deep learning solutions are often based on embedding techniques for representing the meaning of, for example, text, graphs and images. The *Vector Store* stores these embeddings in vector databases. Vector databases facilitate the availability and search of vectors, reduce the need for re-computing them and provide similarity functions that can be used to optimise information retrieval. Some vector databases offer the possibility of storing the metadata associated with the vector (e.g., if multiple news-relevant items have been used to generate the vectors, we can add their IRIs as metadata). This can enhance the level `Trustworthy-AI` on JKPs, as it allows for more explainable AI by providing `Provenance` to the vector representations and their resulting outcomes. At the same time, interlinking all databases and vector representations with IRIs simplifies data collection and generation for solutions that integrate `Sub-/symbolic-AI` and update ML/DL models.

These storage services must handle large volumes of data, intensive write and read operations in real-time (`Velocity`), and horizontally scale (`Scalability`). For example, distributed databases like *Apache HBase*[8] and *Cassandra*[9] can store large data volumes. Although many of the open-source graph databases and triple stores with support for RDF and SPARQL do not provide support for scaling horizontally, some of them can hold more than one billion (10^9) triples [81] (e.g., *Blazegraph*[10] and *Jena TDB*[11]). Strategies like partitioning the graph databases according to resource types/predicates, temporal aspects, themes and geolocations, or a combination of these can be employed for distributing graph databases.

6.3.3. Curator

The main purpose of the Curator is to make the Knowledge Base as useful as possible for journalistic purposes. The *Enricher* enhances `Knowledge-representation` using external information from the LOD (e.g., Wikidata). It enriches the Knowledge Graph by adding linked data retrieved from the LOD cloud to expand the represented news-relevant information and events (`Enriching`) and updates or corrects them following the latest advances (`Knowledge-evolution`). The *Privacy Manager* monitors incoming data to identify and propagate prohibitions, permissions, obligations and violations [18] and outputs alerts that need to be rectified by the user. The *Licensing manager* controls the data copy-rights and licensing in the Knowledge Base. For example, when different permissions are merged, the licensing manager maintains data usage obligations and restrictions, identifies the `Data-ownership` conflicts and adds the corresponding missing information.

Additional services can be added to analyse news-relevant information and events and produce newsworthy information for journalists. For example, the *Event Detector* detects newsworthy events from social media and other sources. An aggregator service can incrementally relate and cluster news-relevant items into more comprehensive and reliable event representations or storylines. A network analyser service can identify and analyse different types of connections between actors and their relations with the events. An angle detector service can derive the news angles that fit an event [17,82]. An analogy service can find analogies between different news-relevant items [83]. The Curator can also contain services to provide explanations of the AI results to users (`Trustworthy-AI`).

6.3.4. Learner

The *Learner* provides services to keep the AI models and schemas up-to-date. The *Model Updater* uses continuous-learning techniques such as incremental or online training approaches to improve those models that depend on the frequency and context of words and entities or user preferences. The process of `Model-updating` can be triggered when a significant frequency of an unknown word/entity is detected to incorporate it into the model (e.g., the first mentions of "COVID-19"), use the information from the last week to evolve the model (e.g., after unveiling a corruption case some politicians should be placed closer to the corruption theme), or adapt the recommendations following the current work of a journalist (e.g., when the journalist starts working on a new story). To generate training materials, the Model Updater can access the stored data in the Knowledge Base and the latest version of external repositories such as Wikidata and Dbpedia or the most recent and current events in the Current Window. The resulting updates can be stored in the Vector Store or change the models used by other services, creating a new version of the model. The *Schema updater* evolves current schemas or mappings like themes or categories to include newer elements or remove obsolete ones based on the incoming or on-demand data (`Schema-evolution`).

II

6.3.5. Feeder

The Feeder monitors streams of linked data coming from the Current Window to `Push` live information to journalists. It continuously pushes newsworthy feeds and alerts based on users' preferences. For example, the *Emerging Stories* service can be implemented to identify live stories that are gaining attention from various publishers or social media platforms, and a *Story Tracker* to push stories that are related to the current journalists' work. As the Feeder is intended to push information to journalists as soon as it is captured or generated, it improves the `Velocity` of information transmission and discovery and reduces delays. The Feeder implement end-points where other services can connect to get live feeds and alerts (`Interoperability`).

6.3.6. Retriever

Retriever allows users to `Pull` information from the JKP on demand. For example, the *Query Engine* facilitates querying and analysis of data from the knowledge base, generating data visualisations, access to taxonomies, and retrieval from news and multimedia archives. It can provide an end-point with pre-packaged queries for particular purposes, like finding news stories related to a particular person and retrieving relevant information for a given event. The *Knowledge Explorer* provides access to background and related information from external sources. Additional services can provide tools such as currency and time converters, related story retrieval, and suggestions to enhance news stories, including news angles. These services are also exposed as API to allow external users and systems to pull information from the JKP (`Interoperability`).

7. Validation

As explained in the Method section, we have validated our proposed SRA in two ways: (1) by mapping between each of the required qualities from Section 5 and our SRA; and (2) by iteratively developing and testing a prototype implementation of the SRA.

7.1. Mapping

We established mappings to verify that the proposed SRA fulfils all the required qualities. To do so, we examine which components contribute towards each quality and how.

[8] hbase.apache.org

[9] cassandra.apache.org

[10] www.blazegraph.com

[11] jena.apache.org

Table 3
Mapping between functional required qualities and components.

Functional quality	Ingestor	Knowledge Base	Curator	Learner	Retriever	Feeder
`Annotating`	News items annotation					
`Push`						Feeds and alerts
`Pull`					Query on demand	
`Model-updating`				Learning techniques		
`Enriching`			LOD addition			
`Knowledge-representation`	News items to graphs		LOD addition			
`Storage`		Persistent databases				
`Schema-evolution`				Schema updates		
`Data-ownership`	Terms-of-use metadata		Terms-of-use monitoring			
`Privacy`			Data privacy monitoring			
`Provenance`	Tracing metadata	IRI	Tracing metadata	Tracing metadata	Tracing metadata	Tracing metadata

Table 4
Mapping between non-functional required qualities and components.

Non-functional quality	Component	Principle
`Interoperability`		Microservices, Semantic Tech.
`Modularity`		Microservices
`Scalability`		Liquid Architecture, Microservices, Blackboard Model
`Velocity`	Feeder	Liquid Architecture
`Variety`	Ingestor, Knowledge Base	Blackboard Model, Semantic Tech.
`Knowledge-evolution`	Curator, Learner	
`Sub-/symbolic-AI`		Blackboard Model, Semantic Tech.
`Trustworthy-AI`	Curator	

Functional qualities:. To validate that all required functional qualities (Section 5.2) are covered by the SRA, we mapped them to the architecture components. Table 3 shows the components responsible for providing or realising each functional quality and highlights the key aspects that support it. Certain qualities may be associated with multiple components.

As shown in Table 3, the SRA defines Ingestor components for `Annotating` news items and transforming them into `Knowledge-representations`, which are enriched in the Curator with LOD. It also defines components such as the Feeder to `Push` live feeds and alerts and the Retriever to `Pull` information on demand. To keep ML/DL models and schemas up-to-date, the SRA defines the Learner that employs different learning techniques to evolve them. The news-relevant information is persisted in specialised databases in the Knowledge Base. To keep track of `Data-ownership`, the Ingestor adds terms-of-use information to each news item and the Curator monitors them, as well as, potential `Privacy` violations. All components add tracing metadata to provide `Provenance` and utilise IRIs to facilitate identification across services.

Non-functional qualities:. The non-functional qualities (Section 5.3) depend on the architectural principles, the development decisions and specific components, as shown in Table 4. To achieve `Interoperability`, the SRA is based on semantic technologies and vocabularies, schemas, linked data and open standards. These technologies provide language neutrality, formal data representations, open definitions and clear `Knowledge-representation`. This brings data understanding at a conceptual level and facilitate data integration and fusion.

To achieve `Modularity`, the SRA is based on microservice principles. The different components must have clear functional boundaries, so they can be deployed independently and interact between them effortlessly. This facilitates the replacement and addition of new components without affecting the current ones or modifying them. Well-defined boundaries also facilitate communication with external users.

The data ingestion part of the SRA is inspired by the Liquid architecture and microservice principles to handle big data `Scalability` and `Velocity` needs. The Liquid architecture combined with microservices offers a development pattern for designing scalable systems. The annotation components are built upon the blackboard model which is designed for parallel processing and ease scalability. Data `Variety` can be handled by adding new processes to the Lifter for the different types of data and formats. To combine different types of data and annotations, the SRA for JKPs uses semantic technologies to represent them. In addition, the SRA is designed with a knowledge base that can integrate specialised databases for the different types of data. In addition, the SRA enhances the response time with components like the Feeder.

We designed the SRA with the Curator and the Learner components to manage the `Knowledge-evolution`. The Curator contains services for maintaining and improving knowledge representations using external and internal knowledge. While the Learner contains services for updating the ML/DL models to adapt knowledge extraction and representation to current developments.

The integration of `Sub-/symbolic-AI` is accomplished by using the blackboard model and semantic technologies. The SRA makes the source data and its symbolic and sub-symbolic representations available in the knowledge base, while it keeps the involved concepts aligned using IRIs. This allows the design of solutions that can exploit both types simultaneously.

Table 5
Mapping between architecture components and projects.

Project	Ingestor	Knowledge Base	Curator	Learner	Retriever	Feeder
PlanetOnto	✓	✓		✓	✓	✓
Neptuno	✓	✓		✓	✓	
Annoterra	✓	✓	✓	✓	✓	
SemNews	✓	✓		✓	✓	
Hermes	✓	✓	✓	✓	✓	
BBC CMS	✓	✓	✓		✓	
NEWS	✓	✓	✓	✓	✓	✓
Even Registry	✓	✓	✓	✓	✓	
NewsReader	✓	✓	✓	✓	✓	
Reuters Tracer	✓	✓	✓	✓	✓	✓
SUMMA	✓	✓	✓		✓	✓
INJECT	✓	✓	✓		✓	
ASRAEL	✓	✓	✓		✓	

We designed the SRA to favour `Trustworthy-AI` through services that control `Privacy` and `Data-ownership`, provide `Provenance` and enhance explainability through the usage of semantic technologies and symbolic representation.

Comparison with existing JKPs. Finally, to validate that our proposed SRA is able to account for all the elements of the various JKPs reported in the literature (Section 4.1), we have reviewed them carefully and mapped their elements into the corresponding parts of our SRA, as shown in Table 5. Because many of the analysed projects built pipeline-based systems with little architectural description, the elements we mapped were sometimes suggested solutions and processing steps based on their goals and functionalities. We managed to map all the related JKPs into our SRA which indicates that our proposed SRA for JKPs is able to account for existing JKPs reported in the research literature.

7.2. Prototype

In order to evaluate the feasibility of our proposal, we developed a prototype platform that instantiates the SRA for JKPs (Fig. 4).

