Semantic Knowledge Graphs for the News: A Review

ANDREAS L. OPDAHL, University of Bergen, Norway
TAREQ AL-MOSLMI, Independent Researcher, Norway
DUC-TIEN DANG-NGUYEN, University of Bergen, Norway
MARC GALLOFRÉ OCAÑA, University of Bergen, Norway
BJØRNAR TESSEM, University of Bergen, Norway
CSABA VERES, University of Bergen, Norway

ICT platforms for news production, distribution, and consumption must exploit the ever-growing availability of digital data. These data originate from different sources and in different formats; they arrive at different velocities and in different volumes. Semantic knowledge graphs (KGs) is an established technique for integrating such heterogeneous information. It is therefore well-aligned with the needs of news producers and distributors, and it is likely to become increasingly important for the news industry. This paper reviews the research on using semantic knowledge graphs for production, distribution, and consumption of news. The purpose is to present an overview of the field; to investigate what it means; and to suggest opportunities and needs for further research and development.

CCS Concepts: • **Computing methodologies → Semantic networks**; • **Information systems → Information systems applications**.

Additional Key Words and Phrases: News, Journalism, News Production, News Distribution, News Consumption, Knowledge Graphs, Ontology, Semantic Technologies, Linked Data, Linked Open Data, Semantic Web, Literature Review

1 INTRODUCTION

Journalism relies increasingly on computers and the Internet [34]. Central drivers are the big and open data sources that have become available on the Web. For example, researchers have investigated how news events can be extracted from big-data sources such as tweets [28] and other texts [27] and how big and open data can benefit journalistic creativity during the early phases of news production [35].

Semantic knowledge graphs and other semantic technologies [3] offer a way to make big and open data sources more readily available for journalistic and other news-related purposes. They offer a standard model and supporting resources for sharing, processing, and storing factual knowledge on both the syntactic and semantic level. Such knowledge graphs thus offer a way to make big, open, and other data sources better integrated and more meaningful. They make it possible to integrate the highly heterogeneous information available on the Internet and to make it more readily available for journalistic and other news-related purposes.

This paper will *systematically review the research literature on semantic knowledge graphs in the last two decades*, from the time when the Semantic Web — an important precursor to semantic knowledge graphs — was first proposed [6]. The purpose is to *present an overview of the field*; *to investigate what it means*; and to *suggest opportunities and needs for further research and development.* We understand both semantic knowledge graphs and

Authors' addresses: Andreas L. Opdahl, University of Bergen, Bergen, Norway, Andreas.Opdahl@uib.no; Tareq Al-Moslmi, Independent Researcher, Oslo, Norway, tareqmail19@gmail.com; Duc-Tien Dang-Nguyen, University of Bergen, Bergen, Norway, Duc-Tien.Dang-Nguyen@uib.no; Marc Gallofré Ocaña, University of Bergen, Bergen, Norway, Marc.Gallofre@uib.no; Bjørnar Tessem, University of Bergen, Bergen, Norway, Bjornar.Tessem@uib.no; Csaba Veres, University of Bergen, Bergen, Norway, Csaba.Veres@uib.no.

0360-0300/2022/6-ART
https://doi.org/10.1145/3543508

the news in a broad sense. Along with *semantic knowledge graphs*, we include facilitating semantic technologies like RDF, OWL, and SPARQL and their uses for semantically Linked (Open) Data and Semantic Web.[1] We also include all aspects of production, distribution and consumption of *news*. More precise inclusion and exclusion criteria will follow in Section 2.

To the best of our knowledge, no literature review has previously attempted to cover this increasingly important area in depth. Several reviews have been published recently on computational journalism in its various guises (e.g., [12, 17, 49, 54]), but none of them go deeply into the technology in general nor into semantic knowledge graphs in particular. Also, recent overviews of knowledge graphs (e.g., [13, 19, 24, 26, 41]) do not consider the specific challenges and opportunities for journalism or the news domain. Among the few papers that discuss the relation between semantic technologies and news, [45] discusses how Linked Data can be integrated into and add value to news production processes and value chains in a non-disruptive way It presents use cases from dynamic semantic publishing at BBC with attention to professional scepticism towards technology-driven innovation. More recently, *Newsroom 3.0* [40] builds on an international field study of three newsrooms — in Brazil, Costa Rica and the UK — to propose a framework for managing technological and media convergence in newsrooms. The framework uses semantic technologies to manage news knowledge, attempting to support interdisciplinary teams in their coordination of journalistic activities, cooperative production of content, and communication between professionals and news prosumers. *Transitions in Journalism* [44] discusses how new technologies are constantly challenging well established journalistic norms and practices, discussing ways in which *semantic journalism* can exploit semantic technologies for everyday journalism.

Compared to these targeted efforts, this paper presents the first systematic review of semantic knowledge graphs for news-related purposes in a broad sense. We ask the following research questions:

RQ1: *Which research problems and approaches are most common, and what are the central results?* For example, the different research contributions may produce different types of results; use different research methods; target different users; focus on different news-related tasks using different input data; use different semantic and other techniques; and address different news domains, languages, and phases of the news life-cycle.

RQ2: *Which research problems and approaches have received less attention, and what types of contributions are rarer?* Where are the green fields and other areas where knowledge is limited and further research needed?

RQ3: *How is the research evolving?* Different problems, result types, and approaches may be more or less prominent at different times, and each of them may be dealt with differently at different times.

RQ4: *Which are the most frequently cited papers and projects, and which papers and projects are citing one another?* For example, how is the research literature organised; which earlier results are cited most by the main papers; and which main papers are most cited in the broader literature?

To answer these questions, the rest of the paper is organised as follows: Section 2 outlines the literature-review process. Section 3 reviews the main papers. Section 4 discusses the main papers, answers the research questions, and offers many paths for further work. Section 5 concludes the paper. The paper is supported by an online *Addendum*[2] that: describes our systematic review method in further detail; provides additional analyses of the main papers and related papers; and offers further readings about the resources and tools that are mentioned in the papers we review. These further readings are marked with an ”A” in the main text, for example “RDF^{A121}”.

[1] Hence, the paper will use the term “semantic knowledge graph” or “semantic KG” in an inclusive way that also covers semantic technologies, computational ontology, Linked Open Data (LOD), and Semantic Web.

[2] doi:10.5281/zenodo.6611518

2 METHOD

To answer our research questions, we conduct a *systematic literature review (SLR)* [31]. In line with our aim to present an overview of the field, we review the research literature in *breadth* in order to cover as many salient research problems, approaches, and potential solutions as possible. A detailed description of our systematic review method is available in the online *Addendum*[2] (Section A).

Our review covers research on *semantic knowledge graphs* for the *news* understood in a wide sense. We include papers that use semantic technologies like RDF[A121], OWL[A117], and SPARQL[A124] [3] and practices like Linked (Open) Data [7] and Semantic Web [6], but we exclude papers that use graph structures only for computational purposes isolated from the semantically-linked Web of Data. We also include all aspects of production, distribution, and consumption of news, but we exclude research that uses news corpora only for evaluation purposes.

We search for literature through the five search engines ACM Digital Library[A14], Elsevier ScienceDirect[A32], IEEE Xplore[A63], SpringerLink[A83], and Clarivate Analytics' Web of Science[A18]. We also conduct supplementary searches using Google Scholar[A52]. We search using variations of the phrases "knowledge graph", "semantic technology", "linked data", "linked open data", and "semantic web" combined with variations of "news" and "journalism" adapted to each search engine's syntax. We select peer-reviewed and archival papers published in esteemed English-language journals or in high-quality conferences and workshops.

The search results are screened in three stages, so that each selected paper is in the end considered by at least three co-authors. In the first stage we screen search results based on title, abstract, and keywords. In the second stage, we skim the full papers and also consider the length, type, language, and source of each paper. In the third stage, we analyse the selected papers in detail according to the framework described below (Table 1). When several papers describe the same line of work, we select the most recent and comprehensive report. In the end, more than six thousand search results are narrowed down to 80 fully analysed *main papers*. They are listed near the end of this paper, right before the Reference list, and we distinguish them from other references by the letter "M", e.g., [M37].

Through a pilot study, we establish an analysis framework that we continue to revise and refine as the analysis progresses [10]. Table 1 lists the ten *top-level themes* in the final framework, along with examples of *sub-themes* that we use to describe and compare the main papers in Section 3. For example, many main papers address specific groups of intended users. *Intended users* therefore becomes a top-level theme in our framework, with more specific groups of users, such as *journalists*, *archivists* and *fact checkers*, as sub-themes.

We make the detailed paper analyses along with their metadata available as a semantic knowledge graph through a SPARQL endpoint at http://bg.newsangler.uib.no. To support impact analysis, the metadata includes all incoming and outgoing citations of and by our main papers. The complete graph contains information about 4238 papers, 9712 authors, and 699 topics from Semantic Scholar[A36]. The online Addendum[2] (Section A) provides further details, and presents examples of SPARQL queries that can be used to explore the graph (Table 10).

3 REVIEW OF MAIN PAPERS

This section reviews the 80 main papers according to the themes of Table 1. Our review and discussion is based on careful manual reading, analysis, marking, and discussion of the main papers, organised by the evolving themes and sub-themes in our analysis framework.

3.1 Technical result types

As shown in Figure 1a, the main papers present a wide variety of technical research results. Further details are available in Table 6.

Pipelines and prototypes: A clear majority of main papers develop ICT architectures and tools for supporting news-related information processing with semantic knowledge graphs and related techniques. Most common are

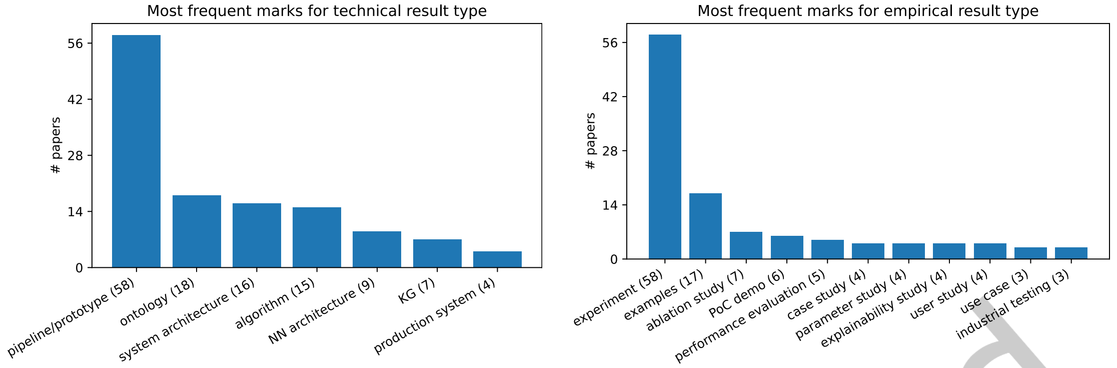
Table 1. Analysis framework.

Technical result type:	What type of technical result does the paper present (e.g., pipelines/prototypes, industrial platforms, algorithms, or information resources such as ontologies and knowledge graphs)?
Empirical result type:	Which research methods are used (e.g., experiments, case studies, or industrial testing)?
Intended users:	Who are the intended users or direct benefactors of the research result (e.g., the general news users, journalists, archivists, or knowledge workers in general)?
Task:	What kind of news-related tasks does the research attempt to support or improve (e.g., semantic annotation, event detection, relation extraction, or content retrieval, provision, and enrichment)?
Input data:	Which sources and types of input data are used (e.g., digital news articles, social media messages, or multimedia news)?
News life cycle:	Which phases of the news cycle are targeted (e.g., future news, emerging news, breaking news, developing news, or already published news)?
Semantic techniques:	Which semantic resources, techniques, and tools are used to create, manage, and exploit semantic knowledge graphs (including information exchange standards, ontologies and vocabularies, semantic data resources, and processing and storage techniques)?
Other techniques:	Which other computing techniques and standards are used in combination with semantic knowledge graphs, including news standards and techniques for natural-language processing (NLP), machine learning (ML), and deep learning (DL)?
News domain:	Does the research target a specific news domain (e.g., economy/finance, the environment, or education)?
Language and region:	Does the research focus on a specific language or combination of languages?

research prototypes and experimental pipelines. For example, the *Knowledge and Information Management (KIM)* platform [M37] is an early and much cited information extraction system that annotates and indexes named entities found in news documents semantically and makes them available for retrieval. To allow precise ontology-based retrieval, each identified entity is annotated with both a specific instance in an extensive knowledge base and a class defined in the associated *KIM Ontology (KIMO)*[3]. which defines around 250 classes and 100 attributes and relations. The platform offers a graphical user interface for viewing, browsing and performing complex searches in collections of annotated news articles. Another early initiative is the *News Engine Web Services (NEWS)* project [M15], which presents a prototype that automatically annotates published news items in several languages. The aim is to help news agencies provide fresh, relevant, and high-quality information to their customers. NEWS uses a dedicated ontology (the *NEWS Ontology* [M17]) to facilitate semantic search, subscription-based services, and news creation through a web-based user interface. *Hermes* [M4] supplies news-based evidence to decision makers. To facilitate semantic retrieval, it automatically identifies topics in news articles and classifies them. The topics and classes are defined in an ontology that has been extended with synonyms and hypernyms from WordNet [18, 37] to improve recall.

Production systems: Some main papers take one step further and present industrial platforms that have run in news organisations, either experimentally or in production. The earliest example is *AnnoTerra* [M57], a system developed by NASA to enhance earth-science news feeds with content from relevant multimedia data sources. The system matches ontology concepts with keywords found in the news texts to identify data sources and support semantic searches. Also, [50] reports industrial experience with NEWS at EFE, a Spanish international news agency. The most recent example is VLX-Stories [M14], a commercial, multilingual system for event detection and information retrieval from media feeds. The system harvests information from online news sites; aggregates

[3] Later superseded by the PROTON ontology[A1].



(a)

(b)

Fig. 1. The most frequent (a) technical and (b) empirical result types.

them into events; labels them semantically; and represents them in a knowledge graph. The system is also able to detect emerging entities in online news. VLX-Stories is deployed in production in several organisations in several countries. Each month, it detects over 9000 events from over 4000 news feeds from seven different countries and in three different languages, extending its knowledge graph with 1300 new entities as a side result.

System architectures: Whether oriented towards research or industry, another group of papers propose system architectures. The *World News Finder* [M34] presents an architecture that is representative of many systems that exploit KGs for managing news content. Online news articles in HTML format are parsed and analysed using GATE (General Architecture for Text Engineering)[A92] and ANNIE (A Nearly New Information Extraction system)[A91] with the support of JAPE (Java Annotations Pattern Engine)[A93] rules and ontology gazetteering lists. A domain ontology is then used in combination with heuristic rules to annotate the analysed news texts semantically. The annotated news articles are represented in a metadata repository and made available for semantic search through a GUI.

Algorithms: Another group of papers focus on developing algorithms that exploit semantic knowledge graphs and related techniques, usually supported by proof-of-concept prototypes that are also used for evaluation. Inspired by Google's PageRank algorithm[A25], [M16] proposes the *IdentityRank* algorithm for named entity disambiguation in the NEWS project [M15]. IdentityRank dynamically adjusts its weights for ranking candidate instances based on news trends (the frequency of each instance in a period of time) and semantic coherence (the frequency of the instance in a certain context), and it can be retrained based on user feedback and corrections. [M54] takes trending entities in news streams as its starting point and attempts to identify and rank other entities in their context. The purpose is to represent trends more richly and understand them better. One unsupervised and one supervised algorithm are compared. The unsupervised approach uses a personalised version of the PageRank algorithm[A25] over a graph of trending and contextual entities. The edges encode directional similarities between the entities using embeddings from a background knowledge graph. The supervised, and better performing, approach uses a selection of hand-crafted features along with a learning-to-rank (LTR) model, LightGBM[A79]. The selected features include positions and frequencies of the entities in the input texts, their co-occurrences and popularity, coherence measures based on TagMe[A90] and on entity embeddings, and the entities' local importance in the text (or salience). *NewsLink* [M75] processes news articles and natural-language (NL) queries from users in the same way, using standard natural-language processing (NLP) techniques. Co-occurrence between entities in a news article or query is used to divide it into segments, for example corresponding to sentences. The entities in each segment are mapped to an open KG from which a connected sub-graph is extracted to represent the segment. The sub-graphs are then merged to represent the articles and queries as KGs that can be compared for similarity to support more robust and explainable query answering. Hermes also provides an algorithm (to be presented later) for ranking semantic search results [M25].

Neural-network architectures: Rather than proposing algorithms, many recent main papers instead exploit semantic knowledge graphs for news purposes using deep neural-network (NN) architectures. These papers too are supported by proof-of-concept prototypes, which are usually evaluated using gold-standard datasets and information retrieval (IR) metrics. *Heterogeneous graph Embedding framework for Emerging Relation detection (HEER)* [M79] detects emerging entities and relations from text reports, i.e., new entities and relations in the news that have so far not been included in a knowledge graph. The challenges addressed are that new entities and relations appear at high speed, with little available information at first, and without negative examples to learn from. HEER represents incoming news texts as graphs based on entity co-occurrence and incrementally maintains joint embeddings of the news graphs and an open knowledge graph. The result is positive and unlabelled (PU) entity embeddings that are used to train and maintain a PU classifier[A31] that detects emerging relations incrementally.

Context-Aware Graph Embedding (CAGE) [M66] is an approach for session-based news recommendation. Entities are extracted from input texts and used to extract a sub-knowledge graph from an open knowledge graph (the paper uses Wikidata[A47]). Knowledge-graph embeddings are calculated from the sub-knowledge graph, whereas pre-trained word embeddings and Convolutional Neural Networks (CNNs) [22] are used to derive content embeddings from the corresponding input texts. The knowledge-graph and content embeddings are concatenated and combined with user embeddings and refined further using Convolutional Neural Networks (CNNs). Finally, an Attention Neural Network (ANN) [55] on top of Gated Recurrent Units (GRUs) [22] are used to recommend articles from the resulting embeddings, taking short-term user preferences into account.

Deep Triple Networks (DTN) [M42] use a deep-network architecture for topic-specific fake news detection. News texts are analysed in two ways in parallel: The first way is to use word2vec [36] embeddings and self-attention [55] on the raw input text. The second way is to extract triples from the text and analyse them using TransD [30] graph embeddings, attention and a bi-directional LSTM (Long Short-Term Memory) [22]. A CNN is used to combine the results of the two parallel analyses into a single output vector. Background knowledge has been infused into the second way by training the TransD graph embeddings, not only on the triples extracted from the input text, but also on related triples from a 4-hop DBpedia [5] extract. Maximum and average biases from the graph triples are concatenated with the CNN output vector and used to classify news texts as real or fake. The intuition behind this and other bias-based approaches to fake news detection is that, if the input text is false, triples learned only from the input text will have smaller bias than triples learned from the same text infused with true (and thus conflicting) real-world knowledge.

Ontologies: Almost half the papers include a general or domain-specific ontology for creating and managing other semantic knowledge graphs. For example, the NEWS project uses OWL to represent the NEWS Ontology [M17], which standardises and interconnects the semantic labels used to annotate and disseminate news content. The *Semantics-based Pipeline for Economic Event Detection (SPEED)* [M24] uses a finance ontology represented in OWL to ensure interoperability between and reuse of existing semantic and NLP solutions. [M71] represents the IPTC (International Press Telecommunications Council) News Codes[A68] as SKOS[A123] concepts in an OWL ontology, and discusses its uses for semantic enrichment and search. The Evolutionary Event Ontology Knowledge (EEOK) ontology [M45] represents how different types of news events tend to unfold over time. The ontology is supported by a pipeline that mines event-evolution patterns from natural-language news texts that report different stages of the same macro event (or storyline). The patterns are represented in OWL and used to extract and predict further events in developing storylines more precisely.

Knowledge graphs: A few papers even present a populated, instance-level semantic knowledge graph or other linked knowledge base as a central result. For example, K-Pop [M36] populates a semantic knowledge graph for enriching news about Korean pop artists. The purpose is to provide comprehensive profiles for singers and groups, their activities, organisations and catalogues. As an example application, the resulting entertainment KG is used to power *Gnosis*, a mobile application for recommending K-Pop news articles. CrimeBase [M67] presents a knowledge graph that integrates crime-related information from popular Indian online newspapers. The purpose is to help law enforcement agencies analyse and prevent criminal activities by gathering and integrating crime entities from text and images and making them available in machine-readable form. *ClaimsKG* [M69] is a live knowledge graph that represents more than 28000 fact-checked claims published since 1996, totalling over 6 million triples. It uses a semi-automatic pipeline to harvest fact checks from popular fact-checking websites; annotate them with entities from DBpedia; represent them in RDF[A121] according to a semantic data model in RDFS[A122]; normalise the validity ratings; and resolve co-references across claims. [M12] uses hashtags and other metadata associated with tweets and tweeters to build an RDF model of over 900.000 French political tweets, totalling more than 20 million triples that describe facts, statements, and beliefs in time. The purpose is to trace how actors propagate knowledge — as well as misinformation and hearsay — over time.

Formal models: A small final group of papers propose formal models of various types and for different purposes. For example, [M22] presents a formal model for managing inconsistencies that arise when live news streams are represented incrementally using description logic. A trust-based algorithm for belief-base revision is presented that takes users' trust in information sources into account when choosing which inconsistent information to discard.

Summary:

Our review suggests that the most common types of results are pipelines and prototypes. In addition, many papers propose ontologies, system architectures, algorithms and neural-network architectures. A few papers also introduce new knowledge graphs. There has been a shift in recent years from research on algorithms and system architectures towards papers that propose deep neural-network architectures. A few of those recent papers also mention explainability.

3.2 Empirical result types

As shown in Figure 1b, a large majority of the papers include an empirical evaluation of their technical proposals.

Experiments: As shown in the previous section, a majority of papers develop pipelines or prototypes, which are then evaluated empirically. The most common evaluation method is controlled experiments using gold-standard datasets and information retrieval (IR) measures such as precision (P), recall (R), and accuracy (A). For example, KOPRA [M70] is a deep-learning approach that uses a Graph Convolutional Network (GCN) [11, 23] for news recommendation. An initial entity graph (called interest graph) is created for each user from entities mentioned in the news titles and abstracts of that user's short- and long-term click histories. A joint knowledge pruning and Recurrent Graph Convolution (RGC) mechanism is then used to augment the entities in the interest graph with related entities from an open KG. Finally, entities extracted from candidate news texts are compared with entities in the interest graphs to predict articles a user may find interesting. The approach is evaluated experimentally with Wikidata as the open KG and using two standard datasets (MIND and Adressa). *RDFLiveNews* [M21] aims to represent RSS data streams[A22] as RDF triples in real time. Candidate triples are extracted from individual RSS items and clustered to suggested output triples. Components of the approach are evaluated in two ways. The first way measures RDFLiveNews' ability to disambiguate alternative URIs for named entities detected in the input items. Disambiguation results are evaluated against a manually crafted gold standard using precision, recall and F1 metrics and by comparing them to the outputs of a state-of-art NED tool (AIDA[A59]). The second way measures RDFLiveNews' ability to cluster similar triples extracted from different RSS items. The clusters are evaluated against the manually crafted gold standard using sensitivity (S), positive predictive value (PPV), and their geometric mean.

