Harnessing Il Manifesto Newspaper Archive for Knowledge Base Creation: Techniques and Findings in the MeMa Project

Robert J. Alexander[1], Matteo Bartocci[2], Oriana Persico[3] and Guido Vetere[1,4]

[1]Isagog S.R.L., Via Faà di Bruno 54, 00195 Roma, Italy
[2]Il Manifesto Soc. Coop., Via Angelo Bargoni 8, 00153 Roma, Italy
[3]Human Ecosystems Relazioni S.R.L, via Umberto Guarnieri 15, 00177 Roma, Italy
[4]Università Guglielmo Marconi, Via Plinio 44, 00193 Roma, Italy

Abstract

English.The historical archive of the newspaper "il Manifesto" is a valuable asset protected by the Italian Ministry of Cultural Heritage. The *MeMa* project aims to create an "intelligent archive" using AI principles, fostering collaboration and transparency. The platform, built around Apache Jena and open linguistic technologies, addresses the newspaper community's specific needs. This paper presents the platform's architecture, knowledge base construction process, and future directions, emphasizing journalism enhancements through AI while respecting "Il Manifesto"'s principles. **Italiano.**L'archivio storico del quotidiano "il Manifesto" è tutelato dal Ministero dei Beni Culturali. Il progetto *MeMa* mira a creare un "archivio intelligente" basato su una intelligenza artificiale che favorisce la collaborazione e la trasparenza. La piattaforma, costruita attorno ad Apache Jena e tecnologie linguistiche aperte, risponde alle esigenze specifiche della comunità del giornale. Questo contributo presenta l'architettura della piattaforma, il processo di costruzione della base di conoscenza e le direzioni future, discutendo il potenziamento del giornalismo attraverso l'intelligenza artificiale nel rispetto dei principi de "Il Manifesto".

Keywords

AI in journalism, Open linguistic technologies, Knowledge graphs, Newspaper community

1. Introduction

The historical archive of the newspaper "il Manifesto" is an asset protected by the Italian Ministry of Cultural Heritage as of particular interest [1]. The archive includes a paper collection starting from 1971, and a digitized collection starting from the 1990s. The resource is now entrusted to the "Nuovo Manifesto Società Cooperativa Editrice", which publishes the newspaper and its digital editions since 2013. The cooperative is committed to maintain and improve the archive, as well as to guarantee free access and digital consultation facilities to anyone interested in it [2]. The digital archive, produced in different phases over the years, reflects the historical and technological evolution of the publishing sector. The database initially included 10,013 digitized files containing about 160,000 articles, with few gaps in the years 1985-1986 and 1994-2002. Il Manifesto considers an "intelligent archive" to be the cornerstone of its digital strategy, and for this reason seeks to align it with new technologies with appropriate investments in research and development. The *MeMa* (Memoria Manifesta) project started in 2020 by a partnership with Salvatore Iaconesi [3] and Oriana Persico, with the aim of developing new archive infrastructure based on Artificial Intelligence. This would be a "Community AI" [1] based on the principles of openness, transparency, collaboration and non-extractiveness, thus being able to establish productive relationships between the archive, the editorial staff, the user communities and society in general [2].

When, in 2023, the project was resumed, the new board decided to continue the original plan by making it evolve in the direction of Linked Open Data, and taking advantage of the latest advances in language and knowledge technologies. The idea was to build a standards-based Knowledge Graph (KG) using editorial metadata and structured information extracted from article text. By itself, this idea is by no means new [3] [4] [5]. Also, there are commercial platforms that have been offering solutions for the newspaper industry some years now, such as Neo4j [6] or Ontotext [7]. However, we realized that the success of the project depended significantly on how the platform would adapt to the way content is produced, extracted, organised, enriched and experienced by the professional and user communities gathered around

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
bob@isagog.com (R.J. Alexander); bartocci@ilmanifesto.it (M. Bartocci); oriana.persico@he-r.it (O. Persico); g.vetere@isagog.com (G. Vetere)
0000-0002-6703-7276 (G. Vetere)

[1]Legislative Decree 42/2004 with provision of the Regional Director for the Cultural Heritage of Lazio (24/2013, 12 March 2013)
[2]https://archiviopubblico.ilmanifesto.it/