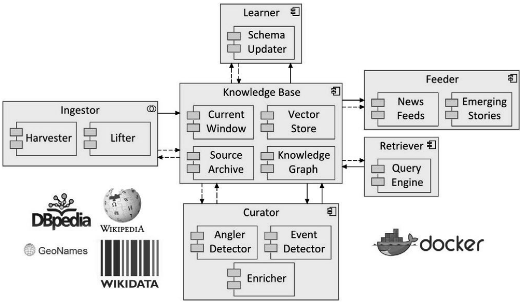


Fig. 4. The instantiated architecture for the JKP prototype.

Ingestor. We implemented the Ingestor with services for Harvesting and Lifting. Our Harvester crawls news-related websites and harvests RSS feeds, Twitter accounts, NewsAPI[12] and GDELT[13]. The Twitter API provides real-time tweet streams from specific accounts, geographical areas or topics. NewsAPI aggregates and provides streams of news articles from over 80000 news sources and blogs. GDELT provides semi-structured information about conflict events, collected from news all around the world and automatically translated into English from 65 different languages. We observed that RSS support is declining among news organisations. This makes aggregation services like NewsAPI and GDELT a solid alternative to consider, as they provide access to a larger number of news sources. We also observed that data from news organisations' and journalists' Twitter accounts cannot be used straightforwardly as many messages only provide links to news articles or little information. Hence, these messages need to be aggregated in chunks of information before they are processed and the information from the links must be downloaded. Harvesting news from different sources also creates duplicates that need to be filtered out. Many of them can be easily identified using URLs. However, it is not always trivial to filter similar news, because the same URL can provide updated content or a different URL can report the same news with a few newsworthy modifications in the content or even contradictions.

Our Lifter [12] transforms the news and event streams into semantic knowledge representations in real time according to an event-description ontology [77]. It combines out-of-the-box NLP systems such as DBpedia Spotlight[14] and SpaCy[15] and different end-to-end deep learning models for semantically annotating potentially news-relevant textual items with named entity linking, relation extraction, sentiment and topic annotations, and links to Wikidata and DBpedia. To integrate these annotations, we use the NLP Interchange Format (NIF) [78]. As these components have been designed as microservices with clear functional boundaries to facilitate `Interoperability` and `Modularity`, variants of the same components can be used to transform unstructured data from RSS feeds and structured data from GDELT to knowledge graphs. Furthermore, to scale the prototype in order to handle the large amount of data produced by GDELT, we replicated the GDELT Lifter components to avoid bottlenecks.

Knowledge base. The Knowledge Base includes a Source Archive service implemented with Apache Cassandra, a Knowledge Graph implemented with Blazegraph, a Current Windows built with Apache Kafka and ksqlDB as a stream store. We are also in the process of incorporating a vector store implemented with Vald. Cassandra is used to store the textual information together with the IRIs of the news-relevant items represented in the knowledge graph. This decision allows us to reduce the data stored in the knowledge graph; provide provenance by tracking news representations back to their source; and facilitate new training material for ML models based on the current state of our system. The Knowledge Graph is distributed over four instances of Blazegraph: one dedicated to storing news-relevant items from news articles and Twitter messages, and three to storing the events from GDELT. These four instances ingest around 11M ($11 \cdot 10^6$) triples daily from news and tweets, and another 11M ($11 \cdot 10^6$)

[12] newsapi.org
[13] www.gdeltproject.org
[14] www.dbpedia-spotlight.org
[15] spacy.io

from GDELT events. In a period of 6 months, it can ingest more than 4B ($6 \cdot 10^9$) triples in total. We have observed that these large amounts of triples cannot be held in a single triple-store instance without affecting its performance.

Curator. The Curator implements an Enricher, an Event detector and an Angle detector. Our Enricher extends the annotated items with location-related background information extracted from DBpedia, Wikidata and other LOD sources. Our Event Detector provides journalists with aggregated and real-time events detected from GDELT streams. The Angle Detector analyses the representations of the news items to identify location angles for a set of selected locations of interest [77]. As the Angler Detector analyses the news-relevant item representations from all sources, we had to replicate it to meet the velocity and volume demands. The Learner implements a Schema updater that monitors incoming GDELT events to identify new themes and update our themes hierarchy and mappings accordingly.

Feeder and Retriever. The Feeder provides an API that exposes a feed of annotated news-relevant items from the Knowledge Base and allows external users to interact with our system. To explore co-development with external contributors, we ran a research challenge[16] where external developers were invited to submit solutions using live feeds directly from the knowledge base [84]. In addition, we collaborated in another research challenge[17] where participants had access to news and images from our system to explore the connection between text and images. Our Retriever exposes several APIs to access the Knowledge Base and other services of the system. On top of the Retriever API, we developed an editing interface for journalists to recommend relevant information for the story the journalist is working on and provide background information for the entities present in the text. We have observed that extracting background information from Wikidata and DBpedia presents many challenges as entities from the same categories are not in general represented following the same structure and using the same properties.

Infrastructure. Our prototype runs on 28 cloud instances (with a total of 94 vCPU, 312 GB RAM and 20 TB disk)[18]. We used Ansible and Terraform to automatically set up the instances and Docker Swarm for orchestrating a total of 114 services as containerised applications (i.e., 70 services related to the JKP and the rest for monitoring these services and the cloud instances). These services run as containerised applications and are exposed through APIs that facilitate their replacement with newer versions without affecting the performance of the platform. We also decided to decouple the ML/DL models from the applications by exposing them through APIs, allowing us to switch the model at any time. By following these principles, we developed a system that allowed us to experiment with and meet the `Scalability` and `Velocity` requirements. Our prototype downloads news items and transforms them into graphs within an average of 21.112 seconds per news item, with a standard deviation of 9.906 seconds, including sleep and network waiting times. If we only take the system and user CPU time, our prototype takes an average of 0.246 seconds per item, with a standard deviation of 0.083 seconds. The text of the news items varies in length, with an average length of 3305.18 characters per item and a standard deviation of 3742.22. This reflects on the extracted graphs that have an average of 712.38 triples per graph, with a standard deviation of 550.355. To further evaluate these qualities, we plan to conduct stress-testing experiments with our prototype by, for example, processing all tweets produced by Twitter in a single day or re-processing in a single day the equivalent of the harvested news in a month.

To support communication between services, we serialised the messages using *JSON-LD*[19] and semantic vocabularies, as well as employed Apache Kafka as a message broker. JSON-LD is a popular extension of JSON for serialising linked data. Apache Kafka is a framework where independent services communicate through subscription to topics and production and consumption of messages with associated metadata (e.g., topic, key and timestamp). To add or duplicate a service, we only needed to assign it to the desired message stream. This also allowed us to duplicate services to meet specific workloads or add new ones effortlessly.

Earlier JKP prototypes, that ran on a simpler infrastructure without an equally carefully planned architecture [36], have already implemented additional functionalities, which we plan to adapt into the current prototype. In the development of the prototype, different people contributed to adapting the old components and creating new ones while the JKP was running. This was in part possible by meeting the modularity requirement.

7.3. Threats to validity

Completeness. Completeness deals with the selection of earlier projects that our work as based on. Our selection of JKP projects has been systematically carried out in the context of an extensive literature review of research on knowledge graphs for the news [26]. We have not included non-JKP projects in our work because JKPs exhibit a unique combination of characteristics that we have not seen in other domains, such as real time, web-scale data volumes, social media, text, multimedia, reference information from other sources and evolving concepts and stories. A limitation is that our work only covers the English-language literature, and we have not identified any relevant JKPs developed in geographical regions outside Europe, Canada and USA. We have also only covered the research literature. Although there are many commercial tools available for journalists and newsrooms, they tend to be focused on single tasks and not on the platform and architecture levels we address in this work.

Project access. Project access deals with the availability and reliability of information about earlier projects that our work is grounded in. We have selected primary accounts of JKP projects that are published in and available through reputed and peer-reviewed international journals and conferences. However, as we did not have access to the code of the related JKPs for re-implementing them following our SRA, nor access to their input and output data for comparing them with our proposed solution, we could not run more detailed comparative evaluations.

Internal validity. Internal validity deals with the clarity of the connections between evidence and conclusions. The qualitative validation is based on our own reading and interpretation of the requirements as presented in the primary studies. We traced each requirement (both functional and non-functional) backwards to its source, both in Section 5 and in the Appendix, and mapped each requirement forward to a specific design decision in Section 7.1. In addition to these mappings, the prototype is a direct instantiation of the SRA for JKPs, and we have shown that it adheres to all of the architecture principles outlined in Section 6.1. Although the authors of this work are involved in the development of the News Hunter platform, we have sought to reduce bias by limiting the contribution of News Hunter to supporting and extending the analysis of the literature and design of the SRA.

[16] https://multimediaeval.github.io/editions/2021/tasks/emergingnews

[17] https://multimediaeval.github.io/editions/2022/tasks/newsimages

[18] The cloud instances run on different models of CPU (Intel Xeon CPU E5-2680 v3 @ 2.50 GHz, Intel Xeon CPU E5-2680 v4 @ 2.40 GHz, Intel Xeon Gold 6226 CPU @ 2.70 GHz, Intel Xeon Gold 5317 CPU @ 3.00 GHz, AMD EPYC 7452 @ 2.35 GHz) and memory speeds (2133 MT/s, 2400 MT/s, 2933 MT/s, 3200 MT/s) respectively.

[19] www.w3.org/TR/json-ld11

External validity. External validity deals with the usefulness and validity of our results in other JKP contexts and other big-data and AI domains beyond journalism. To ensure usefulness and validity in a broad range of JKP contexts, we have taken all the relevant research projects we have found into account and we have collaborated with industrial users of more focussed journalistic tools. To facilitate usefulness and validity in other domains with similar required qualities, we have presented a high-level view of the SRA that can potentially be adapted to other application areas. However, our SRA has only been validated for JKPs, and it would need further validation to be used for other purposes. Language and region aspects should not pose a problem in terms of generalisation, but multimedia analysis remains an area for further research.

8. Conclusion

Grounded in the existing literature and supported by our practical experience, we have proposed an empirically-grounded SRA for JKPs. The purpose was to make it easier for news organisations to evolve their existing independent systems for news production towards integrated journalistic knowledge platforms and to direct further research. Although the SRA has been driven by the needs of journalism and news organisations, we have also presented a more high-level view of the SRA that can potentially serve as a proposal for a generic architecture for big-data and knowledge-based AI systems in other domains.

Our architectural decisions are based on reported experiences with existing platforms, supported by our own experience developing a JKP in collaboration with industry partners. The SRA is based on proven architecture concepts and is designed to be technology independent, open-ended and long-lasting, with components and services that can be replaced and integrated with other systems. It covers those components and functionalities that are essential for JKPs and introduces Learner and Curator components that are not considered in the previous literature. It provides a vocabulary to compare and understand different realisations of JKPs.

To demonstrate the feasibility of the proposed SRA, we have implemented a proof-of-concept prototype of JKP that instantiates it. We have developed the prototype iteratively and incrementally in order to continuously evaluate our SRA design. Validating the SRA in a newsroom production environment and assessing its real and perceived value for practitioners is left for further work.

In further work, we also want to explore the combination of knowledge graphs and vector databases and its implication for our architectural decisions. One possible benefit is improved explainability. We want to investigate how the results of machine learning techniques that employ vectors can be explained by analysing the knowledge representations related to those vectors. Moreover, we want to explore the benefits of expanding the Learner to learn from whole system, to learn not only from the AI models and knowledge representations but also from the usage and performance of each component. This will allow the Learner to adapt and personalise the components to the user's needs and the domain of the AI system.