Performance evaluation: A smaller number of experimental papers collect performance measures such as execution times and throughput in addition to or instead of IR measures. For example, the scalability of RDFLive-News [M21] is also measured using run times for different components of the approach on three test corpora. The results suggest that, with some parallelisation, it is able to handle at least 1500 parallel RSS feeds. The performance of KnowledgeSeeker [M39], an ontology-based agent system for recommending Chinese news articles, is measured through execution times on three datasets for a given computer configuration and using the performance of a vanilla TF-IDF-based approach as comparison baseline. The throughput of SPEED [M24] is benchmarked on a corpus of 200 news messages extracted from Yahoo!'s business and technology news feeds[A126].

Ablation, explainability, and parameter studies: Many recent papers also include ablation studies ([M66,M70,M75,M54]), explainability studies ([M45,M70,M75]), and parameter and sensitivity studies ([M79]). A common theme is that they all use deep or other machine learning techniques. We will present more examples later (e.g., [M80,M40,M74,M73,M19,M78]).

Industrial testing: A few papers present case studies or experience reports from industry. We have already mentioned the commercial VLX-Stories [M14] system. [M44] extends the news production workflow at VRT (Vlaamse Radio- en Televisieomroep), a national Belgian broadcaster, in order to support personalised news recommendation and dissemination via RSS feeds. A semantic version of the IPTC's NewsML-G2[A72] standard is proposed as a unifying (meta-)data model for dynamic distributed news event information. As a result, RDF/OWL and NewsML-G2 can be used in combination to automatically categorise, link, and enrich news-event metadata. The system has been hooked into the VRT's workflow engine, facilitating automatic recommendation of developing news stories to individual news users. [M68] semantically enriches the content of archival news texts. The proposed system identifies mentions of named entities along with their contexts; links the contextualised mentions to entities in a knowledge base; and uses the links to retrieve further relevant information from the knowledge base. The system has been deployed and applied to ten years of archival news in a local Italian newspaper. And as already mentioned, a prototype of the NEWS system [M15] has run experimentally at EFE, alongside their legacy production system, introducing a semi-automatic workflow that lets journalists validate the annotations suggested by the system [50].

Case studies and examples: Other papers present realistic examples based on industrial experience. For example, the *MediaLoep* project [M10] (involving many of the authors behind [M44], and [M9] to be presented later), discusses how to improve retrieval and increase reuse of previously broadcast multimedia news items at VRT, the national Belgian broadcaster, both as background information and as reusable footage. The paper reports experiences with collecting descriptive metadata from different news production systems; integrating the metadata using a semantic data model; and connecting the data model to other semantic data sets to enable more powerful semantic search.

Proof-of-concept demonstrations and use cases: Similar types of qualitative evaluations, but with less focus in industrial-scale examples, are proof-of-concept demonstrations and hypothetical use cases (e.g., [M65]).

User studies: A final group of papers presents user studies and usability tests. [M76] represents news articles as small knowledge graphs enriched with word similarities from WordNet [18, 37]. Overlaps between the sub-graphs of new articles and of articles a user has found interesting in the past are used to recommend new articles to the user. Sub-graphs are compared using Jaccard similarity. The approach is evaluated on a collection of Japanese news articles. 20 users were asked to rate suggested articles in terms of relevance and of interest, breaking the latter down into curiosity and serendipity.

Summary: Our review shows that experimental evaluation of proposed pipelines/prototypes is the most used research method. Experiments most often use information retrieval measures, but usability and performance measures are also employed. In recent years, experiments are increasingly often supplemented by studies of ablation, explainability, and parameter selection. Other used research methods are industrial testing, case studies and examples, proof-of-concept demos, use cases, and user studies.

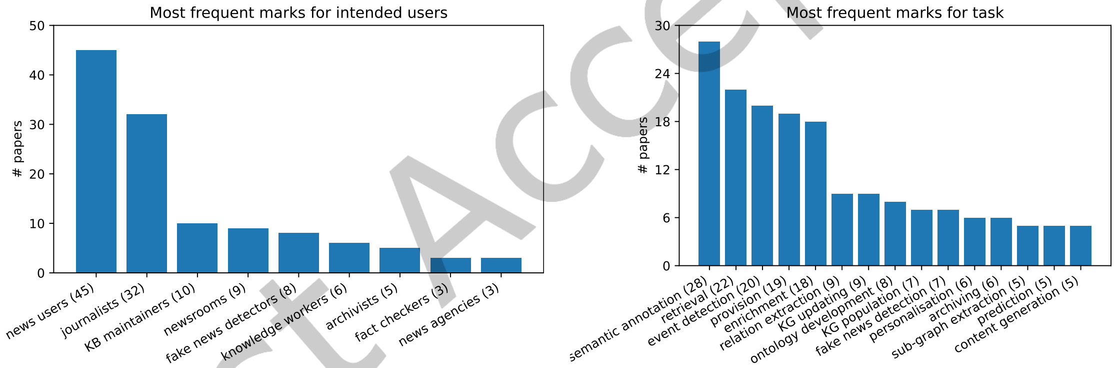
3.3 Intended users

The most frequent types of intended users — or immediate beneficiaries — of the results from our main papers are shown in Figure 2a.

News users: More than half the main papers aim to offer news services to the general public. An early example is *Rich News* [M11], a system that automatically transcribes and segments radio and TV streams. Key phrases extracted from each segment are used to retrieve web pages that report the same news event. The web pages are annotated semantically using the KIM platform [M37], whose web interface is used to support searching and browsing news stories semantically by topic and playing the corresponding segments of the associated media files.

Journalists, newsrooms, and news agencies: The second largest group of papers aim to support journalists and other professionals in newsrooms and news agencies. Several projects mentioned already belong to this type, including the NEWS project [M15]. The proposals in [M10,M71,M45,M54] also target journalists and other news professionals. The ambition of the *News Angler* project [M46] is to enable automatic detection of newsworthy events from a dynamically evolving knowledge graph, by representing news angles, such as "proximity", "nepotism", or "fall from grace" [43], formally using Common Logic[A37].

Knowledge base maintainers: Rather than supporting news users directly, some papers support knowledge base maintainers on a technical level. For example, [M17] presents a plugin for maintaining the NEWS ontology. *Aethalides* [M58] extends the Hermes framework [M4] with a pipeline for semantic classification using concepts defined in a domain ontology.



(a) (b)

Fig. 2. The most frequently (a) intended users and (b) their tasks.

Archivists: A smaller group of papers targets archivists, who maintain knowledge bases on the content level. For example, *Neptuno* [M7] is an early semantic newspaper archive system that aims to give archivists and reporters richer ways to describe and annotate news materials and to give reporters and news readers better search and browsing capabilities. It uses an ontology for classifying archive content along with modules for semantic search, browsing, and visualisation. The purpose of the formal model for belief-base revision [M22] presented earlier is also to maintain knowledge bases by detecting and resolving inconsistencies.

Fake-news detectors and fact checkers: Several recent papers focus on supporting *fake-news detectors* and *fact checkers*. We have already mentioned Deep Triple Networks (DTN) [M42]. [M5] detects fake news through a hybrid approach that assesses sentiments, entities, and facts extracted from news texts. ClaimsKG [M69], the large knowledge graph of French political tweets, can be used to trace how knowledge — along with misinformation and hearsay — is propagated over time. Several of the recent deep-NN approaches we will present later also target fake-news detection and fact checking.

Knowledge workers: A smaller group of papers target general knowledge workers and information professionals outside the news profession. For example, KIM [M37] aims to improve news browsing and searching for knowledge workers in general. Other papers aim to support specific information professions. The *Automatic Georeferencing Video (AGV)* pipeline [M13] makes news videos from the RAI archives available for geography education. Audio is extracted from video using ffmpeg[A4] and transcribed using Ants[A3]. Apache OpenNLP[A46] is used to extract named entities mentioned in the video segment. Google's Knowledge Graph is used to add representative images and facts about related people and places. The places are in turn used to make the videos and their metadata available through Google Street Map[A54]-based user interfaces. The pipeline is tested on a dataset of 10-minute excerpts from 6600 videos from a thematic RAI newscast (Leonardo TGR). *AnnoTerra* [M57] uses ontologies and semantic search to improve NASA's earth-science news feeds, targeting both experts and inexperienced users of earth-science data. *CrimeBase* [M67] uses rules to extract entities from text and associated image captions in multimodal crime-related online news. The extracted entities are correlated using contextual and semantic similarity measures, whereas image entities are correlated using image features. The resulting knowledge base uses an OWL ontology to integrate crime-related information from popular Indian online newspapers. Other main papers (to be presented later) target professionals in domains such as economy and finance [M51], environmental communication [M63], and medicine [M23].

Summary: Our review indicates that the most frequently intended users (or beneficiaries) of the main-paper proposals are general news users and journalists. Other intended users/beneficiaries are newsrooms, knowledge-base maintainers, archivists, fake-news detectors and fact checkers, and different types of knowledge workers.

3.4 Tasks

As shown in Figure 2b, the main papers target a wide range of news production, dissemination, and consumption activities, such as search, recommendation, categorisation, and event detection.

Semantic annotation: Many of the earliest approaches focus on adding semantic labels to entities and topics mentioned in published news texts. We have already introduced KIM [M37], which labels named entities found in news items with links to instances in a knowledge base and to classes defined in the KIM Ontology (KIMO). We have also introduced NEWS [M15], which annotates news items with named entities linked to external sources like Wikipedia[A49], ISO country codes[A38], NASDAQ company codes (e.g.,[A34]), the CIA World Factbook[A27], and SUMO/MILO[A84]. It also categorises the news items by content and represents news metadata using standards and vocabularies such as the Dublin Core (DC)[A29] vocabulary, the IPTC's News Codes[A68], the News Industry Text Format (NITF)[A71], NewsML[A72], and PRISM — the Publishing Requirements for Industry Standard Metadata[A114].

Enrichment: A smaller group of papers instead focus on enriching annotated news items with Linked Open Data or information from other semantically labelled sources. For example, [M2] extends the life of TV content by integrating heterogeneous data from sources such as broadcast archives, newspapers, blogs, social media and encyclopedia and by aligning semantic content metadata with the users' evolving interests. AGV [M13] annotates TV news programs with geographical entities to make archival video content available through a map-based user interface for educational purposes. In addition to representing the IPTC News Codes using SKOS, [M71] discusses how multimedia news metadata can be augmented using natural-language and multimedia analysis techniques and enriched with Linked Data, such as facts from DBpedia [5] and GeoNames[A51]. Contributions that represent news texts as sub-graphs of open KGs like Wikidata (e.g., CAGE [M66], KOPRA [M70], and NewsLink [M75]) can also be considered enrichment approaches. We will present a few similar approaches later ([M40,M80]).

Content retrieval: Other papers use semantic annotations (or "semantic footprints") to support on-demand ("pull") or proactive ("push") dissemination of news content. On the *retrieval* (*on-demand*, *pull*) side, a clear majority of the main papers support tasks such as searching for and otherwise retrieving news items. Projects like KIM [M37], NEWS [M15], and Hermes [M4] all have content provision as central tasks. The *Hermes Graphical Query Language (HGQL)* [M25] makes it simpler for non-expert users to search semantically for content available in the Hermes framework. It is based on RDF-GL[A61], a SPARQL-based graphical query language for RDF, and also provides an algorithm for ranking search results. The *World News Finder* [M34] uses a World News Ontology along with heuristic rules to automatically create metadata files from HTML news documents to support semantic user queries. The aim of *NewsLink* [M75] is to support more robust as well as *explainable* query answering.

Content provision: On the *provision* (*proactive*, *push*) side, another large group of papers focus on actively propagating news to users. For example, [M33] aims to provide more accurate content-based recommendations. It uses existing tools for entity discovery and linking to represent news messages as sub-graphs by adding edges from Freebase[A2]. A new human-annotated data set (CNREC) for evaluating content-based news recommendation systems is made available and used to evaluate the approach. [M6] aims to deal with data sparsity and cold-start issues in news recommender systems. It enriches semantic representations of news items and of users with Linked Data in order to provide more input to recommendation algorithms. Focusing on the user-profiling (or personalisation) side of news recommendation, [M26] uses semantic annotations of news videos to profile users' evolving information needs and interests in order to recommend the most suitable news stories. Context-Aware Graph Embedding (CAGE) [M66] focuses on providing session-based recommendations, whereas KOPRA [M70] aims to take both users' short- and long-term behaviours into account.

Event detection: Several more recent approaches go beyond semantic labelling and enrichment of news content, attempting to extract events or relations (triples, facts) from news items in order to represent their meaning on a fine-grained level. *NewsReader* [M72] is a cross-lingual system (or "reading machine") that is designed to ingest high volumes of news articles and represent them as *Event-Centric Knowledge Graphs (ECKGs)* [M59]. Each graph describes an event, and perhaps how it develops over time, along with the actors and other entities involved in the event. The graphs are connected through shared entities and temporal overlaps, and the entities are linked to background information in knowledge bases such as DBpedia. The *ASRAEL* project [M60] maps events described in unstructured news articles to structured event representations in Wikidata[A47], which are used to enrich the representations of the articles. Because Wikidata's event hierarchy is considered too fine grained for use in search engines, a hierarchical clustering step follows, after which the more coarsely categorised events are made available for querying and navigation through an event-oriented knowledge graph. To keep the Hermes [M4] knowledge base up to date, [M64] represents lexico-semantic patterns and associated actions as rules that are used to semi-automatically detect and semantically describe news events. The approach is implemented in the *Hermes News Portal (HNP)*, a realisation of the Hermes framework that lets news users browse and query for relevant news items. The Evolutionary Event Ontology Knowledge (EEOK) ontology [M45] aims to support event detection by suggesting which event types to look for next in a developing storyline. [M38] identifies and reconciles named events from news articles and represents them in a semantic knowledge graph according to textual contents, entities, and temporal ordering. The commercial tool VLX-Stories [M14] also detects events in media feeds.

Relation extraction: Other papers instead focus on relation extraction, detecting triples (or facts) that can be used to build new or update existing RDF graphs. An early proposal for deeper text analysis is *SemNews* [M30], which extracts textual-meaning representations (TMRs) from RSS[A22] news items using the OntoSem tool (see, e.g.,[A33]), which represents each text as a set of facts about: which actions that are described in the text; which agents, locations and themes each action involves; and any temporal relations between the actions. The SemNews tool transforms the TMRs into OWL to support semantic searching, browsing and indexing of RSS news items. It also powers an experimental web service that provides semantically annotated news items along with news summaries to human users. *BKSport* [M49] automatically annotates sports news using language-pattern rules in combination with a domain ontology and a knowledge base built on top of the KIM platform [M37]. The tool extracts links and typed entities as well as semantic relations between them. It also uses pronoun recognition to resolve co-references. [M55] represents the sentences in a news item as triples, analysing not only top-level but also subordinate clauses. The triples are run through a pipeline of natural language tools that fuse and prioritise them. Finally, selected triples are used to summarise the underlying event reported in the news item. [M18] identifies novel statements in the news, building on ClausIE and DBpedia to propose a semantic novelty measure that takes individual user-relevance into account.

Sub-graph extraction: An alternative to extracting relations from news texts is to represent texts by sub-graphs extracted from open knowledge graphs. An early example is [M33], which uses standard techniques to discover and link entities and adds edges from Freebase to represent news messages as sub-graphs to support content-based news recommendation. *AnchorKG* [M40] represents news articles as small anchor graphs, which consist of entities that are prominently mentioned in the news text, along with relations between those entities taken from an open knowledge graph, and along with those entities' k-hop neighbourhoods in the graph. One aim is to improve news recommendation by making real-time knowledge reasoning scalable to large open knowledge graphs. Another aim is to support explainable reasoning about similarity. Reinforcement learning is used to train an anchor-graph extractor jointly with a news recommender, using already recognised and linked named entities as inputs. The approach is evaluated using the MIND[A80] dataset and a private dataset extracted from Bing News[A78] with Wikidata as reference graph. CAGE [M66] represents news texts as sub-graphs extracted from an open reference knowledge graph to support session-based news recommendation. KOPRA [M70] extracts an entity graph (called interest graph) for each user from seed entities that are mentioned in the news titles and abstracts in the user's short- and long-term click histories. *NewsLink* [M75] represents both news articles and user queries as small KGs that can be compared for similarity.

KG updating: Several recent contributions use deep and other machine-learning techniques to keep evolving knowledge graphs up-to-date by identifying new (emerging, dark) entities and new (or emerging) relations between (the new or existing) entities. We have already mentioned HEER [M79]. *PolarisX* [M77] automatically expands language-independent knowledge graphs in real time with representations of new events reported by news sites and on social media. It uses a relation extraction model based on pre-trained multilingual BERT [16] to detect new relations. Challenges addressed are that available reference knowledge graphs have limited size and scope and that existing techniques are not able to deal with neologisms based on human common sense. *Text-Aware MUlti-RElational learning method (TAMURE)* [M78] also extends a knowledge graph with relations that emerge in the news. It addresses the source heterogeneity of structured knowledge graphs and unstructured news texts by learning joint embeddings of entities, relations, and texts using tensor factorisation implemented in TensorFlow[A13]. TAMURE is linear in the number of parameters, making it suitable for large-scale KGs and live news streams. [M61] empirically investigates the prevalence of entities in online news feeds that cannot be identified by DBpedia Spotlight or by Google's Knowledge Graph API[A53]. Out of 13,456 named entities in an RSS sample, 378 were missing from DBpedia, 488 were missing from Google's Knowledge Graph, and 297 were missing from both.

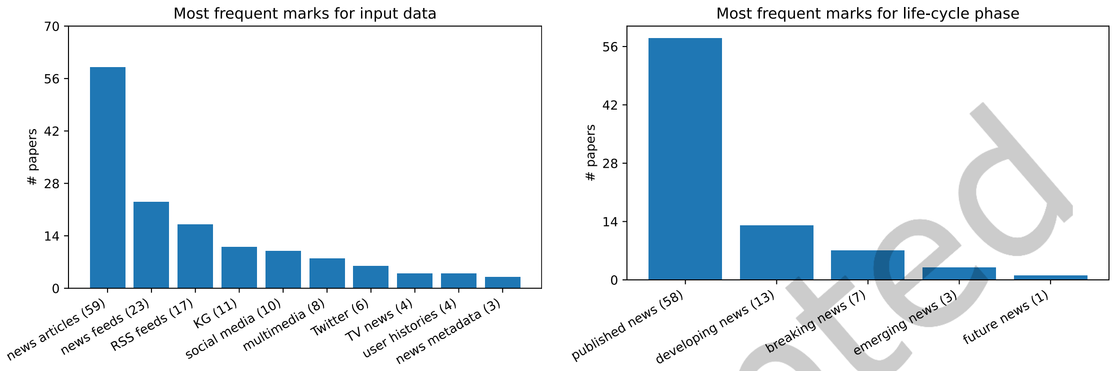
Ontology development: In various ways, several main papers support ontology development. Early projects like KIM [M37] and NEWS [M15] focus on developing new domain ontologies, whereas [M37] integrates existing IPTC standards and vocabularies into the LOD cloud. More recent efforts, such as [M45], use machine learning techniques to automate ontology creation and maintenance.

Fake-news detection and fact checking: Several recent papers focus on the detection of *fake news*, such as [M5]. Another proposal is [M52], which uses graph embeddings of news texts to identify fake news. [M48] presents a multimodal approach to quantify whether real-world news texts and their associated images represent the same or connected entities, suggesting that low coherence is a possible indicator of fake news. [M23] lifts medical information from non-trusted sources into semantic form using FRED [21] and reasons over the resulting description logic representations using Racer[A88] and HermiT[A89]. Reasoning inconsistencies are taken to indicate potential "medical myths" that are verbalised and presented to human agents along with an explanation of the inconsistency. KLG-GAT [M80] uses an open knowledge graph to enhance fact checking and verification. Constituency parsing is used to find entity mentions in the claims, which are used to retrieve relevant Wikipedia articles as potential evidence. A BERT-based sentence retrieval model is then used to select the most relevant evidence for the claim. TagMe is used to link entities in the claims and in the evidence sentences to the Wikidata5M[A56] subset of Wikidata and extract triples whose entities are mentioned in the claim and/or evidence. The triples are further ranked using a BERT-based learning-to-rank (LTR) model. High-ranked triples are used to construct a graph of the central claim, its potential evidence sentences, and triples that connect the claim to the evidence sentences. A two-level multi-head graph attention network is used to propagate information between the claim, evidence and knowledge (triple) nodes in the graph as input to a claim classification layer.

Content generation: Targeting news content generation, *Tweet2News* [M3] extracts RDF triples from documentary (headline-like) tweets using the IPTC's rNews vocabulary[A69], organises them into storylines, and enriches them with Linked Open Data in order to facilitate news generation in addition to retrieval. The *Pundit* algorithm [M56] even *suggests plausible future events* based on descriptions of current events. Structured representations of news titles are extracted from a large historical news archive that covers more than 150 years, and a machine-learning algorithm is used to extract causal relations from the structured representations. Although the authors do not propose specific journalistic uses of Pundit, their algorithm might be used in newsrooms to anticipate alternative continuations of developing events. [M31] aims to auto-generate human-quality news image captions based on a corpus of news texts with associated images and captions. Each news image is represented as a feature vector using a pre-trained CNN, and each corresponding article text is split into sentences containing named entities that are processed further in two ways. One line of analysis enriches the sentences and entities with related information from DBpedia. Another line instead replaces the named entities with type placeholders, such as PERSON, NORP, LOC, ORG, and GPE, producing generic sentences that are compressed using dependency parsing and represented as TF-IDF weighted bags-of-words. Correlations are then established between the generic-sentence representations and the features of the associated images in the corpus. An LSTM model is trained to generate matching caption templates for images on top of the pre-trained CNN. Finally, the semantically-enriched original sentences are used to fill in individual entities for the type placeholders. The approach is evaluated on two public datasets, Good News[A21] (466k examples) and Breaking News[A100] (110k examples), that include news images and captions along with article texts. [M55] (presented earlier) uses the triples that have been extracted, fused and prioritised from news sentences to generate new sentences that *summarise* the underlying news events. The *News Angler* project [M46] represents news angles in order to support automatic detection of newsworthy events from a knowledge graph.

Prediction: Prediction is the focus of a small group of papers that include Pundit [M56] and EEOK [M45]. In order to predict stock prices, *EKGStock* [M41] uses named-entity recognition and relation extraction to represent news about Chinese enterprises as knowledge graphs. Embeddings of the enterprise-specific graphs are then used to estimate connectedness between enterprises. Sentiments of news reports that mention an enterprise are then fed into a Gated Recurrent Unit (GRU) [22] model that predicts stock prices, not only for the mentioned enterprise, but also for its semantically related ones. Recent predictive approaches include deep-neural network-based recommendation papers, such as [M73] (more later), that are trained to predict click-through rates (CTR).

Other tasks: In addition to these most frequent uses of knowledge graphs for news, several main papers address *semantic similarity*. For example, [M35] uses information extraction techniques to automatically annotate news documents semantically in order to facilitate cross-lingual retrieval of documents with similar annotations. [M9] clusters semantic representations to detect how news items are derived from one another, using the PROV-O ontology[A120] to represent the results semantically. Supporting *visualisation*, the *Visualizing Relations Between Objects (VRBO)* framework [M32] uses semantic and statistical methods to identify temporal patterns between entities mentioned in economic news. It uses the patterns to create and visualise news alerts that can be formalised and used to manage equity portfolios. Neptuno [M7] uses visualisation on the ontology level to show and publish how knowledge-base concepts are organised. *Archiving* and general *information organisation* is a central task of Neptuno [M7] and several other main papers. *Interoperability* and data *integration* is the focus in *MediaLoep* [M10]. Focusing on multimedia and other metadata, [M20] (more later) also has interoperability as a central task, along with contributions such as [M57,M53,M24,M2].