[3]Salvatore Iaconesi (Livorno 1973, Reggio Calabria 2022) has been an engineer, artist, hacker and interaction designer

the newspaper. Rather than forcing these habits to an out-of-the-box commercial platform, we opted to tailor a specific solution. Moreover, as a sociotechnical platform, *MeMa* should be open to user curation and contribution (e.g. from readers, archivists, and journalists), collaboratively contributing to the evolution of the AI, including correcting the inevitable errors of current NLP technologies. Hence, we started designing a custom platform around a core open graph database, namely Apache Jena [4] and a selection of open linguistic technologies suitable for the Italian language. The solution falls into the broad area of Enterprise Knowledge Graphs [8] which are gaining momentum as "rational counterparts" of generative linguistic technologies based on neural models [9]. This work is a first account of what emerged in the first months of analysis, design and development of the solution, and a discussion of our plans to meet the socio-technical requirements we have analyzed so far. Our contribution is a "reality check" of the use of knowledge and language technologies applied to complex texts produced by an Italian publishing community over more than 40 years of work. In general, our research concerns the interaction between digital systems and human beings to make their contents fully transparent and accessible to different user communities. From a linguistic point of view, relevant aspects include the specificity of the texts produced over a wide period of time, characterized by a specific idiolect but also by diachronic variations.

This paper is organized as follows. In Section 2, we present an architectural overview of the platform under development. Section 3 delves into the process of constructing the knowledge base, detailing the steps involved in gathering and organizing the relevant information. In Section 4, we discuss challenges and ideas about the future directions. Note that automatic content generation is not included in the journalism enhancements driven by AI, as intended by "Il Manifesto".

2. System Overview

MeMa's software architecture comprises several components that work together to handle a graph database with indexed attributes, enabling efficient ingestion, analysis, and semantic querying. The key components of this architecture include:

1. Knowledge Graph: The core of the system is a graph database of the RDF (Resource Description Framework) family with inference capabilities, based on Apache Jena, the Pellet OWL reasoner, the search engine Lucene, and custom components, where a number of KG attributes are indexed and embedded to optimize search and retrieval operations.
2. NLP Service: A REST service that provides an abstraction layer over various NLP functionalities to support the system's operations. It wraps capabilities such as text analysis, entity recognition, topic analysis, semantic similarity, and other NLP tasks based on open source transformers [10]. This service collaborates with the ingestion process to extract valuable insights from the content being ingested.
3. Ingestion Processor: A batch process that is responsible for ingesting content into the KG. This process integrates different sources, analyzes texts to extract relevant information using the NLP service, and produces RDF sources to feed the KG according to the *MeMa* ontology.
4. Query and Update Service: A REST service that is responsible for handling queries and update operations on the KG. It integrates similarity searches and SPARQL queries to retrieve relevant graph entities. This service leverages the indexed attributes to optimize query performance and speed up retrieval operations, and the NLP Service to transform user's queries and evaluate response ranking.

This software architecture employs a services and API-based approach, enabling functional evolution, flexible deployment, and seamless scalability. The service architecture is an abstraction of a general functionality that can be applied to a variety of scenarios. Based on this design, we have developed custom application services that can be used in a front-end designed for the editorial staff of the newspaper.

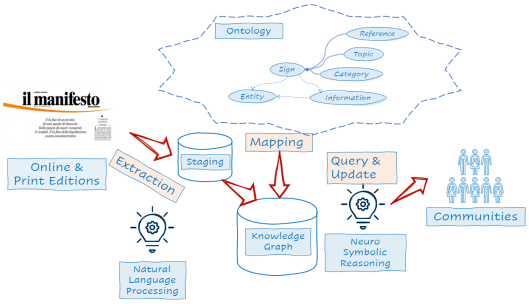

Figure 1: Architectural overview

"Il Manifesto" has a print edition and an online edition, each managed by its own Content Management System (CMS). The two editions largely coincide, however each one may contain articles not present in the other. As a result, the same article (with slight variations) may be available in two different repositories. When consolidating all editorial content into one Knowledge Base, we had to harmonize and integrate the contents from both CMSs.