CRediT authorship contribution statement

Marc Gallofré Ocaña: Conceptualization, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Andreas L. Opdahl:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix. Required qualities and their information sources

Table A.6
Functional required qualities and the projects that dealt with them.

	Annotating	Knowledge-representation	Enriching	Schema-evolution	Model-updating
PlanetOnto	✓	✓	✓	✓	✓
Neptuno	✓	✓	✓	✓	
Annoterra	✓	✓	✓	✓	
SemNews	✓	✓	✓	✓	
Hermes	✓	✓	✓	✓	
BBC CMS	✓	✓	✓		
NEWS	✓	✓	✓	✓	✓
Event Registry	✓	✓	✓		✓
NewsReader	✓	✓	✓	✓	
Reuters Tracer	✓	✓			✓
SUMMA	✓	✓	✓		
INJECT	✓	✓			
ASRAEL	✓	✓	✓		

	Storage	Push	Pull	Data-ownership	Privacy	Provenance
PlanetOnto	✓	✓	✓			
Neptuno	✓	✓	✓			
Annoterra	✓	✓	✓	✓		✓
SemNews	✓	✓	✓			
Hermes	✓		✓			
BBC CMS	✓		✓			
NEWS	✓	✓	✓	✓		✓
Event Registry	✓		✓			
NewsReader	✓		✓			✓
Reuters Tracer	✓	✓	✓			
SUMMA	✓	✓	✓			
INJECT	✓		✓			
ASRAEL	✓		✓			

Table A.7
Non-functional required qualities and the projects that dealt with them.

	Interoperability	Modularity	Scalability	Velocity
PlanetOnto	✓			
Neptuno	✓		✓	✓
Annoterra	✓	✓		
SemNews	✓			
Hermes				
BBC CMS	✓	✓	✓	
NEWS	✓	✓	✓	
Event Registry			✓	
NewsReader	✓	✓	✓	✓
Reuters Tracer			✓	✓
SUMMA	✓		✓	✓
INJECT				
ASRAEL	✓		✓	

	Variety	Knowledge-evolution	Sub-/symbolic-AI	Trustworthy-AI
PlanetOnto				
Neptuno	✓			
Annoterra	✓			
SemNews				
Hermes		✓		
BBC CMS	✓			✓
NEWS	✓	✓		
Event Registry		✓		
NewsReader	✓			
Reuters Tracer				
SUMMA	✓			
INJECT				
ASRAEL				

References

[1] C. Beckett, New Powers, New Responsibilities: a Global Survey of Journalism and Artificial Intelligence, Tech. Rep., Polis, London School of Economics and Political Science, 2019, URL https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities.

[2] J. Vázquez Herrero, S. Direito-Rebollal, A.S. Rodrí guez, X. García, Journalistic Metamorphosis: media Transformation in the Digital Age, Springer International publishing, 2020, http://dx.doi.org/10.1007/978-3-030-36315-4.

[3] 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, 2018, http://dx.doi.org/10.18653/v1/P18-4017.

[4] S.C. Lewis, O. Westlund, Big data and journalism, Digit. Journal. 3 (3) (2015) http://dx.doi.org/10.1080/21670811.2014.976418.

[5] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A.M. Vollmer, S. Wagner, Software engineering for AI-based systems: A survey, ACM Trans. Softw. Eng. Methodol. 31 (2) (2022) http://dx.doi.org/10.1145/3487043.

[6] A. d'Avila Garcez, L.C. Lamb, Neurosymbolic AI: The 3rd wave, 2020, arXiv:2012.05876.

[7] Aidan Hogan, et al., Knowledge graphs, ACM Comput. Surv. 54 (4) (2021) http://dx.doi.org/10.1145/3447772.

[8] M. Gallofré Ocaña, A.L. Opdahl, Supporting newsrooms with journalistic knowledge graph platforms: Current state and future directions, Technologies 10 (3) (2022) http://dx.doi.org/10.3390/technologies10030068.

[9] Y. Raimond, T. Scott, S. Oliver, P. Sinclair, M. Smethurst, Use of semantic web technologies on the BBC web sites, in: Linking Enterprise Data, 2010, http://dx.doi.org/10.1007/978-1-4419-7665-9_13.

[10] 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, 2019, http://dx.doi.org/10.1145/3308560.3316761.

[11] C. Bizer, T. Heath, T. Berners-Lee, Linked data: The story so far, in: Semantic Services, Interoperability and Web Applications: Emerging Concepts, IGI global, 2011, pp. 205–227.

[12] T. Al-Moslmi, M. Gallofré Ocaña, Lifting news into a Journalistic Knowledge Platform, in: Proceedings of the CIKM 2020 Workshops, 2020, URL http://ceur-ws.org/Vol-2699/paper42.pdf.

[13] N. Fernández, D. Fuentes, L. Sánchez, J.A. Fisteus, The NEWS ontology: Design and applications, Expert Syst. Appl. 37 (12) (2010) http://dx.doi.org/10.1016/j.eswa.2010.06.055.

[14] 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.

[15] 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, 2014, http://dx.doi.org/10.1145/2567948.2577024.

[16] P. Vossen, R. Agerri, 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, in: Special Issue Knowledge-Based Systems, Vol. 110, Elsevier, 2016, http://dx.doi.org/10.1016/j.knosys.2016.07.013.

[17] E. Motta, E. Daga, A.L. Opdahl, B. Tessem, Analysis and design of computational news angles, IEEE Access (2020).

[18] M. Gallofré Ocaña, T. Al-Moslmi, A.L. Opdahl, Data privacy in Journalistic Knowledge Platforms, in: Proceedings of the CIKM 2020 Workshops, 2020, URL http://ceur-ws.org/Vol-2699/paper44.pdf.

[19] S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren, D. Valerio, A software reference architecture for semantic-aware big data systems, Inf. Softw. Technol. 90 (2017) http://dx.doi.org/10.1016/j.infsof.2017.06.001.

[20] B. Sena, A.P. Allian, E.Y. Nakagawa, Characterizing big data software architectures: A systematic mapping study, in: Proceedings of the 11th Brazilian Symposium on Software Components, Architectures, and Reuse, SBCARS '17, Association for Computing Machinery, New York, NY, USA, 2017, http://dx.doi.org/10.1145/3132498.3132510.

[21] B. Sena, L. Garcés, A.P. Allian, E.Y. Nakagawa, Investigating the applicability of architectural patterns in big data systems, in: Proceedings of the 25th Conference on Pattern Languages of Programs, PLoP '18, The Hillside Group, USA, 2018.

[22] C. Avci, B. Tekinerdogan, I.N. Athanasiadis, Software architectures for big data: A systematic literature review, Big Data Anal. 5 (1) (2020) 1–53.

[23] P. Ataei, A.T. Litchfield, Big data reference architectures, a systematic literature review, in: ACIS 2020 Proceedings, (30) 2020, URL https://aisel.aisnet.org/acis2020/30.

[24] T.V.R. da Costa, E. Cavalcante, T. Batista, Big data software architectures: An updated review, in: O. Gervasi, B. Murgante, E.M.T. Hendrix, D. Taniar, B.O. Apduhan (Eds.), Computational Science and Its Applications – ICCSA 2022, Springer International Publishing, Cham, 2022, pp. 477–493.

[25] M. Gallofré Ocaña, A.L. Opdahl, Developing a software reference architecture for journalistic knowledge platforms, in: ECSA2021 Companion Volume, 2021.

[26] A.L. Opdahl, T. Al-Moslmi, D.-T. Dang-Nguyen, M. Gallofré Ocaña, B. Tessem, C. Veres, Semantic knowledge graphs for the news: A review, ACM Comput. Surv. (2022) http://dx.doi.org/10.1145/3543508.

[27] S. Angelov, P. Grefen, D. Greefhorst, A framework for analysis and design of software reference architectures, Inf. Softw. Technol. 54 (4) (2012) http://dx.doi.org/10.1016/j.infsof.2011.11.009.

[28] S. Angelov, J.J. Trienekens, P. Grefen, Towards a method for the evaluation of reference architectures: Experiences from a case, in: Software Architecture. ECSA 2008, 2008, http://dx.doi.org/10.1007/978-3-540-88030-1_17.

[29] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, 10.48550/ARXIV.1301.3781.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, (Long and Short Papers).

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[33] Y.A. Malkov, D.A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 824–836, http://dx.doi.org/10.1109/TPAMI.2018.2889473.

[34] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Trans. Big Data 7 (3) (2019) 535–547.

[35] M. Galster, P. Avgeriou, Empirically-grounded reference architectures: A proposal, in: Proceedings of the Joint ACM SIGSOFT Conference – QoSA and ACM SIGSOFT Symposium – ISARCS on Quality of Software Architectures – QoSA and Architecting Critical Systems – ISARCS, Association for Computing Machinery, 2011, http://dx.doi.org/10.1145/2000259.2000285.

[36] A. Berven, O.A. Christensen, S. Moldeklev, A.L. Opdahl, K.J. Villanger, A knowledge-graph platform for newsrooms, Comput. Ind. 123 (2020) http://dx.doi.org/10.1016/j.compind.2020.103321.

[37] 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, URL http://ceur-ws.org/Vol-2316/paper1.pdf.

[38] A. Tverberg, I. Agasøster, M. Grønbæck, R.S. Marius Monsen, K. Eikeland, E. Trondsen, L. Westvang, T.B. Knudsen, E. Fiskerud, R. Skår, S. Stoppel, A. Berven, G.S. Pedersen, P. Macklin, K. Cuomo, L. Vredenberg, K. Tolonen, A.L. Opdahl, B. Tessem, C. Veres, D.-T. Dang-Nguyen, E. Motta, V.J. Setty, WP3 2021 M3.1 Report the Industrial Expectations to, Needs from and Wishes for the Work Package, Tech. Rep., University of Bergen, MediaFutures, 2021.

[39] H.A. Simon, The Sciences of the Artificial, MIT Press, 1996.

[40] A. Hevner, S. Chatterjee, Design science research in information systems, 2010, http://dx.doi.org/10.1007/978-1-4419-5653-8_2.

[41] A.R. Hevner, A three cycle view of design science research, Scand. J. Inf. Syst. 19 (2) (2007) URL https://aisel.aisnet.org/sjis/vol19/iss2/4.

[42] A.R. Hevner, S.T. March, J. Park, S. Ram, Design science in information systems research, MIS Q. 28 (1) (2004) URL http://www.jstor.org/stable/25148625.

[43] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43, URL http://www.jstor.org/stable/26059207.

[44] N. Shadbolt, T. Berners-Lee, W. Hall, The semantic web revisited, IEEE Intell. Syst. 21 (3) (2006) 96–101, http://dx.doi.org/10.1109/MIS.2006.62.

[45] J. Domingue, E. Motta, PlanetOnto: From news publishing to integrated knowledge management support, IEEE Intell. Syst. Appl. 15 (3) (2000) 26–32, http://dx.doi.org/10.1109/5254.846282.