(a)

(b)

Fig. 3. The most frequent types of (a) input data used and (b) news life-cycle phases targeted.

Summary: Our review shows that the research on semantic knowledge graphs for the news support a broad variety of tasks, such as semantic annotation, enrichment, content retrieval and provision, event detection, relation and sub-graph extraction, KG updating, ontology development, fake-news detection and fact checking, content generation, and prediction. The last few years have seen a rapidly growing interest in KGs for fake-news identification. Support for factual journalism is a related area that is growing. Automatic news detection is another emerging area that is becoming increasingly important.

3.5 Input data

As shown in Figure 3a, the proposed approaches rely on a variety of sources and types of primary input data. Note that this section discusses the *data used as input* by the solutions proposed or discussed in each main paper, and not the data used for evaluation.

News articles: The most common input data are textual news articles in digital form. For example, [M47] reads template-based HTML pages and exploits semantic regularities in the templates to automatically annotate HTML elements with semantic labels according to their DOM paths. Online news articles are also used as examples and for evaluation.

RSS and other news feeds: Other main papers take their inputs via RSS feeds or other news feeds. The *Ontology-based Personalised Web-Feed Platform (OPWFP)* [M28] inputs RSS news streams and uses an ontology to provide more precisely customised web feeds. User profiles are expressed using the semantic Composite Capability/Preference Profiles (CC/PP)[A115] and FOAF[A24] vocabularies along with a domain ontology. The three vocabularies and ontologies are used in combination to select appropriate search topics for the RSS search engine.

Social media and the Web: Several main papers use social media and other web resources as input, such as Twitter[A110], Wikinews[A48], Wikipedia[A49], and regular HTML-based web sites. To support personalised news recommendation and dissemination, the extension of VRT's news workflow mentioned earlier [M44], uses OpenID[A41] and OAuth[A9] for identification and authentication. In this way, the system can compile user profiles based on data from multiple social-media accounts, using ontologies such as FOAF and SIOC[A23] to interoperate user data. Focusing on geo-hashtagged tweets, *Location Tagging News Feed (LTNF)* [M8] is a semantics-based system that extracts geographical hashtags from social media and uses a geographical domain ontology to establish relations between the hashtags and the messages they occur in. Wikipedia is also used as a direct source of input in a few papers [M1,M36].

Multimedia news: Several papers use *multimedia data* as input. [M31] analyses news texts in combination with associated images to suggest human-level image captions. To extend the lifetime of TV news, the AGV pipeline [M13] makes news videos from the RAI archives available for geography education.

News metadata: Focusing on multimedia metadata, [M20] inputs metadata embedded in formats such as MPEG-7 for content description and MPEG-21 for delivery and consumption. The approach uses semantic mappings from XML Schema to OWL and from XML to RDF to integrate administrative multimedia metadata in newspaper organisations. As already explained, MediaLoep [M10] also integrates descriptive multimedia news metadata from news production systems semantically.

Knowledge graphs: Many papers use existing knowledge graphs as inputs. The number has risen in the last few years due to the appearance of deep-NN architectures that infuse triples from open KGs to enhance learning from news texts. Indeed, almost all the recent deep learning papers exploit open KGs in this way, e.g., [M40,M66,M42,M31,M80,M70].

User histories: A smaller group of deep-NN papers input user histories, for example in the form of click logs, to train recommenders [M66,M73,M70,M19].

Summary: Our review shows that the research on semantic knowledge graphs for the news exploits a broad range of input sources. Textual news articles in digital form is the most important source. Other frequently used types of input data are RSS and other news feeds, social media and the Web, multimedia news, news metadata, knowledge graphs, and user histories. Multimedia, including TV news, were popular in first years of the study period and have seen a rebound in the deep-learning era. RSS and other news feeds were popular for many years, but have recently been overtaken by social media, including Twitter. In recent years, KGs are being used increasingly often to infuse world knowledge into deep NNs for news analysis. User histories have also been used in recent recommendation papers.

3.6 News life cycle

The main papers also target different phases of the news life cycle, as shown in Figure 3b. The largest group of papers focuses on organising and managing already *published news*. For example, Neptuno [M7] extends the life of published news by annotating reports with keywords and IPTC codes, thereby relating past news reports to current ones that share the same keywords or code. It thus re-contextualises old news in light of more recent events. The MediaLoep data model [M10] supports managing information generated by news production and publishing processes. AGV [M13] makes archival news videos available for geography education.

Focus in recent years has shifted from focusing on already published news to also targeting earlier phases of the news life cycle. As already mentioned, the Pundit algorithm [M56] predicts likely *future news* events based on short textual descriptions of current events. A small group of mostly Twitter-based papers deal with detecting *emerging news*, or potentially newsworthy events or situations that are not yet reported as news but that may be circulating in social media or elsewhere. For example, Tweet2News [M3] identifies emerging news from documentary (or headline-like) tweets and lifts them into RDF graphs, which are then enriched with triples from the LOD cloud and arranged into storylines to generate news reports in real time.

Focusing on *breaking news*, the Semantics-based Pipeline for Economic Event Detection (SPEED) [M24] uses a domain ontology to detect and annotate economic events. The approach combines ontology-based word and event-phrase gazetteers; a word-phrase look-up component; a word-sense disambiguator; and an event detector that recognises event-patterns described in a domain ontology. The Evolutionary Event Ontology Knowledge (EEOK) [M45] ontology presented earlier represents the typical evolution of developing news stories as patterns. It can thereby be used to predict the most likely next events in a developing story and to train dedicated detectors for different event types and phases (such as "investigation", "arrest", "court hearing") in a complex storyline ("fire outbreak"). RDFLiveNews [M21] also follows *developing news* by combining statistical and other machine-learning techniques in order to represent news events as knowledge graphs in real time by extracting RDF triples from RSS data streams.

Summary: Our review shows that all the different phases of the news-life cycle are covered by the research, from predicting future news, through detecting and monitoring emerging, breaking, and developing news, to managing and exploiting already published news. Many main papers attempt to cover several of these life-cycle phases.

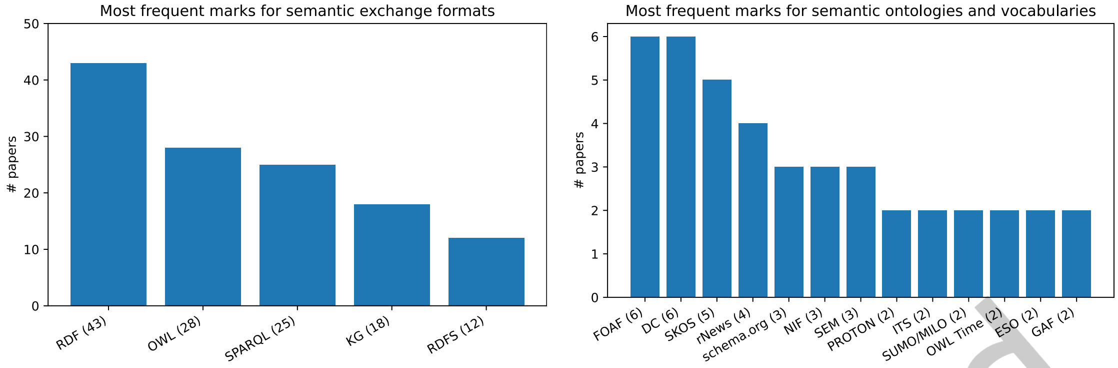
3.7 Semantic techniques and tools

The main papers use a broad variety of semantic techniques and tools. For the purpose of this review, we separate them into *exchange formats*, *ontologies and vocabularies*, *information resources*, and *processing* and *storage techniques*.

Semantic exchange formats: By semantic exchange formats we mean standards for exchanging and storing semantic data. As shown in Figure 4a, RDF, OWL, and SPARQL are most common. More than half of the papers use RDF to manage information. The earliest example is Neptuno [M7], which uses RDF to represent the IPTC's hierarchical subject reference system[A70]. More than a third of the main papers use OWL for ontology representation. For example, the MediaLoep data model [M10] is represented in OWL (using the SKOS vocabulary), and its concepts are linked to standard knowledge bases like DBpedia [5] and GeoNames[A51]. And we have already mentioned the NEWS Ontology [M17], which is represented in OWL-DL[A118], the description logic subset of OWL. SPARQL is also common. It is central in the Hermes project [M4] and in the News Articles Platform [M53]. RDFS[A122] is also widely used, including in the NEWS [M15] project.

Ontologies and vocabularies: By ontologies and vocabularies we mean formal terminologies for semantic information exchange. As shown in Figure 4b, *Dublin Core (DC)* and *Friend of a Friend (FOAF)* are the most used general vocabularies, starting already with KIM [M37], whose ontology is designed to be aligned with them both. The NEWS project [M15] also uses DC, whereas FOAF plays a prominent role in a few approaches that deal with personalisation, in particular in a social context [M44,M28]. Another much used ontology is the *Simple Knowledge Organization System (SKOS)*. It is used by the NEWS Ontology [M17] to align and interoperate concepts from different annotation standards, including the IPTC News Codes[A68]. It is also used for personalised multimedia recommendation in [M26], and for integrating news-production metadata in [M10]. The OWL-representation of IPTC's News Codes in [M71] links to Dublin Core and SKOS concepts to increase precision and facilitate content enrichment. The *Simple Event Model (SEM)*[A111] and *OWL Time*[A119] are used in the NewsReader [M72] and News Angler projects [M46]. *SUMO/MILO*[A97] is used in the NEWS project [M15]. SUMO and *ESO*[A99] are used in NewsReader [M72,M59]. Other general ontologies include *schema.org*, used to contextualise the ClaimsKG [M69] and KIM's PROTON ontology[A1] [M37]. The *Provenance Data Model (PROV-DM)* is used to discover high-level provenance using semantic similarity in [M9]. Although several other papers mention provenance too, they do not explicitly refer to or use PROV-DM, nor its OWL formulation, PROV-O[A120]. However, the NewsReader project [M72] uses the Grounded Representation and Source Perspective (GRaSP) framework, which has at least been designed to be compatible with PROV-DM.

On the news side, the *rNews* vocabulary[A69] is used for semantic mark-up of web news resources in several papers. Whereas most of the papers in this review rely on the older versions of rNews, the ASRAEL project [M60] uses its newer schema.org-based rNews vocabulary. The *Internationalization Tag Set (ITS)*[A116] is also used in a



(a)

(b)

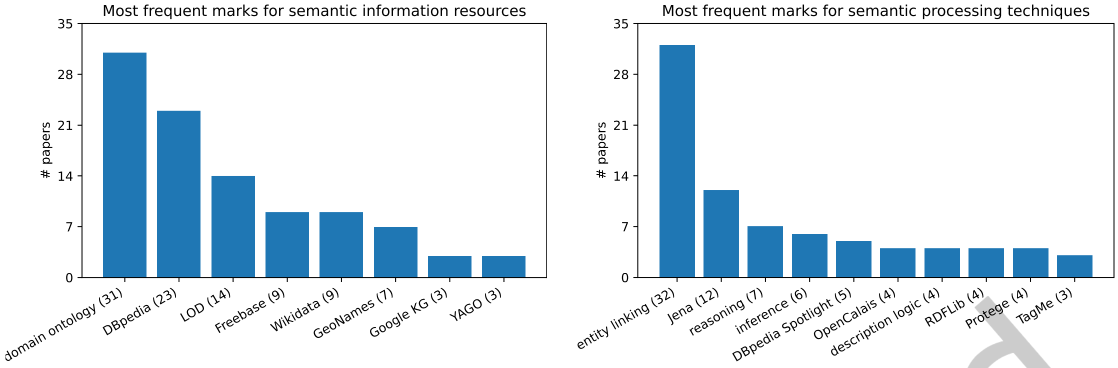
Fig. 4. The most frequently used semantic (a) exchange formats and (b) vocabularies and ontologies.

few papers, for example to unify claims in ClaimsKG [M69], The IPTC's *General Architecture Framework (GAF)*[A66] is used in NewsReader [M72,M59].

On the natural-language side, several proposals [M69,M72,M21] use the RDF/OWL-based *NLP Interchance Format (NIF)*[A87] to exchange semantic data between NLP components. In addition, more than a third of the papers propose their own domain ontologies.

Semantic information resources: By semantic information resources we mean open knowledge graphs, or openly available semantic datasets expressed as triples. As shown in Figure 5a, semantic encyclopedias are most frequently used. More than a quarter of the main papers somehow exploit *DBpedia*. It is, for example, used by NewsReader [M72,M59] for semantic linking and enrichment. *Wikidata* is an alternative that is used in several recent approaches. It is used by ASRAEL [M60] and VLX-Stories [M14] to support semantic labelling, enrichment and search, and it used to detect fake news in [M5]. There is also increasing uptake of *Google's KG*, which is used by VLX-Stories [M14] to detect emerging entities, in [M61] to separate emerging from already-known entities, and by AGV [M13] to provide additional information about entities extracted from educational TV programs. Although it has been seeded into Google's knowledge graph[A104] and is no longer maintained, *Freebase* is still being used for external linking in K-Pop [M36], for content-based recommendation in [M33], for evaluation of TAMURE [M78], and for enriching government data in [M62] (more later). *GeoNames* is used as the reference graph for geographical information in many papers, such as [M44,M71,M46,M72,M62,M10,M63]. With the availability of large one-stop KGs like these, fewer papers than before rely on the LOD cloud in general. An exception is [M26], which exploits the LOD cloud to identify news stories that match users' interests.

Beyond general semantic encyclopedias and other LOD resources, *YAGO2*[A58] and its integration of WordNet event classes is used in [M38] to classify named news events. The initial version of YAGO is used by Pundit [M56] to build a world entity graph for mining causal relationship between news events, and in [M74] to infuse world knowledge into a Knowledge-driven Multimodal Graph Convolutional Network (KMGCN) for fake news detection. Common-sense knowledge from the *Cyc* project [32] is used too, for example to augment reasoning over semantic representations mined from financial news texts [M51] and to predict future events [M56]. PolarisX [M77] uses *ConceptNet 5.5*[A106,A107] as a development case and for evaluating its approach to automatically expand knowledge graphs with new news events. ConceptNet is also used by Pundit [M56]. Several of the general semantic information resources, such as DBpedia, ConceptNet, Cyc, Wikidata, and YAGO, come with their own resource-specific ontologies and vocabularies in addition to the ones mentioned in the previous section.



(a) (b)

Fig. 5. The most frequently used semantic (a) information resources and (b) processing techniques.

On the natural language side, *WordNet* [18, 37] is not natively semantic, but it is used in a third of the main papers — more than any of the natively semantic resources — including Hermes [M4], NewsReader [M72,M59], and SPEED [M24] — although only a single paper ([M58]) explicitly mentions WordNet's RDF version[A125].

Semantic processing techniques: By semantic processing technique we mean programming techniques and tools used to create and exploit semantic information resources. As shown in Figure 5b, entity linking [2] is the most frequently used technique by far. The most used entity linkers are DBpedia Spotlight, Thomson-Reuters' OpenCalais[A109], and TagMe. Beyond entity linking, seven papers employ logical reasoning. Description logic[A82] and OWL-DL is used for trust-based resolution of inconsistent KBs in [M22] and for managing the NEWS Ontology [M17]. Other papers mention general ontology-enabled reasoning without OWL-DL. For example, PWFF [M28] and [M51], which uses Cyc to answer questions about business news. Rule-based inference is also used, e.g., in [M57,M6,M20,M15,M35].

The most common programming API for semantic data processing is Apache's Java-based Jena framework[A44], used in 12 main papers. Only four papers mention Python's RDFlib[A11], most of them from recent years. Protégé[A112] is used in four papers, for example for ontology development in Neptuno [M7] and in the NEWS project [M15].

Semantic storage techniques: Although almost all the main papers mention ontologies or knowledge graphs, few of them discuss storage and none of them focus primarily on the storage side. The two most frequently used triple stores are RDF4J[A39] (formerly Sesame) used by four papers and OpenLink Virtuoso[A105] also used by four papers. AllegroGraph[A65] is employed in two papers [M44,M49]. Used by NewsReader [M72,M59], the KnowledgeStore[A28] is designed to store large collections of documents and link them to RDF triples that are extracted from the documents or collected from the LOD cloud. It uses a big-data ready file system (Hadoop Distributed File System, HDFS[A103]) and databases (Apache HBase[A43] and Virtuoso[A105]) to store unstructured (e.g., news articles) and structured information (e.g., RDF triples) together.

Summary: Our review demonstrates that the research on KG for news exploits a broad variety of available semantic resources, techniques, and tools. The research on KGs for news differs from the mainstream research on KGs mainly in its stronger focus on language (e.g., the ITS and NIF vocabularies), on events (e.g., the SEM ontology) and, of course, on news (the rNews vocabulary). The border between semantic and non-semantic computing techniques is not always sharp. For example, although WordNet is not natively semantic, it is available as RDF, and is used as a semantic information resource in many proposals. A recent trend is that Wikidata is becoming more popular compared to DBpedia.

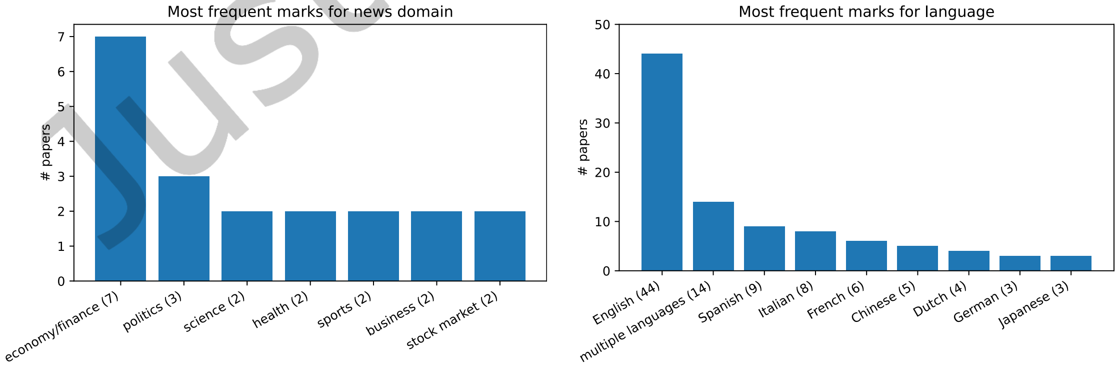
3.8 Other techniques and tools

Most main papers use semantic knowledge graphs in combination with other techniques and tools. Similar to the previous section, we separate them into *exchange formats*, *information resources*, and *processing* and *storage techniques*. The online Addendum[2] (Section B.1) presents a detailed review, which shows that the research on semantic knowledge graphs for the news is technologically diverse. We find examples of research that exploits most of the popular news-related standards and most of the popular techniques for NLP, machine learning, deep learning, and computing in general.

On the news side, the IPTC family of standards and resources[A72,A71,A66] is central. On the NLP side, entity extraction, NL pre-processing, co-reference resolution, morphological analysis, and semantic-role labelling are common, whereas GATE[A92], Lucene[A45], spaCy[A35], JAPE[A93] and StanfordNER[A57] are the most used tools.

On the ML side, the last decade has seen more and more proposals that exploit machine-learning techniques, as illustrated by three early examples from 2012: [M9] uses greedy clustering to automatically detect provenance relations between news articles. The Hermes framework [M29] uses a pattern-language and rule-based approach to learn ontology instances and event relations from text, combining lexico-semantic patterns with semantic information. It is used to analyse financial and political news articles, splitting its corpus of news articles into a training and a test set. Pundit [M56] mines text patters from news headlines to predict potential future events based on textual descriptions of current events. It uses machine learning to automatically induce a causality function based on examples of causality pairs mined from a large collection of archival news headlines. Whereas these early approaches rely on hand-crafted rules and dedicated learning algorithms, more recent proposals use standard machine-learning techniques for word, graph and entity embeddings, such as TransE [9], TransR [33], TransD [30], and word2vec [36].

On the DL side, there has been a sharp rise since 2019 in deep learning [22] approaches. PolarisX [M77] uses pre-trained multilingual BERT model to detect new relations, with the aim of updating its underlying knowledge graph in real time. TAMURE [M78] uses tensor factorisation implemented in TensorFlow[A13] to learn joint embedding representations of entities and relation types. Focusing on click-through rate (CTR) prediction in online news sites, DKN [M73] uses a Convolutional Neural Network [22] with separate channels for words and entities and an attention module to dynamically aggregate user histories. [M19] proposes a deep neural network model that employs multiple self-attention modules for words, entities, and users for news recommendation. [M52] proposes the B-TransE model to detect fake news based on content. The most used deep learning techniques and tools are CNN [22], GRU [22], GCN [11, 23], LSTM [22], BERT [16], and attention [55] are much used.



(a)

(b)

Fig. 6. The most frequently targeted (a) news domains and (b) languages.

The focus on news standards is strongest in the first part of the study period, up to around 2014, when many approaches incorporate existing news standards into the emerging LOD cloud [7]. The second part, from around 2015, sees a shift towards machine learning approaches [39], first focusing on NLP and embedding techniques and, since around 2019, on deep learning [22].

3.9 News domain

As shown in Figure 6a, most of the main papers do not focus on a particular news domain, except as examples or for evaluation. Among the domain-specific papers, *economy/finance* is most common. For example, [M43] presents a semantic search engine for financial news that uses an automatically populated knowledge graph which is kept up to date with semantically annotated financial news items, and we have already mentioned the SPEED pipeline for economic event detection and annotation in real time [M24].

Politics is the theme of over 900.000 French tweets collected in order to trace propagation of knowledge, misinformation, and hearsay [M12]. [M62] uses named entity linking to contextualise open government data, making them available in online news portals alongside related news items that match each user's interests. In the *sports* news domain, [M50] proposes a recommender system based on BKSport [M49] that combines semantic and content-based similarity measures to suggest relevant news items. Other domains targeted by multiple papers include *science* [M57,M13], *business* [M51,M41], *health* [M13,M23], and the *stock market* [M41,M4].

Targeting *entertainment* news, K-Pop [M36] builds on an entertainer ontology to compile a semantic knowledge graph that represents the profiles and activities of Korean pop artists. The artists' profiles in the graph are based on information from Wikipedia[A49] and enriched with content from DBpedia [5], Freebase[A2], LinkedMDB[A6], and MusicBrainz[A40]. They are also linked to other sources that represent not only the artist, but also their activities, business, albums, and schedules. The graph and ontology are used in the Gnosis app to enhance K-Pop entertainment news with information about artists retrieved from the knowledge graph.

WebLyzard [M63] identifies topics and entities in news about the *environment* and uses visualisations to present lexical, geo-spatial, and other contextual information to gain overview of perceptions of and reactions to environmental threats and options. AGV [M13] targets *education*, in particular in science and technology. Other domains include *medicine* [M23], *crime* [M67] and *earth science* [M57].

Summary: Our review suggests that semantic knowledge graphs and related semantic techniques are useful in a broad range of news domains. Most investigated so far is economy and finance. There is little domain-specificity in the research so far: most architectures and techniques proposed for one news domain appear readily transferable to others. The higher interest in the financial and business domains may result from economic opportunities combined with the availability of both quantitative and qualitative data streams in real time.