[4] https://jena.apache.org/

3. The Knowledge Base

Modeling editorial content in a KG requires the adoption of a suitable ontology. Although editorial content modeling has already been studied and tested [11], we did not identify a simple, well-established model that suited our needs. In particular, we aimed to represent how agents interpret specific tokens as referring to entities based on established conventions or procedures. In other words, we were interested in semiotics. At the best of our knowledge, even comprehensive conceptualizations, like the CIDOC Conceptual Reference Model [12], which include linguistic and symbolic objects, do not provide modeling primitives to represent interpretation processes. This is why we decided to develop our own conceptualization, which we will illustrate in the following section. Mappings to existing conceptual frameworks, such as schema.org[5], are preserved as annotations.

3.1. The *MeMa* Ontology

The *MeMa* ontology focuses on the way entities are mentioned, rather than on the characterization of those entities, which is mostly left to external sources. As such, the *MeMa* ontology adopts a semiotic perspective [13] in the line of [14] and [15]. The structure of our ontology is sketched as follows:

- Class: Sign
  An immaterial entity that stands to someone (or something) for some other entity as the outcome of an interpretation
  - Subclass: Category
    A sign standing for a class of entities
  - Subclass: Reference
    A sign standing for a single (even collective) entity
  - Subclass: Topic
    A sign standing for a focus of interest in a larger context
- Class: Information
  An immaterial thing that conveys interconnected signs
  - Subclass: Text
    A textual information object
  - Subclass: Sentence
    Part of a text
  - Subclass: Token
    Part of a sentence
- Class: Entity
  A spatio-temporal thing
  - Subclass: Agent
    An entity that has the capacity to initiate or perform actions
  - Subclass: Location
    An identified portion of space
  - Subclass: Event
    An entity that unfolds in time
  - Subclass: Object
    An entity that unfolds in space

A key feature of this ontology is the distinction of Reference and Token, where the latter instantiates the former [6]. As a Sign, a Reference is based on an interpretation process, whether human or automated, e.g., for *DBpedia Spotlight*, interpreting the string "Aristotle" as the name of the philosopher from Stagira. Sign instances support properties (*interpretation records*) that keep track of these processes. A Token, on the other hand, is a portion of Text, e.g. the string "Aristotle" that appears in a document at a given offset, which may trigger the processes mentioned above. In this way, the semantic qualification of the text is provided with the means to trace the underlying interpretation, be it automatic or human. This is essential for ensuring the traceability and accountability of the knowledge base's content.

3.2. Handling Metadata

Extracting knowledge from newspaper articles essentially consists of working on the both metadata and text in a consistent way. This process has currently generated about 650.000 stored articles and grows roughly by 1000 new articles a month.

[5] https://schema.org/

[6] This aligns with Peirce's distinction of *type* and *token*

According to our ontology, assertions about articles are based on two types of properties, which we call *editorial* and *semantic*. The former includes attributes such as *publication date* or *author*, the latter are generically intended to characterize the content, including standard categorization (*sports*, *business*, etc.), references to people, places and other named entities, and arbitrary classifiers which are typically encoded in freely invented wording. However, this distinction is neither fully aligned with the structure of the legacy metadata schemes, nor fully reflected in how metadata are actually produced. For historical and organizational reasons, in fact, the online and print editions are metadated separately, with different schemes and guidelines. Looking into it, we realized that integrating them could not be done by simply mapping schemes to our ontology, but instead required a thoughtful analysis of the actual data. We carried out qualitative and quantitative analyses which led us to devise an adequate treatment of the metadata content. Here is a summary of the historical archive scheme:

- **ARGOMENTO** (subject) is fed with labels with no semantic relationship amongst them. The raw count for these labels is 792.000 with **4023 distinguished values (0.51%)**, which comprise synonyms, typos, abbreviations, and other variants.
- **CATEGORIA** (category) field, on the other hand, is used with a prevalence of editorial tags (*front page*, *editorial* etc) but again we often encounter values that also belong to the ARGOMENTO field. The raw count usage for CATEGORIA is 828.805, with **1358 different values (0.16%)**, which also comprise synonyms, typos, abbreviations, and other variants.
- **LOCALITA** (location) accommodates editor's or archivist description of what geopolitical entities are involved. They might not be mentioned literally in the article. We observed redundant tagging where many broader geopolitical concepts, which could be inferred, are explicitly stated somewhat arbitrarily (e.g., CUTRO, CR, Italia). Whenever we successfully link a geopolitical mention to GeoNames, this redundancy becomes unnecessary, as GeoNames allows for full hierarchical navigation.
- **RIFERIMENTI** (references) is used as a placeholder for a variety of annotations, which also overlap other fields. Most often, these are short summaries which should facilitate keyword based retrieval. We currently count 949248 occurrences of these annotations, **679760 of which are unique (71,6%)**, thus qualifying by far as the most informative facet.