[46] Y. Kalfoglou, J. Domingue, E. Motta, M. Vargas-Vera, S. Buckingham Shum, Myplanet: An ontology driven web based personalised news service, in: Proceedings of International Joint Conference on Artificial Intelligence, Vol. 2001, International Joint Conferences on Artificial Intelligence, 2001, pp. 44–52, URL http://ceur-ws.org/Vol-47/kalfoglou.pdf.

[47] P. Castells, F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras, J. Lorés, Neptuno: Semantic web technologies for a digital newspaper archive, in: The Semantic Web: Research and Applications. ESWS 2004, 2004, http://dx.doi.org/10.1007/978-3-540-25956-5_31.

[48] D.B. Ramagem, B. Margerin, J. Kendall, AnnoTerra: Building an integrated earth science resource using semantic web technologies, IEEE Intell. Syst. 19 (3) (2004) http://dx.doi.org/10.1109/MIS.2004.3.

[49] A. Java, T. Finin, S. Nirenburg, SemNews: A semantic news framework, in: The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, 2006, URL https://www.aaai.org/Papers/AAAI/2006/AAAI06-316.pdf.

[50] K. Schouten, P. Ruijgrok, J. Borsje, F. Frasincar, L. Levering, F. Hogenboom, A semantic web-based approach for personalizing news, in: Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10, ACM Press, Sierre, Switzerland, 2010, p. 854, http://dx.doi.org/10.1145/1774088.1774264.

[51] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. Lee, Media meets semantic web – how the BBC uses DBpedia and linked data to make connections, in: The Semantic Web: Research and Applications, Vol. 5554, 2009, http://dx.doi.org/10.1007/978-3-642-02121-3_53.

[52] N. Fernández, J.M. Blázquez, J.A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, NEWS: Bringing semantic web technologies into news agencies, in: The Semantic Web - ISWC 2006, 2006, pp. 778–791, http://dx.doi.org/10.1007/11926078_56.

[53] 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, J. Web Semant. 37–38 (2016) 132–151, http://dx.doi.org/10.1016/j.websem.2015.12.004.

[54] Q. Li, S. Shah, X. Liu, A. Nourbakhsh, R. Fang, TweetSift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 2429–2432, http://dx.doi.org/10.1145/2983323.2983325.

[55] X. Liu, Q. Li, A. Nourbakhsh, R. Fang, M. Thomas, K. Anderson, R. Kociuba, M. Vedder, S. Pomerville, R. Wudali, R. Martin, J. Duprey, A. Vachher, W. Keenan, S. Shah, Reuters tracer: A large scale system of detecting & verifying real-time news events from Twitter, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 207–216, http://dx.doi.org/10.1145/2983323.2983363.

[56] X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin, J. Duprey, Reuters tracer: Toward automated news production using large scale social media data, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 1483–1493.

[57] S. 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, http://dx.doi.org/10.1145/3308558.3314135.

[58] N. Maiden, K. Zachos, A. Brown, G. Brock, L. Nyre, A. Nygård Tonheim, D. Apsotolou, J. Evans, Making the news: Digital creativity support for journalists, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–11, http://dx.doi.org/10.1145/3173574.3174049.

[59] F. Frasincar, J. Borsje, L. Levering, A semantic web-based approach for building personalized news services, Int. J. E-Business Res. (IJEBR) 5 (3) (2009) 35–53.

[60] D. Le-Phuoc, H.Q. Nguyen-Mau, J.X. Parreira, M. Hauswirth, A middleware framework for scalable management of linked streams, J. Web Semant. 16 (2012) http://dx.doi.org/10.1016/j.websem.2012.06.003, The Semantic Web Challenge 2011.

[61] M.A. Martínez-Prieto, C.E. Cuesta, M. Arias, J.D. Fernández, The SOLID architecture for real-time management of big semantic data, Future Gener. Comput. Syst. 47 (2015) http://dx.doi.org/10.1016/j.future.2014.10.016, Special Section: Advanced Architectures for the Future Generation of Software-Intensive Systems.

[62] N. Marz, How to beat the CAP theorem, 2011, URL http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html.

[63] P. P, Reference architecture and classification of technologies, products and services for big data systems, Big Data Res. 2 (4) (2015) http://dx.doi.org/10.1016/j.bdr.2015.01.001.

[64] D. Xu, D. Wu, X. Xu, L. Zhu, L. Bass, Making real time data analytics available as a service, in: Proceedings of the 11th International ACM SIGSOFT Conference on Quality of Software Architectures, QoSA '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 73–82, http://dx.doi.org/10.1145/2737182.2737186.

[65] G.M. Sang, L. Xu, P. de Vrieze, A reference architecture for big data systems, in: 2016 10th International Conference on Software, Knowledge, Information Management & Applications, SKIMA, 2016, pp. 370–375, http://dx.doi.org/10.1109/SKIMA.2016.7916249.

[66] L. Heilig, S. Voß, Managing cloud-based big data platforms: A reference architecture and cost perspective, in: Big Data Management, Springer International Publishing, Cham, 2017, pp. 29–45, http://dx.doi.org/10.1007/978-3-319-45498-6_2.

[67] S. Martínez-Fernández, C.P. Ayala, X. Franch, H.M. Marques, Benefits and drawbacks of software reference architectures: A case study, Inf. Softw. Technol. 88 (2017) http://dx.doi.org/10.1016/j.infsof.2017.03.011.

[68] M.K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence: Current trends, 2021, arXiv preprint arXiv:2105.05330.

[69] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Tech. Rep., European Commission, 2019, URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[70] N. Dragoni, S. Giallorenzo, A.L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, L. Safina, Microservices: Yesterday, today, and tomorrow, in: Present and Ulterior Software Engineering, Springer International publishing, 2017, http://dx.doi.org/10.1007/978-3-319-67425-4_12.

[71] R.C. Fernandez, P.R. Pietzuch, J. Kreps, N. Narkhede, J. Rao, J. Koshy, D. Lin, C. Riccomini, G. Wang, Liquid: Unifying nearline and offline big data integration, in: Proceedings of the 7th Biennial Conference on Innovative Data Systems Research, CIDR, 2015.

[72] J. Kreps, Questioning the lambda architecture, 2014, URL https://www.oreilly.com/radar/questioning-the-lambda-architecture.

[73] F. Cerezo, C.E. Cuesta, J.C. Moreno-Herranz, B. Vela, Deconstructing the Lambda architecture: An experience report, in: 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), 2019, http://dx.doi.org/10.1109/ICSA-C.2019.00042.

[74] H.P. Nii, The blackboard model of problem solving and the evolution of blackboard architectures, AI Mag. 7 (2) (1986) 38.

[75] I.D. Craig, Blackboard systems, Artif. Intell. Rev. 2 (2) (1988) 103–118.

[76] F. Lecue, On the role of knowledge graphs in explainable AI, Semantic Web 11 (1) (2020) 41–51.

[77] A.L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, Softw. Syst. Model. (2020) http://dx.doi.org/10.1007/s10270-020-00801-w.

[78] S. Hellmann, J. Lehmann, S. Auer, M. Brümmer, Integrating NLP using linked data, in: The Semantic Web – ISWC 2013, 2013, http://dx.doi.org/10.1007/978-3-642-41338-4_7.

[79] A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W.R. Van Hage, P. Vossen, NAF and GAF: Linking linguistic annotations, in: Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, 2014.

[80] H. Dibowski, S. Schmid, Using knowledge graphs to manage a data lake, in: INFORMATIK 2020, Gesellschaft für Informatik, Bonn, 2021, pp. 41–50, http://dx.doi.org/10.18420/inf2020_02.

[81] W3C, Largetriplestores, 2020, URL https://www.w3.org/wiki/LargeTripleStores.

[82] B. Tessem, M. Gallofré Ocaña, A.L. Opdahl, Construction of a relevance knowledge graph with application to the LOCAL news angle, in: Nordic Artificial Intelligence Research and Development, 2023.

[83] B. Tessem, Analogical news angles from text similarity, in: Artificial Intelligence XXXVI, 2019, http://dx.doi.org/10.1007/978-3-030-34885-4_35.

[84] M. Gallofré Ocaña, A.L. Opdahl, D.-T. Dang-Nguyen, Emerging News task: Detecting emerging events from social media and news feeds, MediaEval, 2021.

Additional Publications

Publication IV

Data Privacy in Journalistic Knowledge Platforms

Marc Gallofré Ocaña, Tareq Al-Moslmi and Andreas L. Opdahl
29th ACM International Conference on Information and Knowledge Management (CIKM 2020), *Proceedings of the CIKM 2020 Workshops.*

Data Privacy in Journalistic Knowledge Platforms

Marc Gallofré Ocaña[a], Tareq Al-Moslmi[a] and Andreas L. Opdahl[a]

[a]*University of Bergen, Fosswinckelsgt. 6, Postboks 7802, 5020 Bergen, Norway*

Abstract

Journalistic knowledge platforms (JKPs) leverage data from the news, social media and other sources. They collect large amounts of data and attempt to extract potentially news-relevant information for news production. At the same time, by harvesting and recombining big data, they can challenge data privacy ethically and legally. Knowledge graphs offer new possibilities for representing information in JKPs, but their power also amplifies long-standing privacy concerns. This paper studies the implications of data privacy policies for JKPs. To do so, we have reviewed the GDPR and identified different areas where it potentially conflicts with JKPs.

Keywords

Privacy, Personal data, Journalistic Knowledge Platforms, GDPR

1. Introduction

Journalistic Knowledge Platforms (JKPs) are an emerging generation of platforms which combine state-of-the-art artificial intelligence (AI) techniques, like knowledge graphs and natural-language processing (NLP) [1, 2] for transforming newsrooms and leveraging information technologies to increase the quality and lower the cost of news production. JKPs exploit and combine news, social media and other information sources, using linked open data (LOD), digital encyclopaedic sources and news archives to construct knowledge graphs and provide fresh and unexpected information to journalists, helping them dive more deeply into information, events and story-lines [3]. JKPs of various kinds are becoming increasingly important in leading news agencies like BBC [4] and Thomson Reuters [5].

However, obtaining and representing knowledge leads to data privacy concerns when personal data from different sources is neither collected directly from the subject nor with the subject's consent, although some countries have exemptions that loosen privacy requirements for journalistic research that is in the public interest or does not identify individuals directly. This exemption becomes even more complex when the national privacy policies that apply to the data sources and the JKP are distinct or the public interest is not crystal clear.

Data privacy has become a central topic of discussion for organisations and projects from private companies and governments to research activities in universities around the globe. Whereas there is no general solution to privacy for everyone and specific solutions vary between different countries, cultures and organisations, privacy is a common concern, which has been discussed from the ethical and philosophical points of view by many different authors [6, 7] and organisations like the European Commission [8, 9]. The EU has established the General Data Protection Regulation (GDPR) which sets up a framework for governing the usage, processing, privacy and security of personal data, granting individuals power over their data and making organisations responsible for data collection and usage practices.

Our group have been developing News Hunter [10, 11, 12], a series of JKP architectures and prototypes. The current News Hunter platform is big-data ready and designed to continually harvest and monitor real-time news feeds (e.g., RSS or web-sites) and social media (e.g., Twitter and Facebook). It aims to analyse and represent news content semantically in knowledge graphs in order to provide better background information for journalists and to suggest news angles [13, 14, 15, 16, 17].