Table 2. The five most frequently cited main papers (recency weighted).

Title	**Year**	**Ref**	**# citations**	**Citation weight**	**# main paper citations**
DKN: Deep knowledge-aware network for news recommendation	2018	[M73]	413	33.15	4
Semantic annotation, indexing, and retrieval	2004	[M37]	523	16.34	1
Learning causality for news events prediction	2012	[M56]	199	9.64	2
Building event-centric knowledge graphs from news	2016	[M59]	111	7.35	2
Content based fake news detection using knowledge graphs	2018	[M52]	72	5.78	1

Table 3. The five papers most frequently referenced by our main papers.

Title	Year	Ref	# main paper refs
The semantic web	2001	[6]	15
WordNet: An electronic lexical database	2000	[18]	11
Translating embeddings for modeling multi-relational data	2013	[9]	8
GATE, a general architecture for text engineering	1997	[15]	7
Distributed representations of words and phrases and their compositionality	2013	[36]	7

3.10 Language

As shown in Figure 6b, the most frequently covered languages beside English are Italian, and Spanish, but neither is supported by more than ten papers. Support for French and German in the main papers appear only in combination with English. Many papers deal with a combination of *several languages*, such as English, Italian, and Spanish in the NEWS project [M15], and a few recent approaches explicitly aim to be *multi-lingual* (or *language-agnostic*). For example, NewsReader [M72] mentions Dutch, Italian, and Spanish in addition to English, whereas PolarisX [M77] aims to cover Chinese, Japanese, and Korean.

Summary: Our review suggests that English is the best supported language by far but, of course, this may be because we use English language as an inclusion criterion. Additional papers addressing other major languages, such as Chinese, French, German, Hindi, and Spanish, may instead be written and published in their own languages. The other most frequently supported languages are Spanish, Italian, French, Chinese, Dutch, German, and Japanese, with many of the Chinese and Japanese papers published in recent years. Many approaches also support more than one language, exploiting the inherent language-neutrality of ontologies and knowledge graphs. There is a growing interest in offering multi-language and language-agnostic solutions.

3.11 Important papers

Our main papers reference 1842 earlier papers and are themselves cited 2381 times according to Semantic Scholar[A36]. Table 2 shows that the most cited of our main papers is the one about KIM [M37] from 2004, with the much more recent DKN paper [M73] from 2018 second. The paper about Pundit [M56] from 2012 is also frequently cited. To account for recency, Table 2 therefore ordered by citation numbers that are weighted against the expected number of citations of a main paper from the same year.[4] Just outside the top five, [M69,M72,M38,M21,M24] are also frequently cited.[5]

Table 3 shows the papers that are referenced most frequently by our main papers.[6] Among the outgoing citations, seminal papers on the Semantic Web [6] and on WordNet [18] are most frequently cited. Also much cited are the central papers on GATE, TransE, and word2vec. Just outside the top five, other frequently referenced papers are [5, 8, 33, 52, 56], confirming the importance of LOD resources and embedding models for the research on semantic KGs for news.[7] Closely related to our main papers, another paper on KIM [48] is cited 6 times, and [29], a precursor to the SemNews paper [M30], is also cited several times.

[4] To weight the citation counts, the three most frequently cited papers ([M37,M73,M56]) are removed as outliers. Average citation counts are calculated for each year for the remaining main papers. A support-vector regression (SVR) model is trained using scikit-learn [39] with a radial-basis function (RBF) kernel, $C = 1000$, and $\gamma = 0.001$. Finally, the citation count for each paper is divided by the count predicted for a paper from that year.

[5] The online *Addendum*[2] (Table 11) presents an extended top-15 list.

[6] We do not report weighted reference counts, because more recent papers are much more frequently cited in our dataset, giving unreasonably high relative weight to older papers even when they are referenced only once or a few times.

[7] The online *Addendum*[2] (Table 12) again presents an extended top-15 list.

Table 4. (a) Authors with three main papers or more and (b) projects with multiple papers.

Author	Main papers	# main papers
F. Frasincar	[M58,M4,M29,M25,M64,M24,M32]	7
F. Hogenboom	[M58,M29,M25,M64,M24]	5
D. Deursen, E. Mannens, R. Walle	[M44,M9,M10]	3
N. García, L. Sánchez	[M16,M17,M15]	3

(a)

Project	# papers	# citations	# main citations
Hermes	7	188	2
NewsReader	2	171	2
"Wuhan"	2	73	0
NEWS	3	73	4
"MediaLoep"	3	45	0
"Chicago"	2	8	0
BKSport	2	8	0

(b)

Table 5. Groups of authors connected by chains of two or more co-authored papers.

Authors	Refs	Project
J. Borsje, F. Frasincar, F. Hogenboom, L. Levering, K. Schouten	[M58,M4,M29,M25,M64,M24,M32]	Hermes
S. Coppens, D. Deursen, E. Mannens, R. Sutter, R. Walle	[M44,M9,M10]	"MediaLoep"
J. Arias-Fisteus, A. Bernardi, N. García, L. Sánchez, J. Toro	[M16,M17,M15]	NEWS
N. Li, C. Li, J. Pan	[M52,M42]	"Wuhan"
Y. Chang, C. Lu, J. Zhang	[M79,M78]	"Chicago"
T. Cao, Q. Nguyen	[M49,M50]	BKSport
I. Aldabe, M. Erp, A. Fokkens, G. Rigau, M. Rospocher, P. Vossen	[M59,M72]	NewsReader

Summary: No paper yet stands out as seminal for the research area. With the exception of the KIM project, none of the main papers or projects are frequently cited by other main papers, suggesting that research on semantic KGs has not yet matured into a clearly defined research area that is recognised by the larger research community.

3.12 Frequent authors and projects

Table 4a shows the most frequent main-paper authors, along with their most centrally related projects. Table 5 also shows co-authorship cliques, defined by chains of at least two co-authored papers. The table shows that repeated co-authorship among the frequent authors occurs exclusively within a small number of research projects (or persistent collaborations), such as NEWS [M15], Hermes [M4], and NewsReader [M72].[8]

The seven cliques cover all the repeated collaborations we have found. Table 4b also shows the cumulative citation counts for each project or collaboration. Hermes and NewsReader are the most frequently cited projects, but there are very few citations to these and to the other projects/collaborations from other main papers. Indeed, none of the main papers from the seven listed projects and collaborations are citing one another (although, of course, such cross-references may still exist between papers from the same projects that we have not included as main papers). Only 43 references in total are from one main paper to another (the online *Addendum*[2] (Figure 10) presents a citation graph).

Summary: The analysis underpins that the research on semantic KGs has not yet matured into a distinct research area. The research is carried out mostly by independently working researchers and groups, although the NewsReader project has involved several institutions located in different countries. There is so far little

[8] Table 5 also introduces informal names such as "MediaLoep" and "Wuhan" for persistent collaborations that are not centred around a single named project.

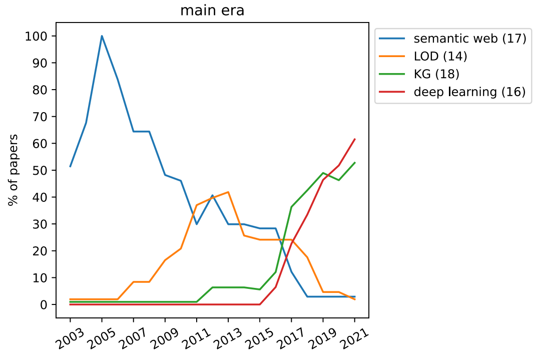


Fig. 7. Timeline for the percentages of main papers from each year that match the sub-themes Semantic Web, LOD, KGs and deep learning.[9]

collaboration and accumulation of knowledge in the area, although the early KIM [M37] proposal has been used in later research.

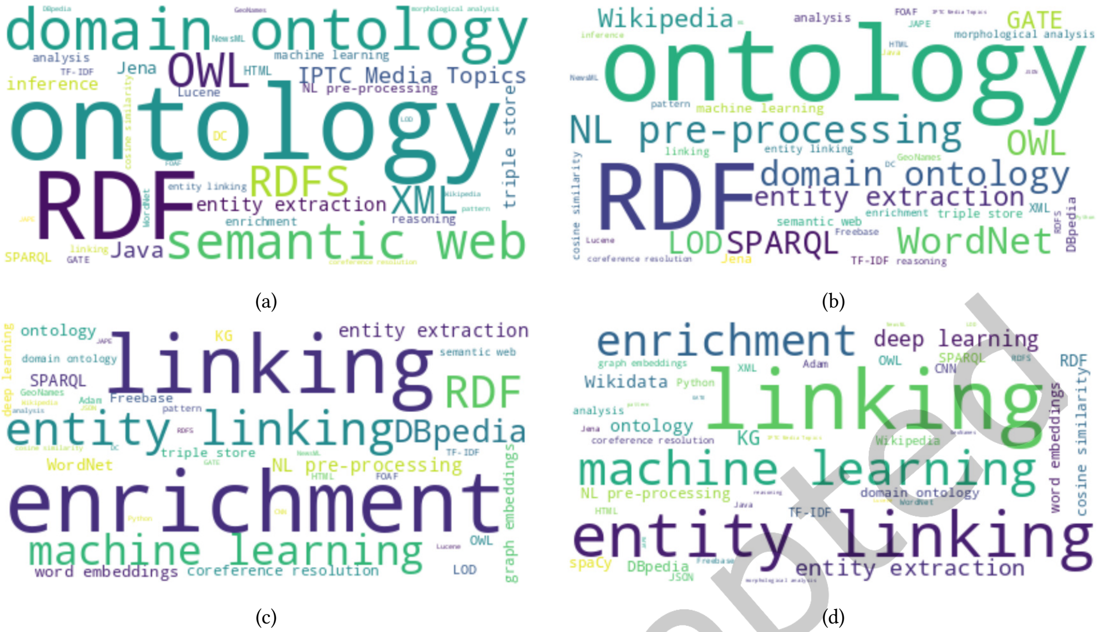
3.13 Evolution over time

The research on semantic knowledge graphs for news can be divided into four *eras* that broadly follow the evolution of knowledge graphs and related technologies in general: the Semantic Web (–2009), Linked Open Data (LOD, 2010–2014), knowledge graphs (KGs, 2015–2018), and deep-learning (DL, 2019–) eras. Figure 7 presents corresponding timelines that show the percentage of main papers from each year that match each theme. To underpin the separation into four eras further, the online Addendum[2] (Section B.2) presents additional timelines that show typical sub-themes from each era.

The first era (until around 2009) is inspired by the *Semantic-Web* idea and early ontology work. Almost all the main papers from this era mention the Semantic Web or Semantic-Web technologies prominently in their introductions. They combine basic natural-language processing with central Semantic-Web ideas such as semantic annotation, domain ontologies, and semantic search applied to the news domain. Many of the papers bring existing news and multimedia publishing standards into the Semantic-Web world, and the IPTC Media Topics[A67] are therefore important. Central semantic techniques are RDF, RDFS, OWL and SPARQL, and important tasks are archiving and browsing. There is also an early interest in multimedia. Figure 8a shows a word cloud of the most prominent sub-themes for papers published during this era.

The main papers in second era (2010–2014, but starting with [M71] already in 2008) trails the emergence of the Linked Open Data (LOD) cloud [7], which many of the papers use to motivate their contributions. Contextualisation and other types of semantic enrichment of news texts is central, aiming to support more precise search and recommendation. Although some papers use Wikipedia and DBpedia for enrichment, the most used information resource is WordNet. In order to link news texts precisely to existing semantic resources, more advanced pre-processing of news texts is used along with techniques such as morphological analysis and vector spaces. GATE is a much used NLP tool in this era, as is OpenCalais for entity linking, and Jena for managing RDF data.

The third era (2015–2018), reflects Google's adoption of the term "Knowledge Graphs" in 2012[A104] and the growing importance of machine learning [39]. One of the first main papers to mention knowledge graphs is [M2] already in 2013, but most of the main papers are published starting in 2015. The research increasingly considers knowledge graphs independently of semantic standards like RDF and OWL, and uses machine learning and



(a)

(b)

(c)

(d)

Fig. 8. Word clouds for (a) the Semantic-Web era (until around 2009) and (b) the Linked Open Data (LOD) era (around 2010–2014) (c) the knowledge-graph era (around 2015–2018), and (d) the deep-learning era (from around 2019).

related techniques to analyse news texts more deeply, for example extracting events and facts (relations). DBpedia and entity linking become more frequently used, along with word and graph embeddings. On the NLP side, co-reference resolution and dependency parsing become more important, along with StanfordNER.

Since around 2019, a fourth and final era starts to emerge. Typical approaches analyse news articles using deep neural network (NN) architectures that combine text- and graph-embedding approaches and that infuse triples from open KGs into graph representations of news texts. Central emerging tasks are fact checking, fake-news detection, and click-through rate (CTR) prediction. Deep-learning techniques like CNN, LSTM, and attention become important, and spaCy is used for NLP. On the back of deep image-analysis techniques, multimedia data also makes a return. Because the boundary between this and the KG era is not sharp, the word cloud in Figure 8d has many similarities to Figure 8c.

4 DISCUSSION

Based on the analysis, this section will discuss each main theme in our analysis framework (Table 1). We will then answer the four research questions posed in the Introduction and discuss the limitations of our paper.

4.1 Conceptual framework

Table 6 shows the conceptual framework that results from populating our analysis framework in Table 1 with the most frequently used sub-themes from the analysis. It is organised in a hierarchy of depth up to 4 (e.g., **Other techniques**: → Other resources → *language* → WordNet). The framework shows which areas and aspects of semantic knowledge graphs for news that have so far been most explored in the literature. It can be used both as an overview of the research area, as grounds for further theory building, and as a guide for further research.

Table 6. Conceptual framework.

Technical result type:	pipeline/prototype (58), ontology (18), system architecture (16), algorithm (15), NN architecture (9), KG (7), production system (4)
Empirical result type:	experiment (58), examples (17), ablation study (7), PoC demo (6), performance evaluation (5), case study (4), parameter study (4), explainability study (4), user study (4), use case (3), industrial testing (3)
Intended users:	news users (45), journalists (32), KB maintainers (10), newsrooms (9), fake news detectors (8), knowledge workers (6), archivists (5), fact checkers (3), news agencies (3)
Task:	semantic annotation (28), retrieval (22), event detection (20), provision (19), enrichment (18), relation extraction (9), KG updating (9), ontology development (8), KG population (7), fake news detection (7), personalisation (6), archiving (6), sub-graph extraction (5), prediction (5), content generation (5), visualisation (4), similarity detection (3), fact checking (3), integration/interoperability (3)
Input data:	news articles (59), news feeds (23), RSS feeds (17), KG (11), social media (10), multimedia (8), Twitter (6), TV news (4), user histories (4), news metadata (3)
Life-cycle phase:	published news (58), developing news (13), breaking news (7), emerging news (3), future news (1)
Semantic techniques:	Semantic exchange formats: *RDF (43)*, *OWL (28)*, *SPARQL (25)*, *KG (18)*, *RDFS (12)* • Semantic ontologies and vocabularies: *FOAF (6)*, *DC (6)*, *SKOS (5)*, *rNews (4)*, *NIF (3)*, *schema.org (3)*, *SEM (3)*, *PROTON (2)*, *ITS (2)*, *SUMO/MILO (2)*, *OWL Time (2)*, *ESO (2)*, *GAF (2)* • Semantic information resources: *domain ontology (31)*, *DBpedia (23)*, *LOD (14)*, *Freebase (9)*, *Wikidata (9)*, *GeoNames (7)*, *Google KG (3)*, *YAGO (3)*, *OpenCyc (2)*, *ConceptNet (2)* • Semantic processing techniques: *entity linking (32)*, *Jena (12)*, *reasoning (7)*, *inference (6)*, *DBpedia Spotlight (5)*, *OpenCalais (4)*, *description logic (4)*, *RDFLib (4)*, *Protege (4)*, *TagMe (3)* • Semantic storage techniques: *Virtuoso (4)*, *RDF4J/Sesame (4)*, *AllegroGraph (2)*, *KnowledgeStore (2)*
Other techniques:	Other exchange formats (general): *XML (13)*, *HTML (10)*, *JSON (5)*, *MPEG-7 (4)*, *CSS (3)*, *XML Schema (2)* • Other exchange formats (news): *NewsML (6)*, *NITF (2)*, *NAF (2)* • Other resources (general): *Wikipedia (14)*, *Twitter (4)*, *Yahoo! Finance (2)*, *ISO country codes (2)*, *CIA WorldFact Book (2)*, *NASDAQ company codes (2)* • Other resources (news): *IPTC Media Topics (7)*, *IPTC NewsCodes (2)* • Other resources (language): *WordNet (22)*, *VerbNet (4)*, *Penn Treebank (2)*, *Predicate Matrix (2)*, *PropBank (2)*, *FrameNet (2)* • Other processing techniques (language): *entity extraction (36)*, *NL preprocessing (33)*, *coreference resolution (11)*, *GATE (10)*, *Lucene (7)*, *spaCy (7)*, *JAPE (6)*, *morphological analysis (6)*, *StanfordNER (5)*, *SRL (5)*, *WSD (4)*, *relation extraction (4)*, *dependency parsing (4)*, *sentiment analysis (4)*, *StanfordNLP (4)* • Other processing techniques (machine learning/deep learning): *word embeddings (13)*, *graph embeddings (8)*, *CNN (6)*, *TransE (5)*, *GRU (4)*, *hierarchical clustering (3)*, *GCN (3)*, *word2vec (3)*, *attention (3)*, *TransR (3)*, *LSTM (3)*, *BERT (3)*, *entity embeddings (3)* • Other storage techniques: *MongoDB (2)*, *MySQL (2)*, *relational DB (2)*, *Heuristic and Deductive Database (2)*
News domain:	economy/finance (7), politics (3), science (2), health (2), sports (2), business (2), stock market (2), earth science (1), technology (1), crime (1), evolutionary events (1), medicine (1), entertainment (1), climate change (1), environment (1)
Language:	English (44), multiple languages (14), Spanish (9), Italian (8), French (6), Chinese (5), Dutch (4), German (3), Japanese (3)

The earliest versions of our framework also contained geographical *region* as a top-level theme, alongside news domain and language, but very few of our main papers were specific to a region, and never exclusively so. For example, although the contextualisation of open government data in [M62] focuses on *Colombian* politics, the proposed solution is straightforwardly adaptable to other regions.

4.2 Implications for practice

For each main theme, this section suggests implications for practice, before the next section proposes paths for further research.

Technical result types: There are already many tools and techniques available that are sufficiently developed to be tested in industrial workflows. Commercial tools like VLX-Stories [M14] and ViewerPro[A102] are also starting to emerge. But most research proposals are either research pipelines/prototypes or standalone components that require considerable effort to integrate into existing workflows before they can become productive. Pilot projects that match high-benefit tasks with low-risk technologies and tools are therefore essential to successfully introduce semantic KGs in newsrooms.

Empirical result types: Although there are examples of tools and techniques that have been deployed in real news production workflows, they are the exception rather than the rule. This poses a double challenge for newsrooms that want to use KGs for news: it is usually not known how robust the proposed techniques and tools are in practice and it is usually not known how well they fit actual industrial needs. Introducing KGs into newsrooms must therefore focus on continuous evaluation both of the technology itself and of its consequences, opening possibilities for collaboration between industry (which wants its projects evaluated) and researchers (who want access to industrial cases and data).

Intended users: The most mature solutions support journalists through tools and techniques for searching, archiving, and content recommendation. The general news user is supported by proposals for news recommendation and to some extent searching.

Tasks: The most mature research proposals target long-researched tasks like semantic annotation, searching and recommendation, both for content retrieval (pull) and provision (push). In particular, annotation of news texts with links to mentioned entities and concepts is already used in practice and will become even more useful as the underlying language models continue to improve. Semantic searching and browsing are also well-understood areas. Semantic enrichment with information from open KGs and other sources is a maturing area that builds on a long line of research, but suffers from the danger of creating information overload. Rising areas that are becoming available for pilot projects are automatic news detection, and automatic provision of background information.

Input data: The most mature tools and techniques are text-based. When multimedia is supported, it is often done indirectly by first converting speech to text or by using image captions only. Newer approaches that exploit native audio and image analysis techniques in combination with semantic KGs may soon become ready for industrial trials. Many newsrooms already have experience with robots [38] that exploit input data from sensors, the Internet of Things (IoT), and open APIs [4]. This creates opportunities to explore new uses of semantic KGs that augment existing robot-journalism tools and techniques. Much of the research that exploits social media is based on Twitter. This poses a challenge, because Twitter-use is dwindling in some parts of the world, sometimes with traffic moving to more closed platforms, such as Instagram, Snapchat, Telegram, TikTok, WhatsApp, etc. In response, news organisations could attempt to host more social reader interactions inside their own distribution platforms, where they retain access to the user-generated content. Semantic KGs offer opportunities through their support for personalisation, recommendation, and networking.

News life cycle: Low-risk starting points for industrial trials are the mature research areas based on already published news, such as archive management, recommendation, and semantically-enriched search. Automated detection of emerging news events and live monitoring of breaking news situations are higher-risk areas that also offer high potential rewards.

Semantic techniques: Because they tend to rely on standard semantic techniques, many of the proposed techniques can be run in the cloud, for example in Amazon's Neptune-centric KG ecosystem[A17] and supported by other Amazon Web Services for NLP and ML/DL[A16]. Cloud infrastructures give newsrooms a way to explore advanced computation- and storage-intensive KG-based solutions without investing heavily upfront in new infrastructure.

Other techniques: The demonstrated ability of KG-based approaches to work alongside a wide variety of other computing techniques and tools suggest that newsrooms that want to exploit semantic KGs should build on what they already have in place, using KG-based techniques to augment existing services and capabilities. For example, KGs are well suited to integrate diverse information sources through exchange standards such as RDF and SPARQL and ontologies expressed in RDFS and OWL. One possibility is therefore to introduce them in newsrooms as part of ML and DL initiatives that need input data from multiple and diverse sources, whether internal or external. Semantic analysis of natural language texts, audio, images, and video is rapidly becoming available as increasingly powerful commodity services. KGs in newsrooms could be positioned to enrich and exploit the outputs of such services, acting as a hub that can represent and integrate the results of ML- and DL-driven analysis tools and prepare the data for journalists and others.

News domain: For newsrooms that want to exploit KGs, the most mature domains are business and finance. For example, ViewerPro[A102], an industrial tool for ontology-based semantic analysis and annotation of news texts, has been applied to gain effective access to relevant finance news. The proposed tools and techniques are often transferable across domains and purposes. Good candidates for industrial uptake are domains that are characterised by data streams that are reliable and high-quality, but insufficiently structured for currently available tools, e.g., for robot journalism [38]. Using KG-techniques to expand the reach and capabilities of existing journalistic robots may be a path to reap quick benefits from KGs on top of existing infrastructures.

Language: Given the focus on English in the research on semantic KG for the news and on NLP in general, international news is a natural starting point for newsrooms in non-English speaking countries that want to explore KG-based solutions. For newsrooms in English and other major-language countries, KG-powered cross-lingual and language-agnostic services can be used to simplify searching, accessing, and analysing minor-language resources, offering a low-effort/high-reward path to introducing semantic KGs.