Overall, the frequency distribution of all these properties exhibits long tails with low frequencies typical of a lack of annotation guidelines and tools. In particular, the RIFERIMENTI field appears to be very heterogeneous, as it mixes editorial tags (e.g. *breve*, *cronaca*), named entities and content summaries. As a result of this analysis, we decided to ignore the formal meaning (if any) of the legacy metadata schema and instead focus on the annotation content. In particular, with respect to our ontology, we want to distinguish among *classifiers* (Sign) and *descriptions* (Information). To this end, we use:

- Two handcrafted tagsets, for editorial marks and standard topics respectively, obtained by clearing and deduplicating the contents of ARGOMENTO, CATEGORIA and the most recurrent RIFERIMENTI
- A lemmatizer for out of tagset values
- A rule-based classifier for multi-word RIFERIMENTI values, which discriminates *descriptions* from multi-word topics

Classifiers are instantiated as either as Category or Topic, and suitably linked to the article, while descriptive summaries are kept as data properties, whose content is indexed. We plan to add a vector representation of summaries to include them in semantic similarity searches and/or clustering.

3.3. Knowledge Extraction

Besides annotated metadata, *MeMa* analyzes the full article text. At the current stage, we only perform entity recognition and linking. There are no limits to the kind of entities that can be mentioned in a newspaper article. However, there are limits to the kinds that can be efficiently retrieved by standard NLP pipelines. One of the richest known inventories [16], includes up to 18 categories, but as a matter of facts the available recognizers for the Italian language, e.g. Spacy [17] and Stanza [18] are limited to just a few of them, such as **PER**(son), **LOC**(alization), and **ORG**(anization). We currently use a combination of Stanford's Stanza [18] (in particular: `tokenize`, `mwt`, `pos`, `lemma`, `depparse`, and `ner` processors), DBPedia Spotlight [19], GeoNames [7], along with a number of custom processing functions. We choose Stanza because of the state-of-the-art performances on Italian benchmarks[8]. We evaluated the NER performance on our sources by randomly choosing 30 articles, manually annotating their content, and matching the pipeline outcome. Results presented in Table 2 align with the current state of the art [20].

For the **PER** class we also adopt a simple co-referencing matching based on the fact that within an article we mostly find a fully named instance of the person and subsequently only the first or last names. Along with the span, we therefore generate a Person co-refernce ID. We then proceed to the grounding attempt against the DBpedia API which we invoke via its Spotlight function. We have found no added precision/recall by giving it more textual context. For both the grounded and the

[7] https://www.geonames.org/

[8] Stanza's performance on NER Corpora https://stanfordnlp.github.io/stanza/ner_models.html

annotation	occurrences
breve (*short*)	5324
cronaca (*news*)	1860
analisi (*analisys*)	901
programma (*program*)	732
scheda (*form*)	691
crisi (*crisis*)	688
scenario (*scenario*)	671
le lettere di oggi (*today's letters*)	662
storia (*history*)	648
ritratto (*portrait*)	575
campagna elettorale (*election campaign*)	564
reazioni (*reactions*)	544
famiglia incertezza e preocupazioni (sic) (*family uncertainty and worries*)	1
oggi sciopero marcia globale per il clima (*global climate march strike today*)	1
giorgio forti, alessandro stoppoloni, christian picucci (proper names)	1

Table 1
An excerpt of both recurrent and unique values of RIFERIMENTI

Type	Precision	Recall	F1 Score
PER	0.9117	0.9612	0.9280
LOC	0.9194	0.8703	0.8763
ORG	0.8071	0.8213	0.7847
Overall	0.8816	0.8868	0.8657