As part of our News Hunter effort, this paper investigates the implications of the GDPR on JKPs. To do so, we asked ourselves which data privacy conflicts can arise when JKPs when are used in journalistic work, in particular when that work may be exempted from some privacy regulations because it is in the public interest. To the best of our knowledge, there is no previous work discussing the possible data privacy conflicts in JKPs. Our contributions are: (1) we review different journalistic scenarios and personal data sources that

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

EMAIL: Marc.Gallofre@uib.no (M. Gallofré Ocaña); Tareq.Al-Moslmi@uib.no (T. Al-Moslmi); Andreas.Opdahl@uib.no (A.L. Opdahl)
ORCID: 0000-0001-7637-3303 (M. Gallofré Ocaña); 0000-0002-5296-2709 (T. Al-Moslmi); 0000-0002-3141-1385 (A.L. Opdahl)

CEUR Workshop Proceedings (CEUR-WS.org)

can conflict with GDPR policies, and (2) we introduce a personal data matrix framework to classify personal data conflicts and discuss the possible uses of this matrix.

This paper is organised as follows: section 2 defines the main privacy concepts, section 3 discusses potential data privacy conflicts in JKPs, section 4 introduces the personal data matrix framework, section 5 summarises the conclusions, and section 6 presents open questions and future work.

2. Background

2.1. Journalistic knowledge platforms

Journalistic Knowledge Platforms (JKPs) leverage and combine news, multimedia content (e.g., TV news channels and podcast) social media (e.g., Twitter and Facebook), web-blogs and information over the net, using linked open data (LOD), digital encyclopaedic sources (e.g., Wikipedia and Wikidata) and news archives to provide fresh and unexpected information to journalists. Projects like Neptuno [18], Event Registry [19], NEWS [20], NewsReader [21], SUMMA [22] and News Angler [11, 16, 12] have presented examples of JKPs.

A typical JKP comprises a knowledge graph [23, 24, 25] along with AI, NLP pipelines, and semantic technology components. In a JKP, the Knowledge graph is filled with potentially news-related histories, information and current and archival news to support journalists in creating newsworthy stories, finding relevant information, events and story-lines, and validating and verifying news. The information in the knowledge graph is represented using standard identifiers and semantic knowledge representations with reasoning capabilities. The usage of standard identifiers facilitates data integration, which is the process of joining and merging different data sets or public data sources like Semantic Web [26, 27], linked open data (LOD) [28], including Wikidata and DBpedia, and Wikipedia. Data integration together with reasoning allows drawing new insights from information from across the data that would be impossible before with isolated datasets. This inherent ability of drawing new insights implies that new personal data may be derived and exposed in the knowledge graph.

2.2. Privacy

Privacy is a historically and culturally situated concept. For example, whereas privacy in Europe is traditionally considered as an inalienable basic right of an autonomous person that states must protect to preserve a democratic society, the concept of privacy in the United States of America is understood primarily as a physical notion that implies the "private space" (e.g., bedroom, bathroom or the entire home) [6]. These differences are reflected in the data privacy regulations of the EU and the USA. The EU states in the *General Data Protection Regulation* (GDPR) [29] that individuals must be notified and have the right to consent when their personal data is collected from either inside or outside EU legislation. In contrast, the US only regulates privacy issues regarding health matters and some financial information, leaving the rest to individual states or businesses which do not need to ask for individuals consent and give the possibility to individuals to resign if they have any reservations about what is being collected from them. On a global scale beyond EU and the USA, differences in how privacy is viewed are even bigger, making it even more challenging to handle privacy regulations when JKPs are used in fully international news organisations that operate across cultural and legal domains.

2.3. GDPR

All actions using or processing personal data of data subjects who are in the European Union shall obey the General Data Protection Regulation (GDPR) [29]. The GDPR is an extensive regulation which sets the basis for dealing with personal data in the EU or using personal data from the EU. This section highlights the most general concepts that restrict what and how to process personal data in JKPs.

The GDPR defines the concepts of *personal data* and *processing* (*Chapter I, Article 4*). *Personal data* is any information that can be employed to identify directly or indirectly a natural person (e.g., name, an identification number or online identifier) or *sensitive data* like health, biometric, genetic, economic, cultural factors or political opinions of a natural person. Data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person is considered as personal data too. By *processing*, the GDPR means processes such as collection, structuring, storage, alteration, consultation, use, disclosure, combination, restriction, erasure or destruction of personal data.

Moreover, the GDPR establishes a set of principles for processing personal data (*Chapter II*) which define how data have to be processed, stored and maintained. These set of principles establish that data shall be processed within the initial purposes and purposes compatible with them (*purpose limitation*), only what is necessary to the purpose (*data minimisation*), personal data shall be accurate and kept up to date (*accuracy*), and stored for no longer periods than the necessary for the purpose (*storage limitation*). It also defines the *lawfulness of processing* which determines when personal data can be processed, e.g., when data subject gives the consent or for a task carried out in public interest. Under the GDPR, some research by journalist and academics is understood as public interest. Likewise, the GDPR limits the processing of sensitive data which is prohibited in general terms but with some exceptions, e.g., when data subject gives explicit consent, it is necessary for reasons of substantial public interest, or the data subject has manifestly made it public.

The GDPR also details when and which information have to be provided to the data subjects (*Chapter III*). In the case of personal data that is not obtained directly from the subject, it determines which data have to be provided, e.g., the source of the personal data and whether it came from publicly accessible sources or the categories of personal data. Nevertheless, it also establishes some exemptions, e.g., when the provision of such information proves to be impossible or is likely to harm the objectives of the processing objective.

3. Privacy conflicts in JKPs

When discussing which scenarios in JKPs can cause a conflict with the GDPR we must consider the source of the personal data, distinguishing between the data gathered directly from the subject, the data harvested from other sources like news or social media and the inferred data.

In the context of GDPR, some data processing by journalists is exempted when it is conducted in the public interest. However, this exemption exclusively applies to journalistically relevant (newsworthy) personal data, not to any personal data processed in the JKP, and sensitive data may be less exempted or not exempted at all. Therefore, we must also consider how relevant the personal data is for the public interest from a news perspective. This includes the assessment of newsworthiness [30] along with the type of news. E.g., a corruption scandal and a private event in the life of a famous person may both be highly newsworthy, but corruption is most likely more important for the public interest.

3.1. Personal data from the subject

When personal data comes directly from a subject and is collected with the subject's explicit consent (e.g., personal data collected during an interviewed), it does not present a problem with the GDPR. However, the data can be made it publicly accessible by the subject itself in social media networks (e.g., posts like tweets in Twitter or forums and groups like Facebook groups) or in the subject's verifiable social media accounts and personal web sites without providing explicit consent for its collection to the JKP. In that case, apart from having to follow the source's data policies, it raises the ethical questions whether the consent is implicit because it is publicly available, when we should consider that it is publicly available and under which conditions.

3.2. Personal data from third parties

When the data is not collected directly from the subject, instead, it has been made accessible by a third party and subject may ignore its existence, we have to consider two possible scenarios:

The first scenario, when news-related information is gathered from the web (e.g., online news, RSS, websites or social media), JKPs can extract personal data from the content to represent and combine it in the knowledge graphs. E.g., from *"We know the classic 7-layer dip, made with Bush's Beans, is a fan favorite for game day snacking celebrations, Kate Rafferty, the consumer experience manager for Bush's, told Fox News."*[1], we can extract information like "Kate Rafferty is a person who works as consumer experience manager at Bush's Beens company at Knoxville, Tennessee" which can be considered as personal data, as it can be used to identify a natural person. According to the GDPR, the subject should be notified, and the JKP has to provide a mechanism for the subject to protest. Even though, on a large scale, two issues arise: the number of notifications that famous people will get and how to contact subjects if the content information is missing.

The second scenario, when personal data is gathered from publicly available sources or open sources like Wikipedia, Wikidata or telephone/address books, it is clear that the personal data is already public. However, it may not be released with subject's consent. In that case, it opens the question about: why should the JKP not be allowed to store copies of personal data which is already public?

3.3. Inferred personal data

When personal data is not gathered from any source, instead, it is inferred using the actual data (either from the data subject, collected from news or gathered from

[1] Drew Schwartz, VICE: This 70-Layer Bean Dip Is the Most Vile Thing I've Ever Seen (https://t.co/qKyyNevpBh)

public sources) and reasoning techniques. E.g., from the text *"The European Court of Justice (ECJ) said that Oriol Junqueras had become an MEP the moment he was elected in May, despite being on trial for sedition."*[2], we can represent the person "Oriol Junqueras" as the entity Q116812 from Wikidata, from which we can derive that he is a member of a political party (P102) and the political party is "the Republican Left of Catalonia" (Q150068). With this information is it possible to infer the subject's political ideology (P1142) from the political party information such as "republicanism" (Q877848) and "Catalan pro-independence movement" (Q893331). In this scenario there is not a direct source of subject's personal data or political opinions, instead, there is a source of related information used for inferring knowledge which can be either in the same knowledge graph or from external sources.

3.4. Possible solutions

To comply with the GDPR's *Chapter III*, in any of the previously discussed scenarios it is important to identify the data source and personal data category (e.g., name, ID number, online identifier, health data, political opinion). Thus, it will be possible to identify both the source and data and take actions accordingly. Although the main responsible of complying with the GDRP in the first place is the data provider (i.e., news website, social media platform or telephone/address books), JKPs should follow the GDPR to safeguard the subjects of privacy and consider the policies and restrictions established by the data provider. The JKP must always take independent responsibility for privacy, and it cannot trust its sources to safeguard privacy. In a truly international and global set-up, where different privacy policies apply, JKPs may have to be designed with different knowledge graphs for different legal domains or geographical regions, each graph only being accessible from its own privacy domain. When this is infeasible, the most restrictive policies to guarantee personal data privacy must be adopted.

Moreover, JKPs should also implement automatic mechanisms to notify subjects with both the personal data and the sources when this information is identified, a process that can be done by email. It is also possible to set up an automatic system for subjects to protest, complain, request or ask about personal data.

Table 1
Personal data matrix

	Consented	Collected	Inferred
Impersonal Data	✔	✔	✔
Personal Data	✔	!	!
Sensitive Data	✔	!!	!!

4. Personal data matrix

After reviewing the previous scenarios, we classified the different situations that can cause a conflict with the GDPR into a two-dimension matrix (figure 1) framework. The personal data matrix aims to help journalists and JKP developers to classify the personal data in JKPs and its possible issues with privacy policies.

The personal data matrix (figure 1) classifies personal data based on the privacy level and the data source. The first dimension (privacy level) classifies the data whenever it does not represent personal data (impersonal data), it represents personal data or it represents "sensitive data". There is an explicit distinction between personal and "sensitive data" because in the GDPR "sensitive data" have much more restricted limitations. The other dimension, the data source dimension, classifies data based on the data collected with the subject's consent (consented), the data directly collected from the content and the inferred data. Only when data is either explicitly consented or it is not personal data its treatment is straightforward. Otherwise, as discussed in the previous section, each of these combinations has its issues and open questions regarding the application of the GDPR and its origin.

The data matrix can be also regarded as a cube, where the public interest represents the third dimension. This third dimension determines to what extent the GDRP exemption to data processing for journalistic purposes in the public interest applies, taking into consideration the newsworthiness component of the data.