4.3 Implications for research

Based on our analysis of main papers, this section proposes paths for further research.

Technical result types: More industrial grade prototypes and platforms are needed in response to the call for industrial testing. Much of the current research, such as the exploration of deep learning and other AI areas for news purposes, is technology-driven and needs to be balanced by investigations of the needs of journalists, newsrooms, news users, and other stakeholders.

Empirical result types: To better understand industrial needs, challenges, opportunities, and experiences, empirical studies are called for, using the full battery of research approaches, including case- and action-research, interview- and survey-based research, and ethnographic studies of newsrooms. Research on semantic knowledge graphs for the news might benefit from the growing and complementary body of literature on augmented, computational, and digital journalism (e.g., [12, 17, 49, 54]), which focuses on the needs of newsrooms and journalists, but goes less into detail about the facilitating technologies, whether semantic or not. Indeed, the research on semantic KGs for the news hardly mentions the literature on augmented/digital/data journalism which, vice versa, does not go into the specifics of KGs.

Most papers that propose new techniques or tools offer at least some empirical evaluation of their own proposals. Experimental evaluations using gold-standard datasets and information-retrieval measures are becoming increasingly common, but there is no convergence yet towards particular gold-standard datasets and measures, which makes it hard to compare proposals and assess overall progress. This is an important methodological challenge for further research. We also find no papers that focus on evaluating tools or techniques proposed by others. Also, the papers that develop pipelines and prototypes are seldom explicit about the design research method they have followed.

Intended users: We found no papers discussing semantic knowledge graphs and related techniques for citizen journalism, for example investigating social semantic journalism as outlined in [25]. Local journalism [42, 53] is also not a current focus, and we found few papers that explicitly mention newsrooms or consider the social and organisational sides of news production and journalism. There is also no mentioning of robot journalism in the main papers.

Tasks: More research is needed in areas that are critical for representing news content on a deeper level, beyond semantic annotation with named entities, concepts, and topics. Central evolving areas are event detection, relation extraction, and KG updating, in particular identification and semantic analysis of dark entities and relations.

There is little research on the quality of data behind semantic KGs for news. Aspects of semantic data quality, such as privacy, provenance, ownership, and terms of use need more attention. Few research proposals target or undertake multimedia analysis natively (i.e., without going through text) and specifically for news.

Input data: The research on social media tends to focus on short texts, which are hard to analyse because they provide less context and use abbreviations, neologisms, and hashtags [51]. More context can be provided by integrating newer techniques that also analyse the audio, image, and video content in social messages. Some research approaches harvest citizen-provided data from social media, but there are no investigations of how to use semantic techniques and tools participatively for citizen journalism [25]. There is little research on KGs for news that exploits data from sensors and from the IoT in general [4], and there is little use of open web APIs outside a few domains (such as business/finance). We have already mentioned the ensuing possibility of combining semantic KGs with robot-journalism tools and techniques. GDELT[A98] is another untapped resource, although data quality and ownership is an issue. Research is needed on how its data quality can be corroborated and improved. Also, the low-level events in GDELT data streams need to be aggregated into news-level events.

News life cycle: Relatively little research target detecting emerging news events, monitoring breaking news situations, and following developing stories. Event detection and tracking as well as detecting emerging entities and relations are important research challenges.

Semantic techniques: Most of the research uses existing news corpora or harvests news articles on-demand. There is less focus on building and curating journalistic knowledge graphs over time. Due to the high volume, velocity, and variety of news-related information, semantic news KGs are a potential driver and test bed for real-time and big-data semantic KGs. More research is therefore needed on combining KGs with state-of-the-art techniques for real-time processing and big data. Yet none of the main papers have primary focus on the design of semantic data architectures/infrastructures for newsrooms, for example using big-data infrastructures, data lakes, web-service orchestrations, etc. The most big-data ready research proposal is NewsReader, through its connection with the big-data ready KnowledgeStore[A28] repository. The *News Hunter* platform developed in the News Angler project [M46] is also built on top of a big-data ready infrastructure [20]. In addition to supporting processing of big data in real time, these architectures and infrastructures must be forward-engineered to accommodate the increasing availability of high-quality, high-performance commodity cloud-services for NLP, ML, and DL that can be exploited by news organisations.

Other techniques: On the research side, few approaches to semantic KGs for news exploit recent advances in image understanding and speech recognition. There is a potential for cross-modal solutions that increase precision and recall by combining analyses of text, audio, images and, eventually, video. These solutions need to be integrated with semantic KGs, and their application should focus on areas where KGs bring additional benefits, such as infusing world and common-sense knowledge into existing analyses. Also, few approaches so far exploit big-data and real-time computing. Although some proposals express real-time ambitions, they are seldom evaluated on real-volume and -velocity data streams and, when they are (e.g., RDFLiveNews [M21] and SPEED [M24]), they do not approach web-scale performance. Although the proposed research pipelines may not be optimised for speed, performance-evaluation results suggest that more efficient algorithms are needed, for example running on staged and parallel architectures. High-performance technologies for massively distributed news knowledge graphs are also called for, for example exploiting big-graph databases such as Pregel[A75] and Giraph[A42].

News domain: Whereas practical applications of KGs may be driven by economical (for economy/finance) and popular (e.g., for sports) interests, there is ample opportunity on the research side for adapting and tuning existing approaches to new and unexplored domains that have high societal value. One largely unexplored domain is corruption and political nepotism, along the lines suggested in [128]. Misinformation is another area of great importance, and in the domain of crises and social unrest, the GDELT data streams may offer opportunities.

Language: Research is needed to make semantic KGs for news available for smaller languages. There is so far little uptake of cross-language models like multi-lingual BERT and little research on exploiting dedicated language models for smaller languages for news purposes.

4.4 Research questions

We are now ready to answer the four research questions we posed in the Introduction.

RQ1: Which research problems and approaches are most common, and what are the central results? Our discussion in Section 4 and Table 6 answers this question for each of the main themes in our framework. The review shows that research on semantic knowledge graphs for news is highly diverse and in constant flux as the enabling technologies evolve. A frequent type of paper is one that develops new tools and techniques for representing news semantically in order to disseminate news content more effectively. In response to the increasing societal importance of information quality and misinformation, there is currently a rapidly growing interest in fake-news detection and fact checking. The tools and techniques are typically developed as pipelines or prototypes and evaluated using experiments, examples or use cases. The experimental methods used are maturing.

RQ2: Which research problems and approaches have received less attention, and what types of contributions are rarer? Our discussion in Section 4, and in 4.3 in particular, answers this question by identifying many under-researched areas. The review shows that there are very few industrial case studies. In our literature searches, we have found few surveys and reviews. There is also little research on issues such as privacy, ownership, terms of use, and provenance, although a few papers mention the latter. Only a few papers focus on evaluating their results in real-time and big-data settings and, when they do, the results are often in need of improvement. Other green-field areas include: exploiting location data and data from the Internet of Things, supporting social and citizen journalism, using semantic knowledge graphs to identify new newsworthy events as in Reuters Tracer[A73], and using semantic knowledge graphs to construct narratives and generate news content.

Although the results suggest that semantic knowledge graphs can indeed support better organisation, management, retrieval, and dissemination of news content, there is still a potential for much larger uptake in industry. Empirical studies are needed to explain why. One possible explanation is that there is a mismatch between what the current tools and algorithms offer and what the industry needs. Another possible explanation is that the solutions themselves are immature, for example that existing analysis techniques are not sufficiently precise or that the often crowd-sourced reference and training data used are perceived as less trustworthy.

RQ3: How is the research evolving? Our analysis in Section 3.13 answers this question by showing that the research broadly follows the development of the supporting technologies used. We identify four eras in the evolution of KGs for news, characterised by (1) applying early Semantic-Web ideas to the news domain, (2) exploiting the Linked Open Data (LOD) cloud for news purposes, (3) semantic knowledge graphs and machine learning and, most recently, (4) deep-learning approaches based on semantic knowledge graphs.

RQ4: Which are the most frequently cited papers and projects, and which papers and projects are citing one another? Our analyses in Sections 3.11 and 3.12 answer this question. The most cited papers are the ones about

DKN [M73] and KIM [M37]. Many recent papers that use deep-learning techniques for fake-news detection or recommendation are already much cited, e.g., [M52, 69]. Among the central projects, main papers related to the Hermes [M4], NewsReader [M72], and NEWS [M15] projects have been most cited. Another much referenced group of papers centres around what we have called the "MediaLoep" collaboration. The citation analysis reported in the online Addendum[2] (Figure 10) shows that the main paper from the Neptuno project [M7] and the effort to make IPTC's news architecture[A66] semantic [M71] also are important.

4.5 Limitations

The most central limitation of our literature review is its scope. We only consider papers that use semantic knowledge graphs or related semantic techniques for news-related purposes, excluding papers that attempt to solve similar problems using other knowledge representation techniques or targeting other domains. There is are also a growing body of research on representing texts in general as semantic knowledge graphs, proposing techniques and tools that could also be used to analyse news. There is another growing body of research on supporting news with knowledge graphs that are not semantically linked, i.e., with knowledge graphs whose nodes and edges do not link into the LOD cloud.

5 CONCLUSION

We have reported a systematic literature review of research on how semantic knowledge graphs can be used to facilitate all aspects of production, dissemination, and consumption of news. Starting with more than 6000 papers, we identified 80 main papers that we analysed in depth according to an analysis framework that we kept refining as analysis progressed. As a result, we have been able answer research questions about past, current, and emerging research areas and trends, and Section 4.3 has offered many paths for further work. We hope the results of our study will be useful for practitioners and researchers who are interested specifically in semantic knowledge graphs for news or more generally in computational journalism or in semantic knowledge graphs.

ACKNOWLEDGMENTS

This research is funded by the Norwegian Research Council's IKTPLUSS programme as part of the News Angler project (grant number 275872) and by MediaFutures partners and the Research Council of Norway as part of MediaFutures: Research Centre for Responsible Media Technology & Innovation (grant number 309339).

CONFLICT OF INTEREST

The authors are themselves involved in the News Angler project reported in [M46].

MAIN PAPERS

[M1] Adeel Ahmed and Syed Saif. 2017. DBpedia based ontological concepts driven information extraction from unstructured text. *International Journal of Advanced Computer Science and Applications* 8, 9 (2017). https://doi.org/10.14569/IJACSA.2017.080954

[M2] Alessio Antonini, Ruggero G. Pensa, Maria Luisa Sapino, Claudio Schifanella, Raffaele Teraoni Prioletti, and Luca Vignaroli. 2013. Tracking and analyzing TV content on the web through social and ontological knowledge. In *Proceedings of the 11th European Conference on Interactive TV and Video — EuroITV'13*. ACM Press, Como, Italy, 13. https://doi.org/10.1145/2465958.2465978

[M3] Francisco Berrizbeita and Maria-Esther Vidal. 2014. Traversing the Linking Open Data cloud to create news from tweets. In *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*, Robert Meersman, Hervé Panetto, Alok Mishra, Rafael Valencia-García, António Lucas Soares, Ioana Ciuciu, Fernando Ferri, Georg Weichhart, Thomas Moser, Michele Bezzi, and Henry Chan (Eds.). Vol. 8842. Springer, Berlin, Heidelberg, 479–488. https://doi.org/10.1007/978-3-662-45550-0_48

[M4] Jethro Borsje, Leonard Levering, and Flavius Frasincar. 2008. Hermes: a semantic web-based news decision support system. In *Proceedings of the 2008 ACM Symposium on Applied Computing — SAC'08*. ACM Press, Fortaleza, Ceara, Brazil, 2415. https://doi.org/10.1145/1363686.1364258

[M5] A.M. Braşoveanu and R. Andonie. 2019. Semantic fake news detection: a machine learning perspective. In *Proceedings of the International Work-Conference on Artificial Neural Networks 2019*. Springer, 656–667.

[M6] Iván Cantador, Pablo Castells, and Alejandro Bellogín. 2011. An enhanced semantic layer for hybrid recommender systems: Application to news recommendation. *International Journal on Semantic Web and Information Systems (IJSWIS)* 7, 1 (2011), 44–78.

[M7] Pablo Castells, Ferran Perdrix, E. Pulido, Rico Mariano, R. Benjamins, Jesús Contreras, and J. Lorés. 2004. Neptuno: Semantic web technologies for a digital newspaper archive. In *Proceedings of the European Semantic Web Symposium*. Springer, Berlin, Heidelberg, 445–458.

[M8] Mohammad Hossein Davarpour, Mohammad Karim Sohrabi, and Milad Naderi. 2019. Toward a semantic-based location tagging news feed system: Constructing a conceptual hierarchy on geographical hashtags. *Computers & Electrical Engineering* 78 (2019), 204–217. https://doi.org/10.1016/j.compeleceng.2019.07.005

[M9] Tom De Nies, Sam Coppens, Davy Van Deursen, Erik Mannens, and Rik Van de Walle. 2012. Automatic discovery of high-level provenance using semantic similarity. In *Provenance and Annotation of Data and Processes (Lecture Notes in Computer Science)*, Paul Groth and James Frew (Eds.). Springer, 97–110.

[M10] Pedro Debevere, Davy Van Deursen, Dieter Van Rijsselbergen, Erik Mannens, Mike Matton, Robbie De Sutter, and Rik Van de Walle. 2011. Enabling semantic search in a news production environment. In *Semantic Multimedia (Lecture Notes in Computer Science)*, Thierry Declerck, Michael Granitzer, Marcin Grzegorzek, Massimo Romanelli, Stefan Rüger, and Michael Sintek (Eds.). Springer, 32–47.

[M11] Mike Dowman, Valentin Tablan, Hamish Cunningham, and Borislav Popov. 2005. Web-assisted annotation, semantic indexing and search of television and radio news. In *Proceedings of the 14th International Conference on World Wide Web — WWW'05*. ACM Press, Chiba, Japan, 225. https://doi.org/10.1145/1060745.1060781

[M12] Ludivine Duroyon, François Goasdoué, Ioana Manolescu, François Goasdoué, and Ioana Manolescu. 2019. A linked data model for facts, statements and beliefs. In *Companion Proceedings of the 2019 World Wide Web Conference* (San Francisco USA, 2019-05-13). ACM, 988–993. https://doi.org/10.1145/3308560.3316737

[M13] Francesca Fallucchi, Rosario Di Stabile, Erasmo Purificato, Romeo Giuliano, and Ernesto William De Luca. 2021. Enriching videos with automatic place recognition in Google Maps. *Multimedia Tools and Applications* (2021), 1–17.

[M14] Dèlia Fernàndez-Cañellas, Joan Espadaler, David Rodriguez, Blai Garolera, Gemma Canet, Aleix Colom, Joan Marco Rimmek, Xavier Giro-i Nieto, Elisenda Bou, and Juan Carlos Riveiro. 2019. VLX-Stories: Building an online event knowledge base with emerging entity detection. In *Proceedings of the International Semantic Web Conference (ISWC'2019)* (2019). Springer, 382–399.

[M15] Norberto Fernández, José M. Blázquez, Jesús A. Fisteus, Luis Sánchez, Michael Sintek, Ansgar Bernardi, Manuel Fuentes, Angelo Marrara, and Zohar Ben-Asher. 2006. NEWS: Bringing semantic web technologies into news agencies. In *The Semantic Web — ISWC 2006*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora M. Aroyo (Eds.). Vol. 4273. Springer, Berlin, Heidelberg, 778–791. https://doi.org/10.1007/11926078_56

[M16] Norberto Fernández, José M. Blázquez, Luis Sánchez, and Ansgar Bernardi. 2007. IdentityRank: Named entity disambiguation in the context of the NEWS project. In *The Semantic Web: Research and Applications*, Enrico Franconi, Michael Kifer, and Wolfgang May (Eds.). Vol. 4519. Springer, Berlin, Heidelberg, 640–654. https://doi.org/10.1007/978-3-540-72667-8_45

[M17] Norberto Fernández, Damaris Fuentes, Luis Sánchez, and Jesús A Fisteus. 2010. The NEWS ontology: Design and applications. *Expert Systems with Applications* 37, 12 (2010), 8694–8704.

[M18] Michael Färber, Achim Rettinger, and Andreas Harth. 2016. Towards monitoring of novel statements in the news. In *The Semantic Web — Latest Advances and New Domains (Lecture Notes in Computer Science)*, Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange (Eds.). Springer, 285–299.

[M19] Jie Gao, Xin Xin, Junshuai Liu, Rui Wang, Jing Lu, Biao Li, Xin Fan, and Ping Guo. 2018. Fine-grained deep knowledge-aware network for news recommendation with self-attention. In *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, Santiago, 81–88. https://doi.org/10.1109/WI.2018.0-104

[M20] Roberto García, Ferran Perdrix, Rosa Gil, and Marta Oliva. 2008. The semantic web as a newspaper media convergence facilitator. *Journal of Web Semantics* 6, 2 (April 2008), 151–161. https://doi.org/10.1016/j.websem.2008.01.002

[M21] Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2013. Real-time RDF extraction from unstructured data streams. In *Proceedings of the International Semantic Web Conference (ISWC'2013)*, Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz (Eds.). Berlin, Heidelberg, 135–150. https://doi.org/10.1007/978-3-642-41335-3_9

[M22] J. Golbeck and C. Halaschek-Wiener. 2009. Trust-based revision for expressive web syndication. *Journal of Logic and Computation* 19, 5 (Oct. 2009), 771–790. https://doi.org/10.1093/logcom/exn045

[M23] A. Groza and A.-D. Pop. 2020. Fake news detector in the medical domain by reasoning with description logics. In *Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)* (2020-09). 145–152. https://doi.org/10.1109/ICCP51029.2020.9266270

[M24] Alexander Hogenboom, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto van der Meer. 2013. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications* 64, 1 (May 2013), 27–52. https://doi.org/10.1007/s11042-012-1122-0

[M25] Frederik Hogenboom, Damir Vandic, Flavius Frasincar, Arnout Verheij, and Allard Kleijn. 2014. A query language and ranking algorithm for news items in the Hermes news processing framework. *Science of Computer Programming* 94 (Nov. 2014), 32–52. https://doi.org/10.1016/j.scico.2013.07.018

[M26] Frank Hopfgartner and Joemon M. Jose. 2010. Semantic user profiling techniques for personalised multimedia recommendation. *Multimedia Systems* 16, 4-5 (Aug. 2010), 255–274. https://doi.org/10.1007/s00530-010-0189-6

[M27] Klesti Hoxha, Artur Baxhaku, and Ilia Ninka. 2016. Bootstrapping an online news knowledge base. In *Web Engineering*, Alessandro Bozzon, Philippe Cudre-Maroux, and Cesare Pautasso (Eds.). Vol. 9671. Springer, 501–506. https://doi.org/10.1007/978-3-319-38791-8_37

[M28] I-Ching Hsu. 2013. Personalized web feeds based on ontology technologies. *Information Systems Frontiers* 15, 3 (July 2013), 465–479. https://doi.org/10.1007/s10796-011-9337-6

[M29] Wouter IJntema, Jordy Sangers, Frederik Hogenboom, and Flavius Frasincar. 2012. A lexico-semantic pattern language for learning ontology instances from text. *Journal of Web Semantics* 15 (2012), 37–50. https://doi.org/10.1016/j.websem.2012.01.002

[M30] Akshay Java, Sergei Nirneburg, Marjorie McShane, Timothy Finin, Jesse English, and Anupam Joshi. 2007. Using a natural language understanding system to generate semantic web content. *International Journal on Semantic Web and Information Systems* 3, 4 (Oct. 2007), 50–74. https://doi.org/10.4018/jswis.2007100103

[M31] Yun Jing, Xu Zhiwei, and Gao Guanglai. 2020. Context-driven image caption with global semantic relations of the named entities. *IEEE Access* 8 (2020), 143584–143594.

[M32] Maarten Jongmans, Viorel Milea, and Flavius Frasincar. 2014. A semantic web approach for visualization-based news analytics. In *Knowledge Management in Organizations*, Lorna Uden, Darcy Fuenzaliza Oshee, I-Hsien Ting, and Dario Liberona (Eds.). Vol. 185. Springer, 195–204. https://doi.org/10.1007/978-3-319-08618-7_20

[M33] Kevin Joseph and Hui Jiang. 2019. Content based news recommendation via shortest entity distance over knowledge graphs. In *Companion Proceedings of the 2019 World Wide Web Conference* (San Francisco USA, 2019-05-13). ACM, 690–699. https://doi.org/10.1145/3308560.3317703

[M34] Leonidas Kallipolitis, Vassilis Karpis, and Isambo Karali. 2012. Semantic search in the world news domain using automatically extracted metadata files. *Knowledge-Based Systems* 27 (March 2012), 38–50. https://doi.org/10.1016/j.knosys.2011.12.007

[M35] Walter Kasper, Jörg Steffen, and Yajing Zhang. 2008. News annotations for navigation by semantic similarity. In *Proceedings of KI 2008: Advances in Artificial Intelligence*, Andreas R. Dengel, Karsten Berns, Thomas M. Breuel, Frank Bomarius, and Thomas R. Roth-Berghofer (Eds.). Vol. 5243. Springer, Berlin, Heidelberg, 233–240. https://doi.org/10.1007/978-3-540-85845-4_29

[M36] Haklae Kim. 2017. Building a K-Pop knowledge graph using an entertainment ontology. *Knowledge Management Research & Practice* 15, 2 (May 2017), 305–315. https://doi.org/10.1057/s41275-017-0056-8

[M37] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. 2004. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics* 2, 1 (Dec. 2004), 49–79. https://doi.org/10.1016/j.websem.2004.07.005

[M38] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. 2014. A fresh look on knowledge bases: Distilling named events from news. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management — CIKM'14*. ACM Press, Shanghai, China, 1689–1698. https://doi.org/10.1145/2661829.2661984

[M39] Edward H. Y. Lim, Raymond S. T. Lee, and James N. K. Liu. 2008. KnowledgeSeeker — an ontological agent-based system for retrieving and analyzing Chinese web articles. In *Proceedings of the 2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence)*. IEEE, Hong Kong, China, 1034–1041. https://doi.org/10.1109/FUZZY.2008.4630497

[M40] Danyang Liu, Jianxun Lian, Zheng Liu, Xiting Wang, Guangzhong Sun, and Xing Xie. 2021. Reinforced anchor knowledge graph generation for news recommendation reasoning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1055–1065.

[M41] Jue Liu, Zhuocheng Lu, and Wei Du. 2019. Combining enterprise knowledge graph and news sentiment analysis for stock price prediction. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019).

[M42] Jinshuo Liu, Chenyang Wang, Chenxi Li, Ningxi Li, Juan Deng, and Jeff Z Pan. 2021. DTN: Deep triple network for topic specific fake news detection. *Journal of Web Semantics* (2021), 100646.

[M43] Eduardo Lupiani-Ruiz, Ignacio García-Manotas, Rafael Valencia-García, Francisco García-Sánchez, Dagoberto Castellanos-Nieves, Jesualdo Tomás Fernández-Breis, and Juan Bosco Camón-Herrero. 2011. Financial news semantic search engine. *Expert Systems with Applications* 38, 12 (Nov. 2011), 15565–15572. https://doi.org/10.1016/j.eswa.2011.06.003

[M44] Erik Mannens, Sam Coppens, Toon De Pessemier, Hendrik Dacquin, Davy Van Deursen, Robbie De Sutter, and Rik Van de Walle. 2013. Automatic news recommendations via aggregated profiling. *Multimedia Tools and Applications* 63, 2 (March 2013), 407–425. https://doi.org/10.1007/s11042-011-0844-8

[M45] Qianren Mao, Xi Li, Hao Peng, Jianxin Li, Dongxiao He, Shu Guo, Min He, and Lihong Wang. 2021. Event prediction based on evolutionary event ontology knowledge. *Future Generation Computer Systems* 115 (2021), 76–89.