Table 2
Average Precision, Recall, and F1 Score per Type and Overall

ungrounded **PER**sons, we then store the span of surface, a fuzzy score of the match with DBpedia's entity to accommodate typos and variations which are especially common with the Italian rendition of foreign names and the reference to the current article. We therefore have the spans where the surface of the person was mentioned and the grounded/ungrounded reference to the article in a separate collection. A similar process is performed for the **LOC**ation named entities against the GeoNames resource. Linking to the GeoNames resource gives us a wealth of added information amongst which geolocalization and administrative and geographical data. Also for **LOC** we store the spans within the article's and the mentions in their dedicated collection. We also tried using DBpedia Spotlight for **ORG**anizations but the results were not satisfactory. One of the causes may be the lack of precision at the **NER** stage. Also, there are often false positive groundings given that there are several organizations with namesakes or placenames. We didn't conduct a comprehensive analysis of the entity linking performance; however, an initial examination revealed that roughly 10% of the total links were incorrect. Finally, the last stages of our pipeline transforms the staging data into corresponding RDF data (Turtle format). We therefore generate article individuals with metadata from both the historical and the digital corpora leveraging the reconciliation when possible and we also generate individuals, topics and all of their cross-linked mentions. The resulting knowledge base is currently expressed with approximately 12.5 million triples, and loaded into Apache Jena Fuseki to be used as a SPARQL endpoint.

4. Challenges and Ideas

Newspaper articles pose several interpretative challenges [21]. The reporting of events, with their participants and their contextual characterization, are the most relevant parts of their content. Metonymy, regular polysemy and presupposition, even combined, stand out as prominent linguistic phenomena. Take for instance the headline: "Di Maio al Colle, ma non da Mattarella" (≈*"Di Maio at the Colle, but not meeting with Mattarella"*) [9]. "Di Maio" and "Mattarella" can be plainly identified as person mentions and linked to their corresponding individuals (Italian politicians). But what about "Colle"? Even if it were identified as a place (the Quirinal hill in Rome) it is clear that, contextually, the token intends to signify the institutional function of the presidency of the Italian Republic. Also, the people mentioned in the sentence represent their public roles at the time the article was written, rather than any identified human being. This kind of metonymic use of language makes classification of named entities more difficult [22]. As for the news

[9] https://ilmanifesto.it/di-maio-al-colle-ma-non-da-mattarella

in question, note that apparently there is no mention of any event, but presumably something happened. Event mining is also a long-standing challenge of NLP, as well as reasoning about implicature and presupposition [23]. These tasks are usually approached with ML methods [24]. In particular, supervised learning strategies have been implemented in recent years, but they are limited in that they require specific annotated corpora and training procedures [25].

The use of Large Language Models (LLM) with generative transformers are now regarded to as the way forward for the challenges mentioned above [26], so we plan to replace part of the existing NLP machinery with a suitable implementation of these techniques. At present, we are piloting publicly available services. For instance, if suitably promped, GPT4 [27] produces the following analysis of the headline mentioned d above:

1. Di Maio: *This refers to a person, likely Luigi Di Maio, an Italian politician. He could be the subject of the discussion or the context in which the sentence is used.*
2. at the Colle: *Refers to "Colle" or "Quirinale", which is the official residence of the President of the Italian Republic, located in Rome. Therefore, "Di Maio at the Colle" indicates that Luigi Di Maio is at or going to the Quirinale Palace.*
3. Mattarella: *Refers to Sergio Mattarella, who was the President of the Italian Republic as of September 2021. He might be mentioned because he is the incumbent president at the time the sentence is spoken.*

Llama2 [28] gives similar results [10]:

1. Di Maio: *This refers to Luigi Di Maio, an Italian politician who served as the Minister of Foreign Affairs and International Cooperation from 2018 to 2019.*
2. Colle: *This is a reference to the Quirinal Hill, a hill in Rome, Italy, where the President of Italy resides.*
3. Mattarella: *This refers to Sergio Mattarella, the President of Italy from 2015 to 2022.*

In both cases, entities are correctly identified and connected to relevant background knowledge, where their respective professional role are also highlighted. When it comes to implicatures, GPT4 is pretty inventive:

So, the sentence could mean that Luigi Di Maio is going to or present at the Quirinale, but he is not receiving instructions or direction directly from Sergio Mattarella. It could be used in a political or governmental context to express a situation where Di Maio is acting independently of the President of the Republic.