The proposed matrix can be used by JKP researchers and developers to ensure – as automatically as possible, but in practice aided by human data privacy stewards – that privacy regulations are never violated. The matrix should be used in the design of JKPs to ensure that personal data is protected by default. E.g., developers of JKPs can use the matrix to evaluate the system and identify which processes or collected data can lead to privacy conflicts; implement the matrix as part of the news creation workflow so that journalists can automatically check data privacy compliance before collecting, re-combining or using any personal data;

[2] BBC: Jailed Catalan leader 'should have had immunity', rules EU court (https://www.bbc.com/news/world-europe-50808766)

it can be utilized as metadata for each piece of data in the knowledge graph to automatise its recognition and privacy assurance; and the matrix can be used when dealing with data under different regulations to find divergences between them.

5. Conclusion

JKPs need to deal with personal data which in many cases will be integrated into knowledge graphs without the explicit consent from the subject. Thus, JKPs need to safeguard data privacy. For that reason, we have presented a framework for classifying personal data in journalistic knowledge graphs and identified different scenarios and personal data sources that potentially can conflict with the GDPR. We believe the identified scenarios, sources and presented matrix will be helpful as a reference for related projects and similar domains.

6. Future work

We want to continue exploring the open questions highlighted in our discussions in section 3, as well as questions such as how to deal with different privacy regulations that may apply in international settings, how to represent and effectively use GDPR in JKP processes, and how to deal with personal data about children. Data linking transparency is another open question which would help to identify situations that conflict with privacy and identify which data can be stored and which data cannot be stored in JKPs according to the GDPR and other privacy policies and regulations. Besides that, as anonymisation, encryption and blockchain technologies are presented as potential solutions to safeguard privacy and control copyrights and data access, we want to research how effective they are in the context of JKPs and how they can benefit JKPs.

Apart from that, one critical aspect when dealing with data from external sources, which has not been considered in this work, is the copyright and intellectual property regulations which have a direct relation with the data that can be processed and stored. In this context, we want to explore how to effectively manage them in JKPs (e.g., using ontologies).

Acknowledgments

Supported by the Norwegian Research Council IKT-PLUSS project 275872 *News Angler*, which is a collaboration with Wolftech AB, Bergen, Norway.

References

[1] 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.

[2] T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl, C. Veres, Named entity extraction for knowledge graphs: A literature overview, IEEE Access 8 (2020) 32862–32881.

[3] M. Gallofré Ocaña, A. L. Opdahl, Challenges and opportunities for journalistic knowledge platforms, in: Proceedings of the CIKM 2020 Workshops, Galway, Ireland, 2020. To appear.

[4] Y. Raimond, M. Smethurst, A. McParland, C. Lowis, Using the past to explain the present: Interlinking current affairs with archives via the semantic web, in: H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty, K. Janowicz (Eds.), The Semantic Web – ISWC 2013, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 146–161. doi:10.1007/978-3-642-41338-4_10.

[5] B. Ulicny, Constructing knowledge graphs with trust, in: METHOD 2015: The 4th International Workshop on Methods for Establishing Trust of (Open) Data, Bethlehem, PA, 2015.

[6] C. Ess, Digital Media Ethics, Polity, 2014.

[7] D. G. Johnson, Computer Ethics, Prentice Hall, 2001.

[8] European Parliament, Regulation (EU) no 1291/2013 of the european parliament and of the council of 11 december 2013 establishing horizon 2020 - the framework programme for research and innovation (2014-2020) and repealing decision no 1982/2006/ECText with EEA relevance (2013).

[9] European Commission, Policy | science with and for society - research and innovation - european commission, 2019. URL: http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=ethics.

[10] 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.

[11] 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.

[12] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, A knowledge graph platform for newsrooms, Computers in Industry (2020). To appear.

[13] 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.

[14] 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.

[15] 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.

[16] A. L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, Software and Systems Modeling (2020) 1–17.

[17] E. Motta, E. Daga, A. L. Opdahl, B. Tessem, Analysis and design of computational news angles, IEEE Access (2020).

[18] P. Castells, F. Perdrix, E. Pulido, R. Mariano, R. Benjamins, J. Contreras, J. Lorés, Neptuno: Semantic web technologies for a digital newspaper archive, in: European Semantic Web Symposium, Springer, Berlin, Heidelberg, 2004, pp. 445–458.

[19] 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, ACM, New York, NY, USA, 2014, pp. 107–110. doi:10.1145/2567948.2577024.

[20] N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, News: Bringing semantic web technologies into news agencies, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), The Semantic Web - ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 778–791.

[21] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. Palmero 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, Knowledge-Based Systems 110 (2016). doi:10.1016/j.knosys.2016.07.013.

[22] U. Germann, P. v. d. Kreeft, G. Barzdins, A. Birch, The summa platform: Scalable understanding of multilingual media, in: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018.

[23] A. Singhal, Introducing the knowledge graph: things, not strings, 2012. URL: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[24] Ehrlinger, Lisa and Wöß, Wolfram, Towards a definition of knowledge graphs, in: Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), 2016, p. 4. URL: http://ceur-ws.org/Vol-1695/paper4.pdf.

[25] J. Yan, C. Wang, W. Cheng, M. Gao, A. Zhou, A retrospective of knowledge graphs, Frontiers of Computer Science (2016). doi:10.1007/s11704-016-5228-9.

[26] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Scientific American 284 (2001).

[27] N. Shadbolt, T. Berners-Lee, W. Hall, The semantic web revisited, IEEE Intell. Syst. 21 (2006) 96–101.

[28] C. Bizer, T. Heath, T. Berners-Lee, Linked data: The story so far, in: Semantic services, interoperability and web applications: emerging concepts, IGI Global, 2011, pp. 205–227.

[29] The European Parliament and The Council of the European Union, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), Official Journal of the European Union (2016). URL: http://data.europa.eu/eli/reg/2016/679/oj.

[30] 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.

Publication V

Lifting News into a Journalistic Knowledge Platform

Tareq Al-Moslmi and Marc Gallofré Ocaña
29th ACM International Conference on Information and Knowledge Management (CIKM 2020), *Proceedings of the CIKM 2020 Workshops.*

Lifting News into a Journalistic Knowledge Platform

Tareq Al-Moslmi[a], Marc Gallofré Ocaña[a]

[a] *University of Bergen, Fosswinckelsgt. 6, Postboks 7802, 5020 Bergen, Norway*

Abstract

A massive amount of news is being shared online by individuals and news agencies, making it difficult to take advantage of these news and analyse them in traditional ways. In view of this, there is an urgent need to use recent technologies to analyse all news relevant information that is being shared in natural language and convert it into forms that can be more easily and precisely processed by computers. Knowledge Graphs (KGs) offer offer a good solution for such processing. Natural Language Processing (NLP) offers the possibility for mining and lifting natural language texts to knowledge graphs allowing to exploit its semantic capabilities, facilitating new possibilities for news analysis and understanding. However, the current available techniques are still away from perfect. Many approaches and frameworks have been proposed to track and analyse news in the last few years. The shortcomings of those systems are that they are static and not updateable, are not designed for large-scale data volumes, did not support real-time processing, dealt with limited data resources, used traditional lifting pipelines and supported limited tasks, or have neglected the use of knowledge graphs to represent news into a computer-processable form. Therefore, there is a need to better support lifting natural language into a KG. With the continuous development of NLP techniques, the design of new dynamic NLP lifters that can cope with all the previous shortcomings is required. This paper introduces a general NLP lifting architecture for automatically lifting and processing news reports in real-time based on the recent development of the NLP methods.

Keywords

Natural language processing (NLP), Journalistic knowledge platforms, Knowledge Graphs, Computational journalism, Stream data processing, Semantic technologies, Big data

1. Introduction

For several years we have seen how the traditional news press has moved to online content and new online press has appeared, publishing more online content than ever. Social networks enhanced that phenomenon facilitating real-time interactions and sharing, allowing pre-news to come to the surface, and bringing users with newer ways to digest news. Analysing news in real-time for supporting journalist work requires lifting those news to machine-understandable formats. Semantic representation of news using knowledge graphs is one of such formats that could be employed. Since news texts are expressed as natural language, there is a crucial need for processing and lifting these texts into a knowledge graph.

This paper presents an NLP lifting architecture component of the Journalistic Knowledge Platforms (JKP) for lifting natural language news text into knowledge graphs. JKP is a system intended for analysing, lifting, and representing news using knowledge graphs to support journalists exploiting knowledge from and about news being shared on the web and social media networks. JKPs have become crucial for press industry. Yet, many works have proposed to process the news texts in many different ways in order to apply different JKP processes.

Our group have been developing a series of JKP prototypes called News Hunter [1, 2, 3] in collaboration with a developer of newsroom tools for the international market. News Hunter moves forward the JKP to address the journalistic needs proposing a system to harvest real-time news stories from RSS feeds and social media, lifting news using SOTA approaches, and representing stories into knowledge graphs using Semantic Web standard technologies, Linked Open Data and NIF formats. News Hunter also explores detection and suggestion of news angles and exploitation of Semantic Web to support journalistic work [4, 5, 6, 7, 8].

Differently from previous works, our introduced NLP subsystem's architecture for News Hunter aims to lift all processed news into a semantic knowledge graph in real-time. Moreover, two Natural Language Processing (NLP) lifting tracks could be chosen: the traditional pipeline and the end-to-end which follows the state-of-the-art (SOTA) development of deep neural network. That would avoid some limitations reported in previous lifting tasks [9, 10].

The rest of the paper is organised as follows: Section 2 presents the background for our work. Section 3 introduced the general architecture of JKP. Section 4 constitutes the bulk of the paper and introduces the

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

EMAIL: Tareq.Al-Moslmi@uib.no (T. Al-Moslmi);
Marc.Gallofre@uib.no (M. Gallofré Ocaña)
ORCID: 0000-0002-5296-2709 (T. Al-Moslmi); 0000-0001-7637-3303
(M. Gallofré Ocaña)

general NLP lifting process for real-time news lifting to a knowledge graph. Section 5 concludes the paper and outlines plans for future work.

2. Related Work

Current JKPs [11, 12, 13, 14, 15, 16] deal with big data multilingual text and multimedia sources of news-related items from which they have implemented their different NLP pipelines. These JPKs implemented NLP pipelines for lifting news into knowledge graphs and detect events normally by using traditional Named Entity Recognition (NER) and Named Entity Linking (NEL) systems, and pre-processed news text using linguistic techniques such as Part-of-Speech tagging (PoS), tokenisation, lemmatization and translation. In addition, NEWS project [11] used pattern matching to detect events, implemented NEL using PageRank and classified items, concepts and events using IPTC codes. NewsReader [13] used DBpedia Spotlight for NEL and mined opinion, causal, factual, temporal and semantic role information from news. ASRAEL [16] used SpaCy for NER, ADEL for NEL and Wikidata for linking events. SUMMA [15] used support vector machines (SVM) for NEL and classified topics from news. And both EvenRegistry [12] and SUMMA [15] used clustering techniques to detect events.