[M46] Enrico Motta, Enrico Daga, Andreas L Opdahl, and Bjørnar Tessem. 2020. Analysis and design of computational news angles. *IEEE Access* 8 (2020), 120613–120626.

[M47] Saikat Mukherjee, Guizhen Yang, and I. V. Ramakrishnan. 2003. Automatic annotation of content-rich HTML documents: Structural and semantic analysis. In *The Semantic Web — ISWC 2003*, Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Dieter Fensel, Katia Sycara, and John Mylopoulos (Eds.). Vol. 2870. Springer, Berlin, Heidelberg, 533–549. https://doi.org/10.1007/978-3-540-39718-2_34

[M48] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin Ireland, 2020-06-08). ACM, 16–25. https://doi.org/10.1145/3372278.3390670

[M49] Quang-Minh Nguyen and Tuan-Dung Cao. 2015. A novel approach for automatic extraction of semantic data about football transfer in sport news. *International Journal of Pervasive Computing and Communications* 11, 2 (2015), 233–252. https://doi.org/10.1108/IJPCC-03-2015-0018 WOS:000212340300007.

[M50] Quang-Minh Nguyen, Thanh-Tam Nguyen, and Tuan-Dung Cao. 2016. Semantic-based recommendation for sport news aggregation system. In *Research and Practical Issues of Enterprise Information Systems (Lecture Notes in Business Information Processing)*, A Min Tjoa, Li Da Xu, Maria Raffai, and Niina Maarit Novak (Eds.). Springer, 32–47.

[M51] Inna Novalija and Dunja Mladenić. 2013. Applying semantic technology to business news analysis. *Applied Artificial Intelligence* 27, 6 (July 2013), 520–550. https://doi.org/10.1080/08839514.2013.805600

[M52] Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web — ISWC 2018*, Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl (Eds.). Vol. 11136. Springer, 669–683. https://doi.org/10.1007/978-3-030-00671-6_39

[M53] Koralia Papadokostaki, Stavros Charitakis, George Vavoulas, Stella Panou, Paraskevi Piperaki, Aris Papakonstantinou, Savvas Lemonakis, Anna Maridaki, Konstantinos Iatrou, Piotr Arent, Dawid Wiśniewski, Nikos Papadakis, and Haridimos Kondylakis. 2017. News Articles Platform: Semantic tools and services for aggregating and exploring news articles. In *Strategic Innovative Marketing (Springer Proceedings in Business and Economics)*, Androniki Kavoura, Damianos P. Sakas, and Petros Tomaras (Eds.). Springer, 511–519.

[M54] Marco Ponza, Diego Ceccarelli, Paolo Ferragina, Edgar Meij, and Sambhav Kothari. 2021. Contextualizing trending entities in news stories. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 346–354.

[M55] Radityo Eko Prasojo, Mouna Kacimi, and Werner Nutt. 2018. Modeling and summarizing news events using semantic triples. In *The Semantic Web (Lecture Notes in Computer Science)*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.). Springer, 512–527.

[M56] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web — WWW'12*. ACM Press, Lyon, France, 909–918. https://doi.org/10.1145/2187836.2187958

[M57] D. B. Ramagem, B. Margerin, and J. Kendall. 2004. AnnoTerra: Building an integrated earth science resource using semantic web technologies. *IEEE Intelligent Systems* 19, 3 (May 2004), 48–57. https://doi.org/10.1109/MIS.2004.3

[M58] Wouter Rijvordt, Frederik Hogenboom, and Flavius Frasincar. 2019. Ontology-driven news classification with Aethalides. *Journal of Web Engineering* 18, 7 (2019), 627–654. https://doi.org/10.13052/jwe1540-9589.1873

[M59] Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics* 37-38 (March 2016), 132–151. https://doi.org/10.1016/j.websem.2015.12.004

[M60] Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. 2019. Searching news articles using an event knowledge graph leveraged by Wikidata. In *Companion Proceedings of the 2019 World Wide Web Conference*. 1232–1239.

[M61] Tomer Sagi, Yael Wolf, and Katja Hose. 2019. How new is the (RDF) news?. In *Companion Proceedings of the 2019 World Wide Web Conference* (2019). 714–721.

[M62] Daniel Sarmiento Suárez and Claudia Jiménez-Guarín. 2014. Natural language processing for linking online news and open government data. In *Advances in Conceptual Modeling*, Marta Indulska and Sandeep Purao (Eds.). Vol. 8823. Springer, 243–252. https://doi.org/10.1007/978-3-319-12256-4_26

[M63] A. Scharl, D. Herring, W. Rafelsberger, A. Hubmann-Haidvogel, R. Kamolov, D. Fischl, M. Föls, and A. Weichselbraun. 2017. Semantic systems and visual tools to support environmental communication. *IEEE Systems Journal* 11, 2 (June 2017), 762–771. https://doi.org/10.1109/JSYST.2015.2466439

[M64] Kim Schouten, Philip Ruijgrok, Jethro Borsje, Flavius Frasincar, Leonard Levering, and Frederik Hogenboom. 2010. A semantic web-based approach for personalizing news. In *Proceedings of the 2010 ACM Symposium on Applied Computing — SAC'10*. ACM Press, Sierre, Switzerland, 854. https://doi.org/10.1145/1774088.1774264

[M65] Md. Hanif Seddiqui, Md. Nesarul Hoque, and Md. Hasan Hafizur Rahman. 2015. Semantic annotation of Bangla news stream to record history. In *Proceedings of the 2015 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, Dhaka, Bangladesh, 566–572. https://doi.org/10.1109/ICCITechn.2015.7488135

[M66] Heng-Shiou Sheu, Zhixuan Chu, Daiqing Qi, and Sheng Li. 2021. Knowledge-guided article embedding refinement for session-based news recommendation. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[M67] K. Srinivasa and P. Santhi Thilagam. 2019. Crime Base: Towards building a knowledge base for crime entities and their relationships from online news papers. *Information Processing & Management* 56, 6 (2019), 102059. https://doi.org/10.1016/j.ipm.2019.102059

[M68] Andrei Tamilin, Bernardo Magnini, Luciano Serafini, Christian Girardi, Mathew Joseph, and Roberto Zanoli. 2010. Context-driven semantic enrichment of Italian news archive. In *The Semantic Web: Research and Applications*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache (Eds.). Vol. 6088. Springer, Berlin, Heidelberg, 364–378. https://doi.org/10.1007/978-3-642-13486-9_25

[M69] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Stefan Dietze, Benjamin Zapilko, and Konstantin Todorov. 2019. ClaimsKG: A live knowledge graph of fact-checked claims. (2019).

[M70] Yu Tian, Yuhao Yang, Xudong Ren, Pengfei Wang, Fangzhao Wu, Qian Wang, and Chenliang Li. 2021. Joint knowledge pruning and recurrent graph convolution for news recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 51–60.

[M71] Raphaël Troncy. 2008. Bringing the IPTC news architecture into the semantic web. In *The Semantic Web — ISWC 2008*. Springer, 483–498.

[M72] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, Marco Rospocher, and Roxane Segers. 2016. NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 110 (Oct. 2016), 60–85. https://doi.org/10.1016/j.knosys.2016.07.013

[M73] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 1835–1844. http://arxiv.org/abs/1801.08284 arXiv: 1801.08284.

[M74] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 540–547.

[M75] Yueji Yang, Yuchen Li, and Anthony KH Tung. 2021. NewsLink: Empowering intuitive news search with knowledge graphs. In *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 876–887.

[M76] Ryohei Yokoo, Takahiro Kawamura, and Akihiko Ohsuga. 2016. Semantics-based news delivering service. *International Journal of Semantic Computing* 10, 04 (Dec. 2016), 445–459. https://doi.org/10.1142/S1793351X1640016X

[M77] SoYeop Yoo and OkRan Jeong. 2020. Automating the expansion of a knowledge graph. *Expert Systems with Applications* 141 (March 2020), 112965. https://doi.org/10.1016/j.eswa.2019.112965

[M78] Jingyuan Zhang, Chun-Ta Lu, Bokai Cao, Yi Chang, and Philip S. Yu. 2017. Connecting emerging relationships from news via tensor factorization. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, MA, 1223–1232. https://doi.org/10.1109/BigData.2017.8258048

[M79] Jingyuan Zhang, Chun-Ta Lu, Mianwei Zhou, Sihong Xie, Yi Chang, and Philip S. Yu. 2016. HEER: Heterogeneous graph embedding for emerging relation detection from news. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington DC,USA, 803–812. https://doi.org/10.1109/BigData.2016.7840673

[M80] Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3132–3143.

REFERENCES

[R1] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. 2007. Open user profiles for adaptive news systems: help or harm?. In *Proceedings of the 16th International Conference on World Wide Web*. 11–20.

[R2] Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access* 8 (2020), 32862–32881.

[R3] Dean Allemang, James Hendler, and Fabien Gandon. 2020. *Semantic Web for the Working Ontologist*. Elsevier.

[R4] Luigi Atzori, Antonio Iera, and Giacomo Morabito. 2010. The Internet of Things: A survey. *Computer Networks* 54, 15 (2010), 2787–2805.

[R5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*. Springer, 722–735.

[R6] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American* 284, 5 (2001), 34–43.

[R7] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked Data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. IGI Global, 205–227. Also published in International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue on Linked Data.

[R8] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. 1247–1250.

[R9] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems* 26 (2013), 2787–2795.

[R10] Virginia Braun and Victoria Clarke. 2014. What can "thematic analysis" offer health and wellbeing researchers? *International Journal of Qualitative Studies on Health and Well-being* 9, 1 (2014), 1–2. https://doi.org/10.3402/qhw.v9.26152

[R11] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637.

[R12] David Caswell and Chris W. Anderson. 2019. Computational journalism. In *The International Encyclopedia of Journalism Studies*. Wiley Online Library, 1–8.

[R13] Vinay Chaudhri, Chaitanya Baru, Naren Chittar, Xin Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas Lenat, Juan Sequeda, Denny Vrandečić, et al. 2022. Knowledge graphs: Introduction, history and, perspectives. *AI Magazine* 43, 1 (2022), 17–29.

[R14] Hamish Cunningham. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. 168–175.

[R15] Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities* 36, 2 (2002), 223–254.

[R16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[R17] Nicholas Diakopoulos. 2017. Computational journalism and the emergence of news platforms. In *The Routledge Companion to Digital Journalism Studies*. Routledge London, UK, 176–184.

[R18] Christiane Fellbaum. 2000. WordNet: An electronic lexical database. *Language* 76 (2000), 706.

[R19] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. 2020. Introduction: what is a knowledge graph? In *Knowledge Graphs*. Springer, 1–10.

[R20] Marc Gallofré Ocaña and Andreas Lothe Opdahl. 2021. Developing a software reference architecture for journalistic knowledge platforms. In *Companion Proceedings of the 15th European Conference on Software Architecture, ECSA 2021 Companion Volume*, Vol. 2978. Technical University of Aachen/CEUR Workshop Proceedings.

[R21] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.

[R22] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. MIT press.

[R23] Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* 151 (2018), 78–94.

[R24] Claudio Gutiérrez and Juan F Sequeda. 2021. Knowledge graphs. *Commun. ACM* 64, 3 (2021), 96–104.

[R25] Bahareh Rahmanzadeh Heravi and Jarred McGinnis. 2015. Introducing social semantic journalism. *The Journal of Media Innovations* 2, 1 (2015), 131–140.

[R26] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge* 12, 2 (2021), 1–257.

[R27] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*, Vol. 779. Citeseer, 48–57.

[R28] Alan Jackoway, Hanan Samet, and Jagan Sankaranarayanan. 2011. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 25–32.

[R29] Akshay Java, Tim Finin, Sergei Nirenburg, et al. 2006. SemNews: a semantic news framework. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. 1939–1940.

[R30] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long papers)*. 687–696.

[R31] Barbara Kitchenham. 2004. *Procedures for performing systematic reviews*. Technical Report 33. Keele, UK, Keele University. 1–26 pages.

[R32] Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38, 11 (1995), 33–38.

[R33] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 2181–2187.

[R34] Marcel Machill and Markus Beiler. 2009. The importance of the Internet for journalistic research: A multi-method study of the research performed by journalists working for daily newspapers, radio, television and online. *Journalism Studies* 10, 2 (2009), 178–203.

[R35] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the news: Digital creativity support for journalists. In *Proceedings of the 2018 CHI Conference on Human*

Factors in Computing Systems. ACM, 475.

[R36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (2013), 3111–3119.

[R37] George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[R38] Andrey Miroshnichenko. 2018. AI to bypass creativity. Will robots replace journalists? (The answer is "yes"). *Information* 9, 7 (2018), 183.

[R39] Andreas C Müller and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc.

[R40] Benedito Medeiros Neto, Edison Ishikawa, George Ghinea, and Tor-Morten Grønli. 2019. Newsroom 3.0: Managing technological and media convergence in contemporary newsrooms. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019).

[R41] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8 (2019), 36–43.

[R42] Lars Nyre, Solveig Bjørnestad, Bjørnar Tessem, and Kjetil Vaage Øie. 2012. Locative journalism: Designing a location-dependent news medium for smartphones. *Convergence* 18, 3 (2012), 297–314.

[R43] Andreas L Opdahl and Bjørnar Tessem. 2020. Ontologies for finding journalistic angles. *Software and Systems Modeling* (2020), 1–17.

[R44] Kosmas Panagiotidis and Andreas Veglis. 2020. Transitions in journalism — Toward a semantic-oriented technological framework. *Journal. Media* 1 (2020), 1.

[R45] Tassilo Pellegrini. 2012. Semantic metadata in the news production process: achievements and challenges. (2012), 125–133. https://doi.org/10.1145/2393132.2393158

[R46] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[R47] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. 2003. KIM — semantic annotation platform. In *International Semantic Web Conference*. Springer, 834–849.

[R48] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. 2004. KIM — a semantic platform for information extraction and retrieval. *Natural Language Engineering* 10, 3-4 (2004), 375–392.

[R49] Ramón Salaverría. 2019. Digital journalism. In *The International Encyclopedia of Journalism Studies*. Wiley Online Library, 1–11.

[R50] Luis Sánchez-Fernández, Norberto Fernández-García, Ansgar Bernardi, Lars Zapf, Anselmo Penas, and Manuel Fuentes. 2005. An experience with Semantic Web technologies in the news domain. In *Workshop on Semantic Web Case Studies and Best Practices for eBusiness*.

[R51] Amit Sheth and Krishnaprasad Thirunarayan. 2012. Semantics empowered Web 3.0: Managing enterprise, social, sensor, and cloud-based data and services for advanced applications. *Synthesis Lectures on Data Management* 4, 6 (2012), 1–175.

[R52] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*. 697–706.

[R53] Bjørnar Tessem, Lars Nyre, Michel D. S. Mesquita, and Paul Mulholland. 2022. Deep learning to encourage citizen involvement in local journalism. In *Futures of Journalism*, V.J.E. Manninen, M.K. Niemi, and A. Ridge-Newman (Eds.). Palgrave Macmillan, 211–226. https://doi.org/10.1007/978-3-030-95073-6_14

[R54] Neil Thurman. 2019. Computational journalism. In *The Handbook of Journalism Studies* (second ed.). Routledge, New York, 475.

[R55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008.

[R56] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

Addendum to “Semantic Knowledge Graphs for the News: A Review”

Andreas L Opdahl, Tareq Al-Moslmi, Duc-Tien Dang-Nguyen,
Marc Gallofré Ocaña, Bjørnar Tessem & Csaba Veres
{Andreas.Opdahl,Duc-Tien.Dang-Nguyen,Marc.Gallofre,Bjornar.Tessem,Csaba.Veres}@uib.no,
tareqmail19@gmail.com

Intelligent Information Systems (I2S) group
Dept. of Information Science and Media Studies
University of Bergen
P.O.Box 7802, N-5020 Bergen, Norway

This addendum provides more detail about the systematic literature review method we have followed in our review paper “Semantic Knowledge Graphs for the News: A Review”[A95] (Section A). It also provides additional background for the discussion and analysis of the main papers (Section B). The final reference list suggests further readings about the tools and other resources we have mentioned in our review paper.

A RESEARCH METHOD

A.1 Research approach

We adopted the guidelines for *systematic literature reviews (SLR)* in software engineering proposed in [31]. In line with the purpose of and as explained in the Method section (Section 2) of our review paper[A95], we conducted a *descriptive review*, which presents an overview of past and ongoing research, in order to better understand it as an emerging research field and to identify central challenges and opportunities. We therefore focused on reviewing the research literature in *breadth* in order to cover as many salient research problems, approaches, and potential solutions as possible.

We first performed an explorative pre-study in order to find appropriate search terms, to ensure that the research literature was sufficiently comprehensive for our study to be worthwhile, and to establish an initial analysis framework. We then established a review protocol that outlined the: motivation and research questions (the Introduction section (Section 1) in our review paper), research approach (this section), scope (Section A.2), literature sources (Section A.3), paper-selection criteria (Section A.4), and framework for analysing the selected papers (Section A.5).

A.2 Scope

Our scope is research on *semantic knowledge graphs for the news*, where we understand both key phrases — *semantic knowledge graphs* and *the news* — in a wide sense. Along with *semantic knowledge graphs*, we include facilitating semantic technologies like RDF[A121], OWL,[A117] and SPARQL[A124] and their uses for semantically linked (open) data [7] and semantic web [6]. However, we exclude uses of knowledge graphs that do not exploit external semantic linking, but use graphs only as a local data structure, for example for sub-symbolic analyses using graph embedding [11, 23]. We consider the *news* to include all phases of the news life-cycle, from detection through production, distribution, consumption, aggregating, and searching to archiving. However, we exclude research that uses news only for evaluation purposes. Table 7 lists our detailed inclusion and exclusion criteria.

Table 7. Inclusion and exclusion criteria for selecting papers.

Include	Exclude
• Papers that report research on semantic knowledge graphs for the news • Full papers published in English since 2000 (but we found no candidates from before 2000 anyway) • Peer-reviewed and archived research papers published in recognised journals or high-quality conferences and workshops • Final reports of completed research • Primary studies	• Papers that use knowledge graphs only as a local data structure and do not exploit external semantic linking • Papers that use news corpora for evaluation only • Short papers (< 6 pages) • Papers in other languages than English • Papers in lower-quality journals or in lower-quality conferences and workshops (according to http://www.conferenceranks.com/ and http://portal.core.edu.au/conf-ranks/) • Papers superseded by more comprehensive reports of the same work by the same research group • Survey and review papers, prefaces, presentations, and workshop introductions

A.3 Literature search

This section details our literature search process.

Sources: We searched for literature using the following much used search engines:

- ACM Digital Library[A14],
- Elsevier ScienceDirect[A32],
- IEEE Xplore[A63],
- SpringerLink[A83], and
- Clarivate Analytics' Web of Science[A18].

Before completing the study, we also performed supplementary searches through Google Scholar[A52].

Search strings: We searched for papers in the intersection of semantic knowledge graphs and the news using the following search string as our basis:

("semantic web" OR "linked data" OR "linked open data"
OR "semantic technolog*" OR "knowledge graph*")
AND ("news" OR "journalis*")

Initial searches: We adapted the search string to match the syntax and other limitations of each search engine, frequently having to combine the results of several simpler searches. In those search engines that supported it, we searched for matching titles, abstracts, and keywords but, in SpringerLink, we instead had to search the full paper texts. Table 8 shows the number of unique papers returned by each search engine, for each keyword combination and in total (the numbers do not always add up because many papers matched more than one keyword combination and were returned by several search engines). "Semantic web" and "linked data" combined with "news" were the two most common keyword combinations in these initial searches. Our searches returned 6393 unique titles of candidate papers, 5788 of them were returned by the full text searches in SpringerLink.

Supplementary searches: As the detailed paper analyses progressed, we performed supplementary searches through the five search engines and Google Scholar to pick up papers published after our first round of search along with other missing papers. The searches identified 862 additional candidate papers, which were screened in the same way as the candidates from the initial search. A few papers were also identified through snowballing, i.e., by references from other main papers.

Table 8. Hits per search engine for each of our search-string combinations (n — news, j — journalism, kg — knowledge graphs, ld — linked data, lod — linked open data, st — semantic technologies, and sw — semantic web).

	Keyword combination										
Search engine	j-kg	j-ld	j-lod	j-st	j-sw	n-kg	n-ld	n-lod	n-st	n-sw	**Unique**
ACM DL	4	8	4	1	0	82	152	64	37	377	648
IEEE Xplore	1	1	0	2	11	8	8	3	18	78	123
ScienceDirect	0	0	0	0	2	2	0	0	1	16	17
SpringerLink	76	304	121	82	578	438	1526	621	614	4149	5788
Web of Science	1	2	1	1	9	5	14	8	6	54	91
All engines	82	315	126	86	619	537	1711	696	681	4721	6393

A.4 Paper selection

This section details the paper selection process we have followed.

Screening: To select the main papers to review and analyse, we screened the search results in three stages in light of our inclusion and exclusion criteria. At each stage, at least two of the authors considered each candidate paper. While taking into account our different specialities, we also attempted to use different authors in each stage.

Screening for scope: Because our searches had returned so many results, we screened the results of each search engine separately, without attempting to systematically remove duplicates. For each paper returned from ACM Digital Library, IEEE Xplore, ScienceDirect, and Web of Science we screened the title, abstract and, when available, keywords according to our inclusion and exclusion criteria. For the more than 5000 papers returned from SpringerLink we also screened the title and abstract in order of relevance, but only screened the results for each keyword combination until we encountered 30 consecutive excluded results, after which we assumed that the remaining ones would be even less-relevant and safe to exclude. In the end, 2214 candidate papers were thus screened by a pair of raters in this round. The raters agreed to accept 271 and reject 1750 outright, and they disagreed on 193 papers, giving an inter-rater agreement (Cohen’s kappa) of 0.69. After discussion and removal of 33 duplicates, 339 papers were passed on to the next round.

Screening by inclusion and exclusion criteria: In the second stage, we downloaded the 339 full papers and skimmed them according to our inclusion and exclusion criteria. Hence, in addition to research theme, we also considered the length, type, language, and source of each paper in this stage. The two raters agreed to accept 77 and reject 191 of them, and they disagreed on 71 papers, giving a kappa of 0.53. After discussion and removal of a few papers that reported overlapping work, 133 papers were left.

Screening by content: In the third and final stage, we read all the 133 remaining papers cursorily and marked the included ones according to the analysis framework presented in the Method section (Section 2) of our review paper[A95]. Of the 133 papers from the second screening, the two raters agreed to accept 71 and reject 31, and they disagreed on 31 papers, giving a kappa of 0.49. After discussion, 9 more papers were included as main papers. In the end, 80 papers were thus passed on to detailed analysis.

Table 9 summarises the paper selection process. Our Cohen’s kappa values range from 0.69 to 0.49, which is considered “good” to “fair”. The falling inter-rater agreement in each round may reflect the increasing ratio of borderline papers and the difficulty of assessing them in increasing detail.