Llama2 seems to be less imaginative:

Therefore, the entities mentioned in the phrase are two politicians (Luigi Di Maio and Sergio Mattarella) and a geographic location (Quirinal Hill)

These examples show how, using LLMs appropriately, events can also be found in nominal constructions (such as the headline in question), and their participants, along with some other contextual element, can be reliably identified even with little superficial evidence. The LLMs generative ability of "connecting the dots" seem to be particularly effective when dealing with journalistic jargon, which is actually full of elliptical constructions. As for lexical units other than entities and events, framing complex notions such as *not receiving instructions* in a Knowledge Graph may raise ontological challenges, e.g. in this case that of representing negative facts. The "ontological cut-of" operated in the design phase, i.e. the way in which linguistic and logical (conceptual) expressiveness is arranged, plays here a crucial role. Our ontology is such that only basic patterns (e.g. *participation in action*) are ingested into the KG as logic assertions (i.e. triples), while blurry concepts (e.g. *receiving instructions*) are kept at the lexical level. Lexical concepts can be mapped to onto-lexical resources and interleaved by semantic relationships, as well as associated to distributional embeddings. In any case, the "ontological cut-of" requires the division of KG's reasoning into logical and linguistic inference procedures and the integration of their results, which is at the core of our future developments. The current prototype does not include semantic relationships and deep linguistic inference, but we do evaluate semantic similarity based on embeddings of textual fragments (e.g. headlines and summaries), e.g. when re-ranking KG queries results.

To improve knowledge extraction, we are in the process of experimenting LLMs generative models. It is already clear, however, that for giant models available only through remote services, such as those of the OpenAI family, the feasibility of these experiments could be problematic, since the stability of their behaviour seems to be questionable [29]. Also, the use of remote services would not comply with Il Manifesto's digital strategy, due to unwanted bindings to external business entities. Therefore, we are focusing on the use of on-premise open LLMs, trading some functionality for dependability, freedom, control, and cost effectiveness. At the time of writing, although the use of open models such as LLama2 seems promising, we have identified some hallucinations, for example the person "Matteo Meloni", erroneously identified as reference for "Meloni" in the context of "governo Meloni", who looks like a disturbing hybridization of the current Italian Prime Minister and his Deputy. How to deal with invented entities and fancy judgments is a general concern for the productive use of these new NLP methods.Our approach will be to involve editors, archivists and readers in reviewing and amending AI results.

[10] We are using the 13B parameters deployed on a virtual host

5. Conclusion

The construction of *MeMa*'s KG is an opportunity to discuss the state of the art perspective of NLP in the context of a real Italian content production environment. The KG will be made available later this year through a SPARQL endpoint and a dataset collection. At the current stage, our experience shows the potential, but also the limits, of NLP technologies applied to a large corpus of newspaper articles extended over a relevant time interval, which are characterized by a sophisticated use of the Italian language. In general, structured knowledge extraction can be achieved with various levels of granularity by integrating NLP processors, such as named entities recognizers, event recognizers and role labelers, keyword and topic extractors. Pre-trained multilingual LLM-based generative transformers will probably replace the supervised methods that have dominated the technology of these processors the last decade, considerably easing the task of extracting qualified semantic information. However, the new neural technologies do not seem free from errors, mainly due to the kind of inventive linguistic generation that may produce. Giving the user community the ability to "educate" AI, i.e. monitor and correct its results, remains the main route for us. Transparent logical structures such as Knowledge Graphs offer the best support for this type of activity. How information automatically extracted from text can be conceptualized and critically scrutinized by user communities will have a profound impact on the harmonization of AI in human ecosystems.

References

[1] S. Iaconesi, O. Persico, When my child is ai: Learning and experiencing through ai outside the school - the experiences of a community ai, QTimes Journal of Education Anno XIII (2021).

[2] S. Iaconesi, The illustrated principles of nuovo abitare, Medium, 2021. Https://medium.com/@salvatoreiaconesi/the-illustrated-principles-of-nuovo-abitare-f5c0af6d69a5.

[3] M. Lakshika, H. Caldera, Knowledge graphs representation for event-related e-news articles, Machine Learning and Knowledge Extraction 3 (2021) 802–818. URL: https://www.mdpi.com/2504-4990/3/4/40. doi:10.3390/make3040040.

[4] A. L. Opdahl, T. Al-Moslmi, D.-T. Dang-Nguyen, M. Gallofré Ocaña, B. Tessem, C. Veres, Semantic knowledge graphs for the news: A review, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3543508. doi:10.1145/3543508.