The NEWS project [17, 11] aimed to provide fresh multilingual information to news agencies (Spanish EFE and Italian ANSA agencies) analysing both textual and multimedia items. NEWS uses Ontology Ltd. (currently part of EXFO Nova Context real-time active topology platform[1] to implement the NLP pipeline to provide item categorization, concept representation, abstract generation, event recognition and NER using the ITPC codes. The NLP pipeline combines both linguistic techniques (patterns and rules such as PoS tagging) and traditional NER and NEL techniques (statistical techniques and PageRank). For recognizing events, NEWS project used pattern recognition techniques to describe and find the desired events.

The process of recognizing events is a relevant feature of such systems, which is approached in many different ways. For example, Event Registry [12] uses clustering algorithms to detect and group similar articles which represent the same event. Following the central idea of events, NewsReader project [13, 18] proposed a method, tools and a system to automatically leverage and represent events from news.

The NewsReader NLP pipeline performs language specific NER and NEL, event and semantic role detection and temporal relation detection over four different languages dealing and millions of news articles. The NLP pipeline processes each item starting with linguistic techniques (tokenizer, PoS, multiwords tagger), traditional NER and NEL (based on DBpedia Spotlight), opinion miner, semantic role labeler, event resolution, temporal recognizer and causal and factuality relation extraction. To overcome the large amount of news articles, NewsReader implemented its NLP pipeline using Big Data oriented technologies (i.e., Hadoop and Storm) into an scalable and real-time system [14].

Big data, multimedia and multilingual sources together are encountered in SUMMA project [15] which is an open-source platform for automated, scalable and distributed monitoring of real-time media broadcasts to support news agencies work like BBC or Deutsche Welle. The platform is built using big data-oriented technologies and services running in Docker[2] containers. SUMMA converts multimedia sources into text which is translated into English when found in other languages. Then, the text is processed through a NLP pipeline which classify them by topic using a hierarchical attention model, cluster them into storylines using clustering algorithms, and represent them using traditional NER (dependency parsing) and NEL (SVM-Ranking) techniques.

Likewise, the previous works ASRAEL project [16] uses knowledge graphs to represent events in news articles for searching purposes. To do so, they map AFP articles to Wikidata using NER (based on spaCy) and the NEL system ADEL.

As observed in the previous works there is a need for big data, real-time and semantic technologies approaches to deal with high volumes of news items that comes from multilingual and multimedia sources, and a common interest for detecting events among journalists and the different projects. Moreover, the proposed NLP techniques follow traditional approaches and similar pipelines which may not be always suitable for big data and real-time or for providing the best results.

Many approaches for lifting natural language to knowledge graphs are based on previous-generation NER techniques, and new lifting approaches that add disambiguation and linking to recent best-of-breed NE recognisers are needed . There is also a lack of standards for comparing lifting approaches[10]. This can partly be attributed to a lack of commonly accepted benchmarks, but it also a consequence of the recognition-disambiguation-linking pipeline. For ex-

[1] https://www.exfo.com/en/ontology/

[2] www.docker.com

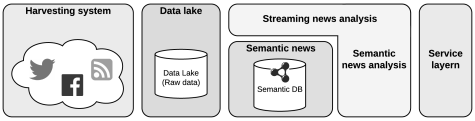


Figure 1: News Hunter architecture [2]

ample, it is hard to fairly compare pure NER with combined NER-NED-NEL techniques, when the latter is restricted to identifying named entities in the KB that is used for disambiguation and linking. Moreover, traditional sequential steps are now being integrated by joint learning or end-to-end processes. Consequently, mentions and entities that were previously analysed in isolation are now being lifted in each other's context. The current culmination of these trends are the deep-learning approaches that reported promising results recently. Most of those developments are not considered in previous works and this paper targets to cope with these gaps.

3. Journalistic Knowledge Platform architecture

In our previous work on News Hunter[2] we have proposed a general architecture for journalistic knowledge platforms (Figure 1) which is intended for big data real-time news lifting and processing. The still evolving architecture consists of 5 main parts: (1) The harvesting system which harvests the news from the web (e.g., RSS feeds, Facebook, Twitter) or daily produced in-house news (e.g., agency daily news activity) and its associated metadata (e.g., URL, source, author, ID, timestamp), and represents them using JSON in order to facilitate its parsing, transferring and simplifying it further processing. (2) The data lake or storage system for big data and real-time which is designed for sharing the news items across the different processes. (3) The semantic news component which contains the NLP lifter and the semantic DB (knowledge graph). (4) The semantic and streaming news analysis services, which due to the importance of social media can provide real-time analysis like trend monitoring, and event detection. (5) The service layer which allows users interact with the JKP.

News items can be collected from multiple sources: online news (e.g., RSS feeds), social media (e.g., Facebook, Twitter), archives or daily produced in-house news (e.g., agency daily news activity). The news crawler is oriented to harvest news from any source or multiple sources of interest. Due to the high amount of news items and their velocity of production, the harvested items are represented using standard lightweight formats like JSON, in order to facilitate its parsing, execution, transfer, sharing between components and temporal storage. News items are gathered together with its associated metadata (e.g., URL, source, author, ID, timestamp) which is included in the JSON files to benefit, speed and simplify its further processing and NLP tasks.

News items are processed according to their source: social media or news agencies . The news histories coming from news agencies (RSS feeds, news websties or archive) in JSON format are lifted into the knowledge graph as RDF triples using the NLP lifter, which can be adapted to the domain specific of the news history (e.g., economics, politics, sports). On the other hand, the news items coming from social media can be either pre-news (i.e., real-time information about events or something that is happening at the moment but not yet or incomplete as news histories) or small summaries/abstracts about news. Thus, identifying the topic they are related to and cluster them into groups of pre-news items that represent the same event and topic facilitates its processing. As these clusters of pre-news items represent a potential event with richer information that a single one item, they can be lifted using NLP techniques into the Knowledge Graphs.

Furthermore, as the social media items are potential real-time pre-news or events which can be breaking news, they are of highly importance for journalists. Yet, the clusters are analysed and monitored in order to find trends or breaking news events, that are reported in real-time to journalists.

In this paper, we are introducing the NLP lifting architecture that received the input from the harvester that have been explained previously[2]. The harvester is taking care of getting the data from different sources and standardise the data type into a unified format like JSON, XML, or NIF. The text can be stored in a big-data oriented databases such as Apache Cassandra [3] or HBase [4], which are oriented for distribution and large-scale processing pipelines. Moreover, the text can be distributed along the different NLP tasks using API or distribution framework like Kafka [5] or RabbitMQ [6]. The NLP liifter then has to deal with the data and lift it into a proper semantic format that will then be inserted to the KG.

[3] https://cassandra.apache.org
[4] https://hbase.apache.org
[5] https://kafka.apache.org
[6] https://www.rabbitmq.com

4. NLP lifter

This section describes the NLP lifter for news natural language texts to knowledge graphs. The NLP lifter which is a component of the JKP architecture consists of the main NLP lifting tasks as well as some additional related tasks. Differently from others proposed systems, our proposed NLP lifter is docker-based and contains the most possible tasks (traditional and recently developed ones) as shown in Figure 2. This allow the development of the platform and ensure using the most recent technology all the time. There will be two main NLP tracks: the traditional pipeline that is updated by recent technologies and the end-to-end track which is the SOTA in many tasks. In addition, there is the ensemble service that could combine more than one lifter to produce better results. The purpose and advantage of this is that the user can choose to use the most suitable track for his case and data as well as the most recent techniques. In the traditional pipeline the tasks like NER, NED, and NEL are implemented separately and mostly using the off-the-shelf software. The off-the-shelf systems are usually based on old approaches and their performance is not the SOTA. Moreover, traditional lifting methods neglect the relations between entity types and entity context. However, there will be a possibility in our introduced architecture to ensure the using of the most updated ones or using newest systems by just replacing or adding their dockers to the related component. The news item annotation ontology that has already been designed by[7] defines how the semantic annotations of news items should be represented in the knowledge graph. Each harvested news item is associated with one or more annotations, which may be, for example, named entities, concepts, topics, times or geolocations or relations between annotations. The ontology also describes how the sources of news items and annotations are represented in the knowledge graph to maintain provenance[7]. We describe the general NLP lifter components as the following:

4.1. Pre-processing

The quality of the data plays a key role in determining the suitable pre-processing techniques. Since we are dealing with the real-time streaming, the cleaning and normalization are required to remove unnecessary or noisy terms (like ASCII codes, currency symbols, hashtags, and so forth). The most frequently used pre-processing techniques are tokenization and POS tagging [19, 20]. Other common steps are sentence splitting, lemmatisation, chunking and dependency parsing, and structural parsing. Recent works indicate that robust lifting systems require accurate tuning of several steps, especially tokenization and semantic similarity [21]. Recently, deep neural networks, especially end-to-end methods, have reduced the need for pre-processing steps. Moreover, using deep neural networks for pre-processing tasks such as tokenization has recently produced promising results [22]. The proposed NLP lifter could include as many pre-processing steps as possible, which will be in separate dockers, so the user can choose all suitable ones for the target data.

4.2. Named entity recognition

Named entity recognition is the task that identifies the named entities contained in the text like persons, locations, organizations, time, date, money, etc. NER approaches could be categorised into three main groups: knowledge-based approaches, learning-based methods, and feature-inferring neural network methods. Despite the existence of recent SOTA NER results (especially recent deep NN approaches) such as [23, 24, 25, 26], these approaches have not been utilized and exploited in the process of lifting natural language to knowledge graphs as mentioned earlier. This paper aims to implement those SOTA NER methods in docker-based components to tackle this shortcoming.

4.3. Named entity linking

NEL annotates each mention in a text with the identifier of its corresponding entity that is described in a KB in the LOD cloud. Our paper has defined NEL as a wider task that includes NED as one of its processes. Many NEL approaches are utilizing off-the-shelf systems for NER task. It is, however, a challenging task to choose which particular model to use for those systems. That is because it requires to estimate the similarity level between the system's training datasets and the dataset that needs to be processed in which we strive to accurately recognize entities, according to [27]. Most recent SOTA systems on AIDA-CoNLL dataset includes [28, 29, 30, 31]. There is no perfect NEL model for all datasets and one model might be the best on one dataset but perform poorly on others. Accordingly, having the top N best SOTA implemented in dockers will allow the user to pick the most suitable model for his data and/or replace or update them at any time when needed.