A.5 Paper Analysis

Analysis framework: To answer the research questions more clearly, we used an analysis framework that comprised a hierarchy of *themes*, i.e., patterns in the main papers that capture something that is important in relation to our

Table 9. Key numbers from the paper selection process.

Round	Papers in	Clear yes	Clear no	Discussion	Remaining	Kappa
1st round	2214	271	1750	193	372[†]	0.69
2nd round	339	77	191	71	133	0.53
3rd round	133	71	31	31	80	0.49

[†]Of which 33 duplicates were removed before the next round.

research questions [10]. An example of a pattern is that many main papers address specific groups of intended users, a *top-level theme* of which the specific patterns of users, such as journalists, archivists, and the general public, become *sub-themes*. We used the pilot study to extract candidate top-level themes, which we continued to revise and refine according to the thematic analysis method [10] as we analysed the main papers in increasing detail. The analysis framework presented in the Method section (Section 2) of our review paper lists the final top-level themes, along with examples of sub-themes that we used to describe and compare the papers in the Review section (Section 3) in our review paper.

Manual marking: We used a template to mark each main paper according to the analysis framework, and iterated the marking process until the top-level themes and sub-themes had stabilised. We merged some similar sub-themes and in the end excluded sub-themes that did not match multiple papers. The conceptual framework presented in the Discussion section (Section 4) of our review paper presents a populated version of the framework resulting from analysing the 80 main papers. For each top-level theme, it lists the most frequently used sub-themes along with the number of times each sub-theme has matched a main paper.

Table 10. SPARQL queries for inspecting the semantic knowledge graph at http://bg.newsangler.uib.no.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ss: <http://semanticscholar.org/>
PREFIX kg: <http://i2s.uib.no/kg4news/>
PREFIX sp: <http://i2s.uib.no/kg4news/science-parse/>
PREFIX th: <http://i2s.uib.no/kg4news/theme/>

SELECT DISTINCT ?t WHERE { ?s rdf:type ?t . }  # All types used
SELECT DISTINCT ?p WHERE { ?s ?p ?o . }        # All properties used
SELECT ?tp ?e WHERE { ?tp rdfs:comment ?e . }  # All explanations of types and properties

SELECT ?p ?t WHERE { ?p rdf:type kg:MainPaper . ?p dc:title ?t . }    # All main papers
SELECT ?a ?n WHERE { ?a rdf:type kg:MainAuthor . ?a foaf:name ?n . }  # All main authors

SELECT ?l WHERE { ?l ^skos:prefLabel / rdf:type ss:Topic . }                   # All topics from Semantic Scholar
SELECT ?t WHERE { "Graph embedding" ^skos:prefLabel / ^dc:subject / dc:title ?t . } # All papers about topic "Graph embedding"