[5] A. Berven, O. A. Christensen, S. Moldeklev, A. L. Opdahl, K. J. Villanger, A knowledge-graph platform for newsrooms, Computers in Industry 123 (2020) 103321. URL: https://www.sciencedirect.com/science/article/pii/S0166361520305558. doi:https://doi.org/10.1016/j.compind.2020.103321.

[6] T. Bratanic, Making sense of news, the knowledge graph way, Neo4j Developer Blog, 2021. Published online on Feb 2, 2021.

[7] M. Yankova, Journalism in the age of open data, Ontotext Blog, 2016. Last accessed on July 2023.

[8] J. Z. Pan, G. Vetere, J. M. Gomez-Perez, H. Wu (Eds.), Exploiting Linked Data and Knowledge Graphs in Large Organisations, Springer, Cham, 2017. doi:10.1007/978-3-319-45654-6.

[9] P. Hitzler, M. K. Sarker, Neuro-symbolic artificial intelligence: The state of the art (2022).

[10] Hugging face, the ai community building the future, 2023. Https://huggingface.co/ (Accessed: 2023-07-20).

[11] N. Fernández, D. Fuentes, L. Sánchez, J. A. Fisteus, The news ontology: Design and applications, Expert Systems with Applications 37 (2010) 8694–8704. URL: https://www.sciencedirect.com/science/article/pii/S0957417410005592. doi:https://doi.org/10.1016/j.eswa.2010.06.055.

[12] C. Bekiari, G. Bruseker, M. Doerr, C.-E. Ore, S. Stead, A. Velios, Definition of the CIDOC Conceptual Reference Model v7.1.1, The CIDOC Conceptual Reference Model Special Interest Group, 2021. Release Date: June 9, 2021.

[13] C. S. Peirce, Collected Papers of Charles Sanders Peirce, Harvard University Press, Cambridge, MA, 1931-1958.

[14] J. Sowa, Ontology, metadata, and semiotics, Conceptual Structures: Logical, Linguistic, and Computational Issues (2000) 55–81.

[15] A. Gangemi, semiotics.owl: A content ontology pattern that encodes a basic semiotic theory, Linked Open Vocabularies, 2007. Https://lov.linkeddata.es/dataset/lov/vocabs/semiotics.

[16] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, E. Hovy, R. Belvin, R. MacIntyre, R. Grishman, et al., Ontonotes release 5.0 ldc2013t19, in: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), 2011, pp. 2525–2530. Https://catalog.ldc.upenn.edu/LDC2013T19.

[17] M. Honnibal, I. Montani, spacy 2: Industrial-strength natural language processing in python, 2020. URL: https://spacy.io.

[18] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics: System Demonstrations, Association for Computational Linguistics, 2020, pp. 272–277. Https://www.aclweb.org/anthology/2020.acl-demos.34.

[19] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia Spotlight: Shedding Light on the Web of Documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, pp. 101–108. URL: https://dbpedia.org/spotlight.

[20] S. Vajjala, R. Balasubramaniam, What do we really know about state of the art ner?, in: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), European Language Resources Association (ELRA), Marseille, 2022, pp. 5983–5993. Conference held on 20-25 June 2022.

[21] T. A. van Dijk, News as Discourse, Lawrence Erlbaum Associates, 1988.

[22] K. Markert, M. Nissim, Semeval-2007 task 08: Metonymy resolution at Semeval-2007, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, 2007, pp. 36–41.

[23] P. Jeretic, A. Warstadt, S. Bhooshan, A. Williams, Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8690–8705. doi:10.18653/v1/2020.acl-main.768, https://aclanthology.org/2020.acl-main.768.

[24] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti, P. S. Yu, A survey on deep learning event extraction: Approaches and applications, IEEE Transactions on Neural Networks and Learning Systems 14 (2022) November 2022. doi:10.1109/TNNLS.2022.xxxxxxx.

[25] K. A. Mathews, M. Strube, A large harvested corpus of location metonymy, in: International Conference on Language Resources and Evaluation, 2020.

[26] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, 2023. arXiv:2304.10428.

[27] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[28] H. Touvron, al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[29] L. Chen, M. Zaharia, J. Zou, How is chatgpt's behavior changing over time?, 2023. arXiv:2307.09009.