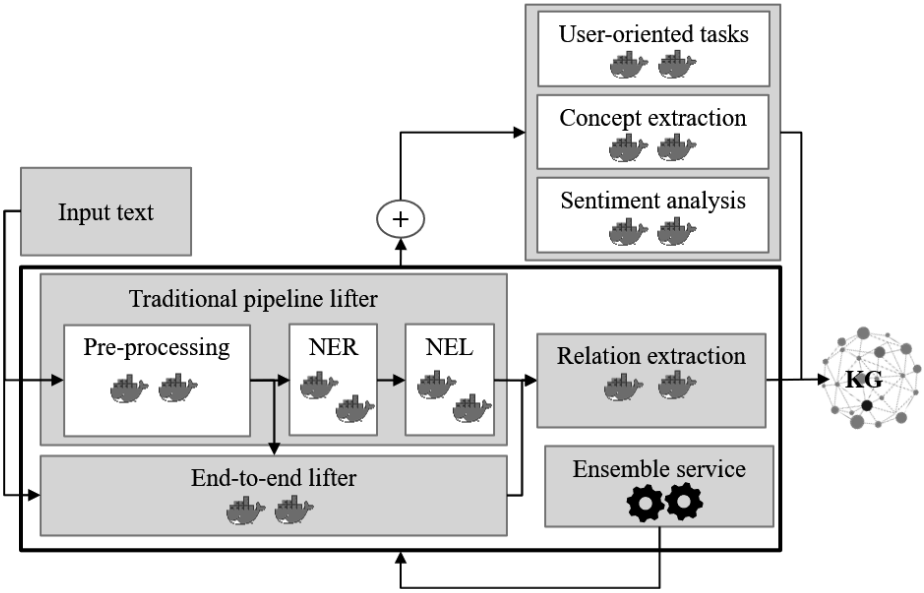


Figure 2: General NLP lifting architecture

4.4. End-to-end track

The majority of previous studies were mostly assuming the availability of mentions and entities and focused on the disambiguation process only. However, leveraging mutual dependency between mentions and their entities is neglected. Moreover, it is not a practical idea in a real-world application. Different from that and to overcome those shortcomings, end-to-end deals with row text and aims to extract all mentions and link them to their entities in the knowledge base. End-to-end entity linking has been recently proposed and is receiving increasing attention. Few studies have been published which claiming the application of the end-to-end approach [32, 33, 34, 35]. The most interesting ones are the most recent neural-based end-to-end linking models [36, 37, 38, 39]. One of the most recent SOTA is [38] followed by [36]. Our NLP lifter aims at including such techniques as an alternative recent track for lifting news texts into a semantic knowledge graph.

4.5. Relation and concept extraction

Our NLP lifter aims at covering lifting of general concepts and of relations between entities. Many recent approaches also lift relations jointly with entities (both named entities and concepts) and reported the SOTA results. Similar to previous components, the proposed lifter will implement those methods and include them as optional tasks as many others for the user.

4.6. User-oriented tasks

User-oriented tasks include those tasks specific and personalised for the project where the NLP lifting architecture is implemented. Apart from including SOTA NLP tasks like the previously described, the NLP lifting architecture takes into account purpose specific tasks such as news angles detection, event detection, IPTC media codes annotation, rumours detection and text completion.

4.7. Knowledge graph

In a knowledge graph, the nodes represent either concrete objects, concepts, information resources, or data about them, and the edges represent semantic relations between the nodes [40]. Knowledge graphs thus offer a widely used format for representing information in computer-processable form. They build on, and are heavily inspired by, Tim Berners-Lee's vision of the semantic web, a machine-processable web of data that augments the original web of human-readable documents [41]. Knowledge graphs can therefore leverage existing standards such as RDF, RDFS, and OWL. Moreover, the constructed knowledge graph could be used to implement more operations like question answering, knowledge graph-based sentence autocompletion, storytelling, fact-checking and so forth using semantic news analysis.

5. Conclusion

Lifting high-volume streams of news texts involves representing their content in machine-understandable formats. KGs is one such formats that has received much attention recently. NLP lifters are an important prerequisite for making the abundance of natural language news on the internet available as computer-processable knowledge graphs. Thus, the presented NLP lifting pipeline provides with an structured and formalised process for transforming natural language text into computer-processable knowledge graphs. The presented pipeline can incorporate any NLP technique like traditional or end-to-end approaches and combining its results or expand them with specific-purpose NLP method like sentiment analysis. Moreover, the introduced NLP lifter is designed to simplify its components replaceability by making use of docker technology, facilitating e.g., the update of all tasks and methods to SOTA approaches. Although the proposed JKP is designed mainly to help journalists, it could be used and customized for the public. The presented NLP lifting architecture aims to be used as reference for developers and researchers of JKP interested in real-time NEL. News organisations may need to adapt their systems, replace components, add new SOTA technologies, or integrate it with other JKP, thus having such NLP lifting pipeline as reference facilitate its management and understanding. Furthermore, it is not restricted to news text and could be used to lift other types of texts.

In our future work, we plan to validate the results of our proposed NLP lifter by using both a manually collected and annotated corpus of news and gold-standards, and compare the results of our proposed lifter with current NEL systems such as ADEL, SpaCy lifter, NewsReader, Stanford CoreNLP or DBpedia Spotlight. Besides, we want to explore the possibilities that validations tools like GERBIL [42] can provide when applied inside our NLP lifter. We believe that validation tools can provide insights about the evolution and performance of the applied NLP processes which can be incorporated to reinforce, improve and keep updated the NLP models.

Acknowledgments

Supported by the News Angler project funded by the Norwegian Research Council's IKTPLUSS programme as project 275872.

References

[1] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, News hunter: building and mining knowledge graphs for newsroom systems, in: NOKOBIT, volume 26, 2018.

[2] 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.

[3] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, A knowledge graph platform for newsrooms, Computers in Industry (2020). To appear.

[4] B. Tessem, A. L. Opdahl, Supporting journalistic news angles with models and analogies, in: 2019 13th RCIS, IEEE, 2019, pp. 1–7.

[5] 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.

[6] B. Tessem, Analogical news angles from text similarity, in: Artificial Intelligence XXXVI, Springer International Publishing, 2019, pp. 449–455.

[7] A. L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, Software and Systems Modeling (2020) 1–17.

[8] E. Motta, E. Daga, A. L. Opdahl, B. Tessem, Analysis and design of computational news angles, IEEE Access (2020).

[9] M. Albared, M. Gallofré Ocaña, A. Ghareb, T. Al-Moslmi, Recent progress of named entity recognition over the most popular datasets, in: 2019 First International Conference of Intelligent Computing and Engineering (ICOICE), 2019, pp. 1–9.

[10] T. Al-Moslmi, M. Gallofré Ocaña, A. L. Opdahl, C. Veres, Named entity extraction for knowledge graphs: A literature overview, IEEE Access 8 (2020) 32862–32881.

[11] N. Fernández, D. Fuentes, L. Sánchez, J. A. Fisteus, The news ontology: Design and appli-

cations, Expert Systems with Applications 37 (2010) 8694 – 8704.

[12] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: learning about world events from news, in: Proceedings of the 23rd WWW, ACM, 2014, pp. 107–110.

[13] P. Vossen, R. Agerri, 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, Knowledge-Based Systems 110 (2016) 60 – 85.

[14] M. Kattenberg, Z. Beloki, A. Soroa, X. Artola, A. Fokkens, P. Huygen, K. Verstoep, Two architectures for parallel processing of huge amounts of text, in: Proceedings of the Tenth LREC'16), European Language Resources Association (ELRA), 2016, pp. 4513–4519.

[15] U. Germann, P. v. d. Kreeft, G. Barzdins, A. Birch, The summa platform: Scalable understanding of multilingual media, in: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018.

[16] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, X. Tannier, Searching news articles using an event knowledge graph leveraged by Wikidata, in: 30th WWW Conference, 13-17 May 2019, 2019.

[17] N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, News: Bringing semantic web technologies into news agencies, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), The Semantic Web - ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 778–791.

[18] 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.

[19] G. Zhu, C. A. Iglesias, Exploiting semantic similarity for named entity disambiguation in knowledge graphs, Expert Systems with Applications 101 (2018) 8 – 24.

[20] M. Fossati, E. Dorigatti, C. Giuliano, N-ary relation extraction for simultaneous t-box and a-box knowledge base augmentation, Semantic Web 9 (2018) 413–439.

[21] M. Conover, M. Hayes, S. Blackburn, P. Skomoroch, S. Shah, Pangloss: Fast entity linking in noisy text environments, in: Proceedings of the 24th ACM SIGKDD, KDD '18, ACM, 2018, p. 168–176.

[22] T. Boros, S. D. Dumitrescu, R. Burtica, NLP-cube: End-to-end raw text processing with neural networks, in: Proceedings of the CoNLL 2018, ACL, 2018, pp. 171–179.

[23] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli, Cloze-driven pretraining of self-attention networks, in: Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP, ACL, 2019, pp. 5359–5368.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[25] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the NAACL, ACL, 2018, pp. 2227–2237.

[26] L. Liu, X. Ren, J. Shang, X. Gu, J. Peng, J. Han, Efficient contextualized representation: Language model pruning for sequence labeling, in: Proceedings of the 2018 Conference on EMNLP, ACL, 2018, pp. 1215–1225.

[27] J. Plu, G. Rizzo, R. Troncy, Enhancing entity linking by combining ner models, in: H. Sack, S. Dietze, A. Tordai, C. Lange (Eds.), Semantic Web Challenges, Springer International Publishing, Cham, 2016, pp. 17–32.

[28] J. Raiman, O. Raiman, Deeptype: Multilingual entity linking by neural type system evolution, 2018. arXiv:1802.01021.

[29] I. Yamada, H. Shindo, Pre-training of deep contextualized embeddings of words and entities for named entity disambiguation, 2019. arXiv:1909.00426.

[30] Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, Y. Liu, Joint entity linking with deep reinforcement learning, in: The WWW Conference, WWW '19, ACM, 2019, p. 438–447.

[31] A. Luo, S. Gao, Y. Xu, Deep semantic match model for entity linking using knowledge graph and text, Procedia Computer Science 129 (2018) 110 – 114. 2017 International Conference on Identification, Information and Knowledge in the Internet of Things.

[32] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Transactions of the ACL 2 (2014) 231–244. arXiv:10.1162/tacl_a_00179.

[33] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, T. Hofmann, Probabilistic bag-of-hyperlinks model for entity linking, in: Proceedings of the 25th WWW, WWW '16, WWW Conference, 2016, p. 927–938.

[34] D. B. Nguyen, M. Theobald, G. Weikum, J-nerd: Joint named entity recognition and disambiguation with rich linguistic features, Transactions of the ACL 4 (2016) 215–229. arXiv:10.1162/tacl_a_00094.

[35] O.-E. Ganea, T. Hofmann, Deep joint entity disambiguation with local neural attention, in: Proceedings of the 2017 Conference on EMNLP, ACL, 2017, pp. 2619–2629.

[36] N. Kolitsas, O.-E. Ganea, T. Hofmann, End-to-end neural entity linking, in: Proceedings of the 22nd Conference on Computational Natural Language Learning, ACL, 2018, pp. 519–529.

[37] Y. Cao, L. Hou, J. Li, Z. Liu, Neural collective entity linking, 2018. arXiv:1811.08603.

[38] P. Le, I. Titov, Improving entity linking by modeling latent relations between mentions, in: Proceedings of the 56th ACL, ACL, 2018, pp. 1595–1604.

[39] P. H. Martins, Z. Marinho, A. F. T. Martins, Joint learning of named entity recognition and entity linking, in: Proceedings of the 57th ACL, ACL, 2019, pp. 190–196.

[40] D. Allemang, J. Hendler, Semantic Web for the Working Ontologist, second edition ed., Morgan Kaufmann, 2011.

[41] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Scientific american 284 (2001) 28–37.

[42] M. Röder, R. Usbeck, A. N. Ngomo, GERBIL - benchmarking named entity recognition and linking consistently, Semantic Web 9 (2018) 605–625.

NORDIC ECOLABEL
241 731
PRINTED MATTER

Graphic design: Communication Division, UIB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 9788230860205 (print)
9788230840405 (PDF)