SELECT ?t ?l WHERE { ?t rdf:type th:TopLevelTheme . ?t skos:prefLabel ?l . }                # All top-level themes
SELECT ?l WHERE { "Domain" ^skos:prefLabel / ^skos:broader / skos:prefLabel ?l . }        # All subthemes of "Domain"
SELECT ?t WHERE { "stock market" ^skos:prefLabel / ^th:theme / ^dc:subject / dc:title ?t . } # All main papers about stocks
```

Metadata harvesting: We searched Semantic Scholar^{A36} to obtain a URI for each main paper. We downloaded paper and author data from the Semantic Scholar API and scraped their web pages using Selenium and BeautifulSoup to retrieve topic descriptions. In addition to metadata for each main paper, such as title, abstract, and authors, we also collect incoming and outgoing references along with the titles and authors of these referencing and referenced papers.

Knowledge graph: From the extracted data and the manual paper markings, we built a knowledge graph. The complete graph contains information about 4238 papers, 9712 authors, and 699 main-paper topics from Semantic Scholar. It also contains the top-level themes, the 226 sub-themes, and the 1711 sub-theme matches with the main papers. It is available at http://bg.newsangler.uib.no. Table 10 presents examples of SPARQL queries that can be used to explore the graph.

Automatic text extraction We proceeded to extract the raw text from each main paper using AllenAI's science-parse tool[10]. We brushed up the raw texts manually to standardise section titles and extracted the core text of each paper, excluding sections such as "Literature", "Related work", "Discussion", and "Further work", which were most likely to mention research efforts outside each paper's main contribution. We automatically searched the core texts for occurrences of themes and sub-themes, taking into account likely spelling variants and synonyms, in order to quality control our manual paper markings.

While our review and discussion are primarily based on *careful manual reading, analysis, marking, and discussion* of papers, we have used the automatic data extraction and analysis as a supplement to inform and corroborate our findings.

B SUPPLEMENTARY ANALYSIS MATERIALS

This section provides additional details about our main paper analyses.

B.1 Other techniques and tools

Section 3.7 of our review paper analyses the semantic techniques and tools used in our main papers. Most of the main papers use them in combination with other, *non-semantic* techniques and tools. Similar to Section 3.7, we separate them into *exchange formats*, *information resources*, and *processing* and *storage techniques*.

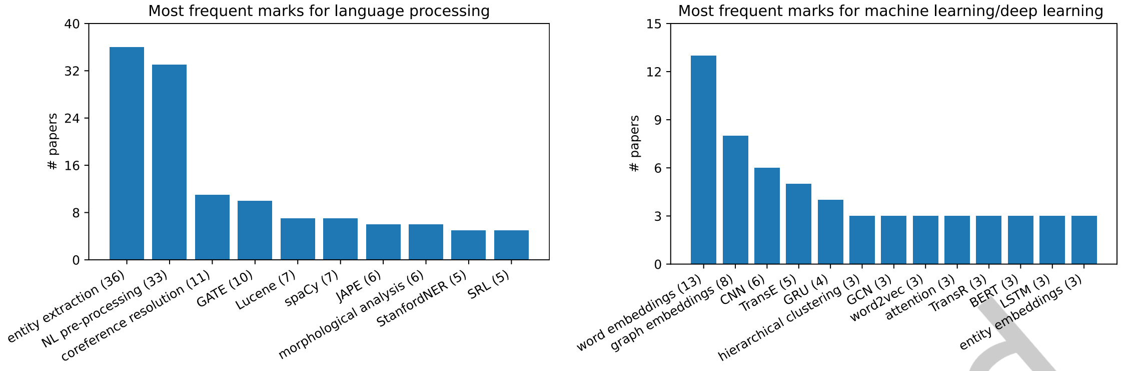
Information exchange formats: On the news side, IPTC standards are popular. *NewsML-G2*^{A72} is used in six papers, the *News Industry Text Format (NITF)*^{A71} in two, and the *News Architecture Framework (NAF)*^{A66} also in two. In addition, 17 papers somehow use *RSS feeds*, either as inputs or for evaluation. Used in four papers, *MPEG-7* is the only frequent format for multimedia exchange. For example, it is one of the input formats to the semantic mappings in [M20]. For information exchange in general, *XML*, *HTML*, *JSON*, and *CSS* are most frequent.

Information resources: On the news side, IPTC standards are again central, and seven papers use the *IPTC's Media Topics*^{A67} (or their precursor, the Subject Codes^{A70}) to categorise and otherwise label content. Whereas most of the papers in this review rely on older versions of the Media Topics, the ASRAEL project [M60] uses the IPTC's more recent SKOS-based version. The more general *IPTC News Codes*^{A68} are also used in two papers.

On the natural-language side, *WordNet* [18, 37] is the most used resource by far (22 papers). *VerbNet*^{A86} is second most used (with four papers). Other NLP resources are *FrameNet*^{A85}, *ProbBank*^{A10} and the *Predicate Matrix*^{A15}, all of them used in NewsReader's machine reading pipeline [M72]. The *Penn Treebank*^{A108} is also used in two papers [M6,M21]. In addition come general non-semantic information resources such as *Facebook*, *Google*, *Twitter*, and *Wikipedia*. *Yahoo! Finance* is used in two papers.

Processing techniques: On the natural-language side, almost all the annotation and information-extraction approaches use standard natural-language processing (NLP) techniques for semantic lifting. Figure 9a shows the most frequent ones. Most central are entity extraction and NL pre-processing in general. Other much

[10] https://github.com/allenai/science-parse, also used by Semantic Scholar^{A36} for metadata extraction.



(a) (b)

Fig. 9. The most frequently used information processing techniques for (a) NLP and (b) machine learning/deep learning.

used techniques are co-reference resolution, morphological analysis, sentiment analysis, relation extraction, dependency parsing, semantic-role labelling (SRL), and word-sense disambiguation (WSD). The most used framework for text processing is the General Architecture for Text Engineering (GATE)[A92], along with its components A Nearly New Information Extraction system (ANNIE)[A91] for information extraction and the Java Annotations Pattern Engine (JAPE)[A93] for patterns. Other frequent tools are spaCy[A35] and its NER component, Lucene[A45] for plain text indexing and searching, and StanfordNER and -NLP[A57].

Unsurprisingly, the last decade has seen more and more proposals that exploit machine-learning techniques, as illustrated by three early examples from 2012: [M9] uses greedy clustering to automatically detect provenance relations between news articles. The Hermes framework [M29] uses a pattern-language and rule-based approach to learn ontology instances and event relations from text, combining lexico-semantic patterns with semantic information. It is used to analyse financial and political news articles, splitting its corpus of news articles into a training and a test set. Pundit [M56] mines text patters from news headlines to predict potential future events based on textual descriptions of current events. It uses machine learning to automatically induce a causality function based on examples of causality pairs mined from a large collection of archival news headlines. Whereas these early approaches rely on hand-crafted rules and dedicated learning algorithms, more recent proposals use standard machine-learning techniques such as word [36, 46] and graph embeddings [11, 23], As shown in Figure 9b, TransE [9] and TransR [33] are most used for graph embedding, along with word2vec [36] for word embedding.

In the years since 2019, there has been a sharp rise in deep learning [22] approaches that employ transformers [16] and attention mechanisms [55]. PolarisX [M77] uses pre-trained multilingual BERT model to detect new relations, with the aim of updating its underlying knowledge graph in real time. TAMURE [M78] uses tensor factorisation implemented in TensorFlow[A13] to learn joint embedding representations of entities and relation types. Focusing on click-through rate (CTR) prediction in online news sites, DKN [M73] uses a Convolutional Neural Network [22] with separate channels for words and entities and an attention module to dynamically aggregate user histories. [M19] proposes a deep neural network model that employs multiple self-attention modules for words, entities, and users for news recommendation. [M52] proposes the B-TransE model to detect fake news based on content. Several other papers that employ deep neural networks have already been mentioned [M80,M70,M66,M40,M31]. The most used deep learning techniques and tools are Convolutional Neural Networks (CNN) [22], Gated-Recurrent Units (GRU) [22], Bidirectional Encoder Representations from

Table 11. The 15 most frequently cited main papers (recency weighted).

Title	**Year**	**Ref**	**# citations**	**Citation weight**	**# main paper citations**
DKN: Deep knowledge-aware network for news recommendation	2018	[M73]	413	33.15	4
Semantic annotation, indexing, and retrieval	2004	[M37]	523	16.34	1
Learning causality for news events prediction	2012	[M56]	199	9.64	2
Building event-centric knowledge graphs from news	2016	[M59]	111	7.35	2
Content based fake news detection using knowledge graphs	2018	[M52]	72	5.78	1
ClaimsKG: A knowledge graph of fact-checked claims	2019	[M69]	49	4.38	0
NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news	2016	[M72]	60	3.97	1
A fresh look on knowledge bases: Distilling named events from news	2014	[M38]	57	3.19	2
Real-time RDF extraction from unstructured data streams	2013	[M21]	60	3.12	1
Semantics-based information extraction for detecting economic events	2013	[M24]	55	2.86	0
A lexico-semantic pattern language for learning ontology instances from text	2012	[M29]	53	2.57	3
Web-assisted annotation, semantic indexing and search of television and radio news	2005	[M11]	78	2.55	1
An enhanced semantic layer for hybrid recommender systems: Application to news recommendation	2011	[M6]	53	2.40	0
Financial news semantic search engine	2011	[M43]	50	2.27	0
Automatic annotation of content-rich HTML documents: Structural and semantic analysis	2003	[M47]	75	2.25	0

Transformers (BERT) [16], Graph Convolution Networks (GCN) [11, 23], Long Short-Term Memories (LSTM) [22], and attention mechanisms [55]. Keras[A5] and TensorFlow[A13] are used for programming.

Storage techniques: The NEWS project [M15,M16] uses a Heuristic and Deductive Database (HDDB) to store and index the textual content of news items in order to support keyword-, category- and entity-based searches. MongoDB[A7] is used in two other projects [M27,M59], and four papers mention MySQL[A96] or relational DBs in general. Beyond KnowledgeStore[A28], mentioned in the previous section, there is so far little use of big-data technologies for large knowledge graphs.

Summary: Our review shows that the research on semantic knowledge graphs for the news is technologically diverse. We find examples of research that exploits most of the popular news-related standards and most of the popular techniques for NLP, machine learning, deep learning, and computing in general. On the news side, the IPTC family of standards and resources is central. On the NLP side, entity extraction, NL pre-processing, co-reference resolution, morphological analysis, and semantic-role labelling are common, whereas GATE, Lucene, spaCy, JAPE and StanfordNER are the most used tools. On the ML side, techniques for word, graph and entity embeddings are popular, such as TransE, TransR, TransD, and word2vec. On the DL side, neural-network techniques such as CNN, GRU, GCN, LSTM, BERT, and attention are much used. The focus on news standards is strongest in the first part of the study period, up to around 2014, when many approaches incorporate existing

Table 12. The 15 papers most frequently referenced by our main papers.

Title	Year	Ref	# main paper refs
The semantic web	2001	[6]	15
WordNet: An electronic lexical database	2000	[18]	11
Translating embeddings for modeling multi-relational data	2013	[9]	8
GATE, a general architecture for text engineering	1997	[15]	7
Distributed representations of words and phrases and their compositionality	2013	[36]	7
Learning entity and relation embeddings for knowledge graph completion	2015	[33]	7
YAGO: A core of semantic knowledge	2007	[52]	6
Freebase: A collaboratively created graph database for structuring human knowledge	2008	[8]	6
Knowledge graph embedding by translating on hyperplanes	2014	[56]	6
DBpedia: A nucleus for a web of open data	2007	[5]	6
A framework and graphical development environment for robust NLP tools and applications	2002	[14]	6
KIM — a semantic platform for information extraction and retrieval	2004	[48]	6
KIM - semantic annotation platform	2003	[47]	5
Open user profiles for adaptive news systems: help or harm?	2007	[1]	5
Knowledge graph embedding via dynamic mapping matrix	2015	[30]	5

news standards into the emerging LOD cloud. The second part, from around 2015, sees a shift towards machine learning approaches, first focusing on NLP and embedding techniques and, since around 2019, on deep learning.

B.2 Citation analysis

Main paper citations: Table 11 shows the 15 most cited among our main papers, in recency-weighted order. Accordingly, Table 12 shows the papers that are referenced most frequently by our main papers. The two tables extend the ones in our review paper[A95].

Main paper citation graph: The directed graph in Figure 10 shows the 43 citations from one main paper to another.[11] The graph corroborates the citation analysis in our review paper. The two most "internally cited" papers are [M73] on DKN and [M15] on the NEWS project, with four citations each from other main papers. [M7,M71,M64,M29] are also cited three times each. [M7] is about Neptuno, whereas [M64,M29] are related to Hermes. [M71] is Troncy's paper about the IPTC's news architecture and the semantic web. When both incoming and outgoing citations are counted, this is the most central main paper, with three incoming and three outgoing citations from and to other main papers.

Evolution of the research area:

Figures 11–14 provide timelines to supplement the word clouds for the four main eras of research on KGs for news presented in Section 3.13 of our review paper.[12]

[11] Rendered with RDF Grapher, https://www.ldf.fi/service/rdf-grapher accessed 2022-03-10.

[12] All timelines depict three-year averages.

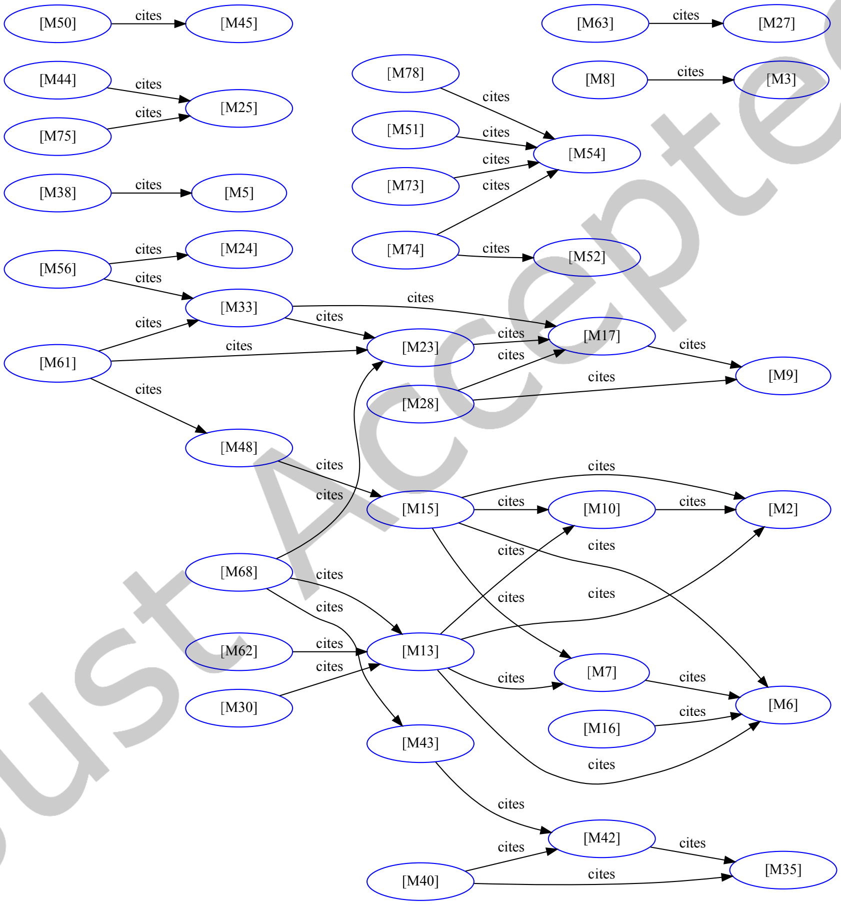
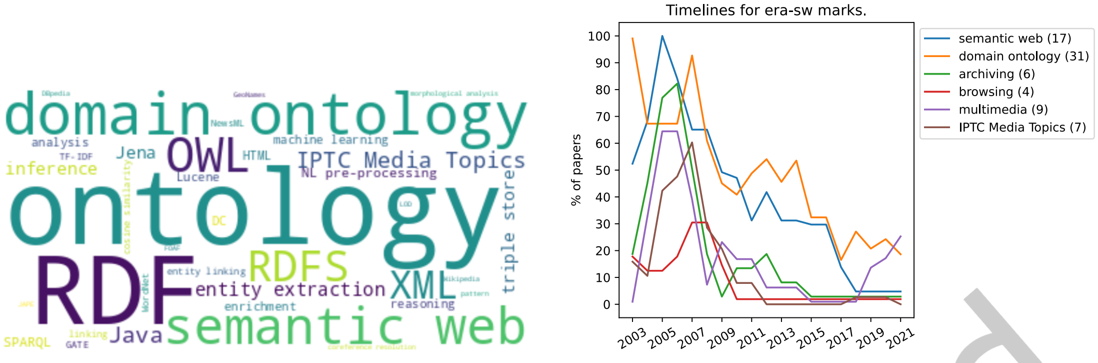
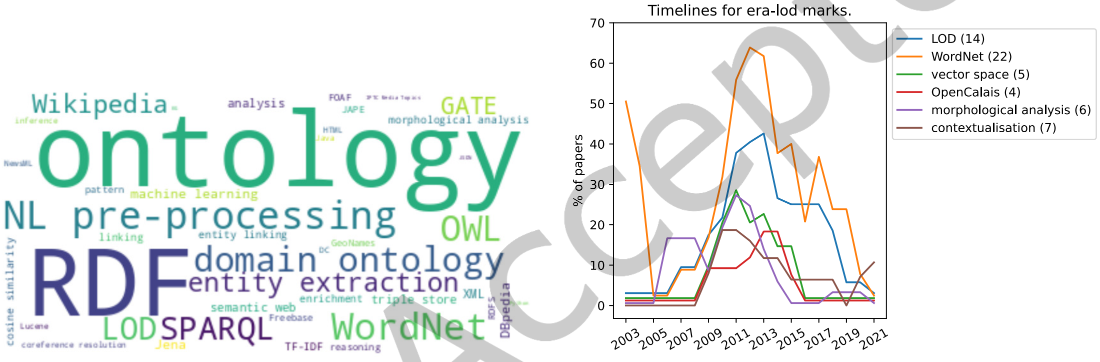


Fig. 10. Citations from one main paper to another.



(a) (b)

Fig. 11. The semantic-web era (until around 2009): (a) wordcloud and (b) timeline of typical themes.



(a) (b)

Fig. 12. The linked-open-data era (around 2010-2014): (a) wordcloud and (b) timeline of typical themes.

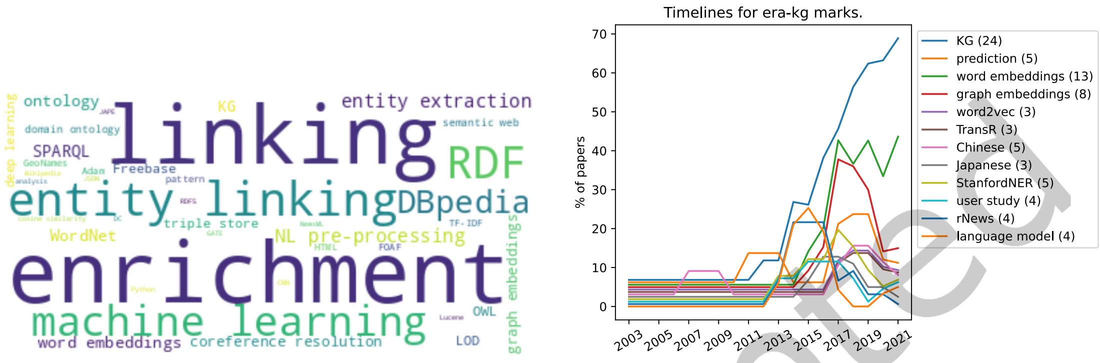
B.3 Related papers

In addition to the main papers presented in Section 3 of our review paper, we analysed several other closely related papers which, in the end, we had to exclude because they did not fully satisfy our inclusion and exclusion criteria.

Technical result types — Ontologies: One related contribution is [A19], which proposes a simple taxonomy of 48 OWL classes that can be used to describe the social context of social media articles with the aim of providing richer inputs to fake news detection. Concepts covered in the taxonomy include “social media agents” like “users” and “advertisers”, different types of “social media objects”, different types of “engagement”, and other metadata.

Technical result types — Knowledge graphs: The *ICIJ Offshore Leaks Database*[A128] represents data from the Panama Papers leaks in a Neo4J[A8] database, linked with DBpedia [5] and GeoNames[A51] and organised in a hierarchy of 65 classes, such as “company”, “offshore entity”, “country”, “person”, “agent”, “site”, etc. The graph makes information about almost 12M companies and 9.5M persons involved in the leaks available for complex querying using SPARQL. The authors present simple SPARQL queries that can provide an overview of and find trends in the dataset.

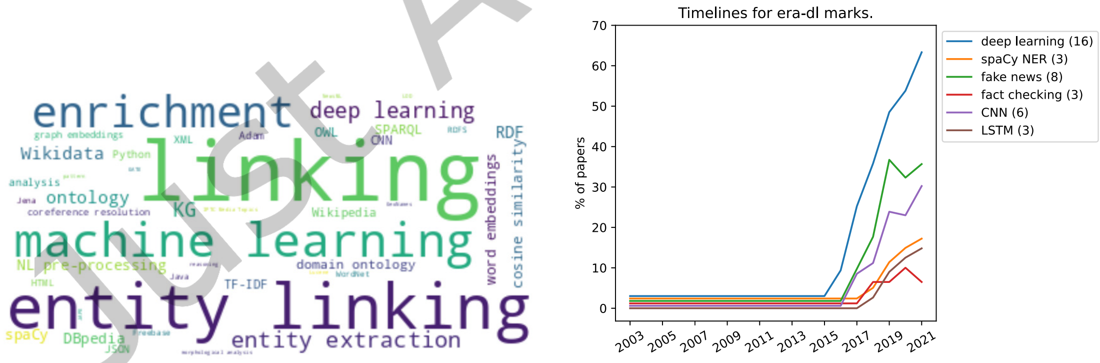
Intended users — Journalists, newsrooms, and news agencies: An early contribution is *RitroveRAI*[A20], a system that annotates multimedia news streams. News items are annotated with semantic metadata collected from topic categorisations or extracted from multimedia transcripts. Annotations are enriched with additional metadata lifted from sources such as HTML pages. The annotations are used to browse and retrieve news content and to maintain personal profiles that describe the interests of users.



(a) (b)

Fig. 13. The knowledge-graph era (from around 2015): (a) wordcloud and (b) timeline of typical themes.

Tasks — Content provision: Another related, but not main, paper describes the *Athena plug-in*, which extends Hermes [M4] with news recommendation capability[A64]. The approach is based both on user profiles of concepts found in the browsed news items and on ontology- and TF-IDF–supported matching. The more recent Bing-CSF-IDF+ recommender[A62] is also based on the Hermes framework.



(a) (b)

Fig. 14. The deep-learning era (from around 2019): (a) wordcloud and (b) timeline of typical themes.

Instead of recommending news content, the user agents in the EKA (Enterprise Knowledge Architecture)[A26] recommend like-minded *people* to one another, based on semantic representations of their news preferences and other interests.

Tasks — Event detection: [A127] uses a non-parametric Bayesian model to detect news events along with an unsupervised graph-embedding model to induce event schemas, which are then used to represent the detected events. To support updating of the Hermes News Portal (HNP) ontology [M64], OULx[A101] is proposed as an extension of the Ontology Update Language (OUL)[A74]. Many of the event-detection proposals build on the existing body of work on extracting events from texts[A60].

Tasks — Relation extraction: Another related contribution is [A76], which uses an ensemble for four tools (DBpedia Spotlight[A77], TagMe[A90], Babelfy[A81], and WAT[A12]) to extract named entities. Relations are extracted using ClausIE[A30] for dependency parsing and Mate-Tools[A94] for semantic-role labelling (SRL). In a final step, semantic roles and verb proximity are used to order entities and find appropriate names for events.

Tasks — Other: Related to *semantic similarity*, work on grouping news items by storyline includes [A50,A113]. Related to *interoperability*, RitroveRAI[A20] (mentioned earlier) supports multimedia data by annotating transcripts of broadcast news.

Input data — Multimedia news: [A55] describes a pipeline that inputs and transcribes TV streams and adapts classical natural-language processing (NLP) techniques to annotate the transcripts semantically. The annotations are used to describe programs; to segment them into topics; and to generate semantic hyperlinks in order to support semantic cross-media navigation. The pipeline has been used to demonstrate such semantic navigation of TV news content. And we have already mentioned RitroveRAI[A20] several times.

Other techniques and tools — Storage: Among the related papers, the Offshore Leaks Database[A128] uses Neo4J[A8] to represent data from the Panama Papers leaks in a KG.

News domain — Business and politics The knowledge-graph representation of data from the Panama Papers leaks[A128] addresses both business and political concerns.

FURTHER READINGS

[A1] 2012. PROTON. www.ontotext.com/documents/proton/Proton-Ver3.0B.pdf. [Online, accessed 2022-05. Supersedes the KIM Ontology (KIMO).].

[A2] 2016. Freebase. Previously http://www.freebase.com. [Offline since May 2016].

[A3] 2022. Ants. http://www.crit.rai.it/CritPortal/progetti/?p=249&lang=en.. [Online, accessed 2022-05].

[A4] 2022. FFmpeg — A complete, cross-platform solution to record, convert and stream audio and video. https://ffmpeg.org/. [Online, accessed 2022-06-02].

[A5] 2022. Keras. https://keras.io/. [Online, accessed 2022-05].

[A6] 2022. LinkedMDB. Available through https://data.world/linked-data/linkedmdb. [Online, accessed 2022-05].

[A7] 2022. MongoDB. https://www.mongodb.com/. [Online, accessed 2022-05].

[A8] 2022. Neo4J. https://neo4j.com/. [Online, accessed 2022-05].

[A9] 2022. OAuth. https://oauth.net/. [Online, accessed 2022-05].

[A10] 2022. ProbBank. https://propbank.github.io/. [Online, accessed 2022-05].

[A11] 2022. RDFlib. https://rdflib.readthedocs.io/en/stable/. [Online, accessed 2022-05].

[A12] 2022. WAT. https://sobigdata.d4science.org/web/tagme/wat-api. [Online, accessed 2022-05].

[A13] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 265–283.

[A14] ACM. 2022. ACM Digital Library. http://dl.acm.org. [Online, accessed 2022-05].

[A15] ADIMEN. 2022. Predicate Matrix. https://adimen.si.ehu.es/web/PredicateMatrix. [Online, accessed 2022-05].

[A16] Amazon. 2022. Amazon Web Services (AWS). https://aws.amazon.com/. [Online, accessed 2022-05].

[A17] Amazon. 2022. Neptune. https://aws.amazon.com/neptune/. [Online, accessed 2022-05].

[A18] Clarivate Analytics. 2022. Web of Science (WoS). http://webofknowledge.com. [Online, accessed 2022-05].

[A19] Anoud Bani-Hani, Oluwasegun Adedugbe, Elhadj Benkhelifa, Munir Majdalawieh, and Feras Al-Obeidat. 2020. A semantic model for context-based fake news detection on social media. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 1–7.

[A20] Roberto Basili, Marco Cammisa, and Emanuale Donati. 2005. RitroveRAI: A web application for semantic indexing and hyperlinking of multimedia news. In *The Semantic Web — ISWC 2005*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann

Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen (Eds.). Vol. 3729. Springer Berlin Heidelberg, Berlin, Heidelberg, 97–111. https://doi.org/10.1007/11574620_10

[A21] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475.

[A22] RSS Advisory Board. 2022. RSS 2.0 Specificaiton. https://www.rssboard.org/rss-specification. [Online, accessed 2022-05].

[A23] Uldis Bojars and John G. Breslin. 2022. SIOC Core Ontology Specification. http://rdfs.org/sioc/spec/. [Online, accessed 2022-05].

[A24] Dan Brickley and Libby Miller. 2022. FOAF. http://xmlns.com/foaf/spec/. [Online, accessed 2022-05].

[A25] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30, 1-7 (1998), 107–117.

[A26] Joshua Church and Ruhollah Farchtchi. 2008. Enterprise Knowledge Agents: Agent-based discovery of like-minded individuals. In *2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*. IEEE, Arlington, VA, USA, 231–238. https://doi.org/10.1109/CECandEEE.2008.96

[A27] U.S. Central Intelligence Agency (CIA). 2022. CIA World Factbook. https://www.cia.gov/library/publications/the-world-factbook/. [Online, accessed 2022-05].

[A28] Francesco Corcoglioniti, Marco Rospocher, Roldano Cattoni, Bernardo Magnini, and Luciano Serafini. 2015. The KnowledgeStore: A storage framework for interlinking unstructured and structured knowledge. *International Journal on Semantic Web and Information Systems* 11, 2 (April 2015), 1–35. https://doi.org/10.4018/IJSWIS.2015040101

[A29] Dublin Core Metadata Initiative (DCMI). 2022. Dublin Core (DC) — Metadata Innovation. https://www.dublincore.org/. [Online, accessed 2022-05].

[A30] Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*. 355–366.

[A31] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 213–220.

[A32] Elsevier. 2022. ScienceDirect. http://www.sciencedirect.com/. [Online, accessed 2022-05].

[A33] Jesse English and Sergei Nirenburg. 2007. Ontology learning from text using automatic ontological-semantic text annotation and the Web as the corpus. In *AAAI Spring Symposium: Machine Reading*. 43–48. Section 2.

[A34] EODData. 2022. NASDAQ codes. Available from https://www.eoddata.com/stocklist/nasdaq.htm?AspxAutoDetectCookieSupport=1. [Online, accessed 2022-05].

[A35] ExplosionAI. 2022. spaCy — Industrial-Strength Natural Language Processing in Python. https://spacy.io/. [Online, accessed 2022-05].

[A36] AI2 Allen Institue for AI. 2022. Semantic Scholar — A free, AI-powered research tool for scientific literature. https://www.semanticscholar.org/. [Online, publication metadata downloaded February-March 2022].

[A37] International Organization for Standardization (ISO). 2022. Common Logic. https://www.iso.org/standard/66249.html. [Online, accessed 2022-05].

[A38] International Organization for Standardization (ISO). 2022. ISO Country Codes. https://www.iso.org/iso-3166-country-codes.html. [Online, accessed 2022-05].

[A39] Eclipse Foundation. 2022. RDF4J. https://rdf4j.org/. [Online, accessed 2022-05].

[A40] MetaBrainz Foundation. 2022. MusicBrainz. https://musicbrainz.org/. [Online, accessed 2022-05].

[A41] OpenID Foundation. 2022. OpenID. https://openid.net/. [Online, accessed 2022-05].

[A42] The Apache Software Foundation. 2022. Giraph. https://giraph.apache.org/. [Online, accessed 2022-05].

[A43] The Apache Software Foundation. 2022. HBase. http://hbase.apache.org/. [Online, accessed 2022-05].

[A44] The Apache Software Foundation. 2022. Jena. https://jena.apache.org/. [Online, accessed 2022-05].

[A45] The Apache Software Foundation. 2022. Lucene. https://lucene.apache.org/. [Online, accessed 2022-05].

[A46] The Apache Software Foundation. 2022. OpenNLP. https://opennlp.apache.org/. [Online, accessed 2022-05].

[A47] Wikimedia Foundation. 2022. Wikidata. http://www.wikidata.org. [Online, accessed 2022-05].

[A48] Wikimedia Foundation. 2022. Wikinews. https://www.wikinews.org/. [Online, accessed 2022-05].

[A49] Wikimedia Foundation. 2022. Wikipedia. https://www.wikipedia.org/. [Online, accessed 2022-05].

[A50] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th International Conference on World Wide Web*. 482–490.

[A51] Unxos Gmbh. 2022. GeoNames. https://www.geonames.org/about.html. [Online, accessed 2022-05].

[A52] Google. [n.d.]. Google Scholar. http://scholar.google.com/.

[A53] Google. 2022. Google Knowledge Graph API. https://developers.google.com/knowledge-graph. [Online, accessed 2022-05].

[A54] Google. 2022. Google StreetView. https://www.google.com/streetview/. [Online, accessed 2022-06-02].

[A55] Guillaume Gravier, Camille Guinaudeau, Gwénolé Lecorvé, and Pascale Sébillot. 2011. Exploiting speech for automatic TV delinearization: From streams to cross-media semantic navigation. *EURASIP Journal on Image and Video Processing* 2011 (2011), 1–17.

https://doi.org/10.1155/2011/689780
[A56] MilaGraph Group. 2022. Wikidata5M. https://deepgraphlearning.github.io/project/wikidata5m. [Online, accessed 2022-05].
[A57] Stanford University NLP Group. 2022. StanfordNER and -NLP. https://nlp.stanford.edu/. [Online, accessed 2022-05].
[A58] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194 (2013), 28–61.
[A59] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 782–792.
[A60] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems* 85 (2016), 12–22.
[A61] Frederik Hogenboom, Viorel Milea, Flavius Frasincar, and Uzay Kaymak. 2010. RDF-GL: a SPARQL-based graphical query language for RDF. In *Emergent Web Intelligence: Advanced Information Retrieval*. Springer, 87–116.
[A62] Lies Hooft van Huijsduijnen, Thom Hoogmoed, Geertje Keulers, Edmar Langendoen, Sanne Langendoen, Tim Vos, Frederik Hogenboom, Flavius Frasincar, and Tarmo Robal. 2020. Bing-CSF-IDF+: A semantics-driven recommender system for news. In *European Conference on Advances in Databases and Information Systems*. Springer, 143–153.
[A63] IEEE. 2022. IEEE Xplore. http://ieeexplore.ieee.org/. [Online, accessed 2022-05].
[A64] Wouter IJntema, Frank Goossen, Flavius Frasincar, and Frederik Hogenboom. 2010. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*. 1–6.
[A65] Franz Inc. 2022. AllegroGraph. https://allegrograph.com/. [Online, accessed 2022-05].
[A66] International Press Telecommunications Council (IPTC). 2022. IPTC General Architecture Framework (GAF). https://iptc.org/standards/news-architecture/. [Online, accessed 2022-05].
[A67] International Press Telecommunications Council (IPTC). 2022. IPTC Media Topics. https://iptc.org/standards/media-topics/. [Online, accessed 2022-05].
[A68] International Press Telecommunications Council (IPTC). 2022. IPTC News Codes. https://iptc.org/standards/newscodes/. [Online, accessed 2022-05].
[A69] International Press Telecommunications Council (IPTC). 2022. IPTC rNews vocabulary. https://iptc.org/standards/rnews/. [Online, accessed 2022-05].
[A70] International Press Telecommunications Council (IPTC). 2022. IPTC Subject Reference System. https://iptc.org/standards/subject-codes/, succeeded by the IPTC Media Topics, https://iptc.org/standards/media-topics/. [Online, accessed 2022-05].
[A71] International Press Telecommunications Council (IPTC). 2022. News Industry Text Format (NITF). https://iptc.org/standards/nitf/. [Online, accessed 2022-05].
[A72] International Press Telecommunications Council (IPTC). 2022. NewsML-G2. https://iptc.org/standards/newsml-g2/. [Online, accessed 2022-05].
[A73] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters Tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1483–1493.
[A74] Uta Lösch, Sebastian Rudolph, Denny Vrandečić, and Rudi Studer. 2009. Tempus fugit. In *European Semantic Web Conference*. Springer, 278–292.
[A75] Grzegorz Malewicz, Matthew H Austern, Aart JC Bik, James C Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. 2010. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. 135–146.
[A76] Jose L. Martinez-Rodriguez, Ivan Lopez-Arevalo, Ana B. Rios-Alvarado, Julio Hernandez, and Edwin Aldana-Bobadilla. 2019. Extraction of RDF statements from text. In *Knowledge Graphs and Semantic Web*, Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado (Eds.). Vol. 1029. Springer International Publishing, 87–101. https://doi.org/10.1007/978-3-030-21395-4_7 Series Title: Communications in Computer and Information Science.
[A77] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. 1–8.
[A78] Microsoft. 2022. Bing News. https://www.bing.com/news. [Online, accessed 2022-06-02].
[A79] Microsoft. 2022. LightGBM. https://lightgbm.readthedocs.io/en/latest/. [Online, accessed 2022-05].
[A80] Microsoft. 2022. MIND — MIcrosoft News Dataset. https://msnews.github.io/. [Online, accessed 2022-06-02].
[A81] Andrea Moro and Roberto Navigli. 2022. Babelfy. http://babelfy.org/. [Online, accessed 2022-05].
[A82] Daniele Nardi, Ronald J Brachman, et al. 2003. An introduction to description logics. *Description Logic Handbook* 1 (2003), 40.
[A83] Springer Nature. 2022. SpringerLink. http://link.springer.com/. [Online, accessed 2022-05].
[A84] Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems-Volume 2001*. 2–9.

[A85] University of Berkeley. 2022. FrameNet. https://framenet.icsi.berkeley.edu/fndrupal/. [Online, accessed 2022-05].

[A86] University of Colorado. 2022. VerbNet. https://verbs.colorado.edu/verbnet/. [Online, accessed 2022-05].

[A87] University of Leipzig. 2022. NLP Interchange Format (NIF) 2.0 — Overview and Documentation. https://persistence.uni-leipzig.org/nlp2rdf/. [Online, accessed 2022-06-02].

[A88] University of Lübeck. 2022. Racer. https://www.ifis.uni-luebeck.de/~moeller/racer/. [Online, accessed 2022-05].

[A89] University of Oxford. 2022. HermiT. http://www.hermit-reasoner.com/. [Online, accessed 2022-05].

[A90] University of Pisa. 2022. TagMe. https://tagme.d4science.org/tagme/. [Online, accessed 2022-05].

[A91] University of Sheffield. 2022. ANNIE (A Nearly New Information Extraction system). https://gate.ac.uk/ie/annie.html. [Online, accessed 2022-05].

[A92] University of Sheffield. 2022. GATE (General Architecture for Text Engineering). http://gate.ac.uk/. [Online, accessed 2022-05].

[A93] University of Sheffield. 2022. JAPE (Java Annotations Pattern Engine). https://gate.ac.uk/sale/tao/splitch8.html. [Online, accessed 2022-05].

[A94] University of Stuttgart. 2022. Mate-Tools. https://www.ims.uni-stuttgart.de/en/research/resources/tools/matetools/. [Online, accessed 2022-05].

[A95] Andreas L Opdahl, Tareq Al-Moslmi, Duc-Tien Dang-Nguyen, Marc Gallofré Ocaña, Bjørnar Tessem, and Csaba Veres. [n.d.]. Semantic knowledge graphs for the news: A review. *Comput. Surveys* ([n. d.]).

[A96] Oracle. 2022. MySQL. https://www.mysql.com/. [Online, accessed 2022-05].

[A97] Adam Pease. 2022. Suggested Upper Merged Ontology (SUMO). https://www.ontologyportal.org/. [Online, accessed 2022-05].

[A98] The GDELT Project. 2022. GDELT — Watching Our World Unfold. https://www.gdeltproject.org/. [Online, accessed 2022-05].

[A99] The NewsReader Project. 2022. ESO. http://www.newsreader-project.eu/results/event-and-situation-ontology/. [Online, accessed 2022-05].

[A100] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2017. BreakingNews: Article annotation by image and text processing. *IEEE transactions on pattern analysis and machine intelligence* 40, 5 (2017), 1072–1085.

[A101] Jordy Sangers, Frederik Hogenboom, and Flavius Frasincar. 2012. Event-driven ontology updating. In *International Conference on Web Information Systems Engineering*. Springer, 44–57.

[A102] SemLab. 2022. ViewerPro. https://www.semlab.nl/portfolio-item/viewerpro-semantic-text-analysis/. [Online, accessed 2022-05].

[A103] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. The Hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. Ieee, 1–10.

[A104] Amit Singhal. 2012. Introducing the knowledge graph: things, not strings. Official Google Blog. https://blog.google/products/search/introducing-knowledge-graph-things-not/ [Online, accessed 2022-05-30].

[A105] OpenLink Software. 2022. Virtuoso. https://virtuoso.openlinksw.com. [Online, accessed 2022-05].

[A106] Robyn Speer et al. 2022. ConceptNet 5.5. https://conceptnet.io/. [Online, accessed 2022-05].

[A107] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An open multilingual graph of general knowledge. arXiv preprint arXiv:1612.03975.

[A108] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: an overview. In *Treebanks*. Springer, 5–22.

[A109] Thomson-Reuters. 2022. OpenCalais. http://www.opencalais.com. [Online, accessed 2022-05].

[A110] Twitter. 2022. Twitter. https://about.twitter.com/. [Online, accessed 2022-05].

[A111] Vrije Universiteit. 2022. Simple Event Model (SEM). https://semanticweb.cs.vu.nl/2009/11/sem/. [Online, accessed 2022-05].

[A112] Stanford University. 2022. Protégé — A free, open-source ontology editor and framework for building intelligent systems. https://protege.stanford.edu/. [Online, accessed 2022-05].

[A113] Arnout Verheij, Allard Kleijn, Flavius Frasincar, and Frederik Hogenboom. 2012. A comparison study for novelty control mechanisms applied to web news stories. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 431–436.

[A114] World Wide Web Consortium (W3C). [n.d.]. PRISM — the Publishing Requirements for Industry Standard Metadata. https://www.w3.org/Submission/prism/.

[A115] World Wide Web Consortium (W3C). 2022. Composite Capability/Preference Profiles (CC/PP). https://www.w3.org/TR/CCPP-struct-vocab/. [Online, accessed 2022-05].

[A116] World Wide Web Consortium (W3C). 2022. Internationalization Tag Set (ITS). https://www.w3.org/TR/its20/. [Online, accessed 2022-05].

[A117] World Wide Web Consortium (W3C). 2022. OWL. http://www.w3.org/TR/owl-ref/. [Online, accessed 2022-05].

[A118] World Wide Web Consortium (W3C). 2022. OWL-DL. https://www.w3.org/TR/2012/REC-owl2-profiles-20121211/. [Online, accessed 2022-05].

[A119] World Wide Web Consortium (W3C). 2022. OWL Time. https://www.w3.org/TR/owl-time/. [Online, accessed 2022-05].

[A120] World Wide Web Consortium (W3C). 2022. PROV-O. https://www.w3.org/TR/prov-o/. [Online, accessed 2022-05].

[A121] World Wide Web Consortium (W3C). 2022. RDF. http://www.w3.org/TR/rdf11-concepts/. [Online, accessed 2022-05].

[A122] World Wide Web Consortium (W3C). 2022. RDFS. https://www.w3.org/TR/rdf-schema/. [Online, accessed 2022-05].

[A123] World Wide Web Consortium (W3C). 2022. SKOS. http://www.w3.org/2009/08/skos-reference/skos.html. [Online, accessed 2022-05].

[A124] World Wide Web Consortium (W3C). 2022. SPARQL. http://www.w3.org/TR/sparql11-query/. [Online, accessed 2022-05].

[A125] World Wide Web Consortium (W3C). 2022. WordNet in RDF. https://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html. [Online, accessed 2022-05].

[A126] Yahoo. 2022. Yahoo! News. https://news.yahoo.com/. [Online, accessed 2022-05].

[A127] Quan Yuan, Xiang Ren, Wenqi He, Chao Zhang, Xinhe Geng, Lifu Huang, Heng Ji, Chin-Yew Lin, and Jiawei Han. 2018. Open-schema event profiling for massive news corpora. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy (CIKM '18)*. ACM, New York, NY, USA, 587–596. https://doi.org/10.1145/3269206.3271674

[A128] Lily Popova Zhuhadar and Mark Ciampa. 2021. Novel findings of hidden relationships in offshore tax-sheltered firms: a semantically enriched decision support system. *Journal of Ambient Intelligence and Humanized Computing* 12, 4 (2021), 4377–4394.