This the accepted version of a book chapter. Please cite it as follows:
Zavarella, V., Consoli, S., Recupero, D.R., Fenu, G. (2025). Exploring Digital Health Trends in the Headlines via Knowledge Graph Analysis. In: Nicosia, G., Ojha, V., Giesselbach, S., Pardalos, M.P., Umeton, R. (eds) Machine Learning, Optimization, and Data Science. LOD ACAIN 2024 2024. Lecture Notes in Computer Science, vol 15508. Springer, Cham. https://doi.org/10.1007/978-3-031-82481-4_9

# Exploring Digital Health Trends in the Headlines via Knowledge Graph Analysis

Vanni Zavarella[1 [0000−0002−4498−2769]], Sergio Consoli[2 [0000−0001−7357−5858]], Diego Reforgiato Recupero[1 [0000−0001−8646−6183]], and Gianni Fenu[1 [0000−0003−4668−2476]]

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09121 Cagliari, Italy v.zavarella@studenti.unica.it {diego.reforgiato,fenu}@unica.it
[2] European Commission, Joint Research Centre (DG JRC),Via E. Fermi 2749, 21027 Ispra, Italy sergio.consoli@ec.europa.eu

**Abstract.** We introduce a method for analyzing digital transformation in the health domain by constructing a Knowledge Graph from a large corpus of 7.8 million English news articles from the Dow Jones Data, News, and Analytics platform, dating from 1987 through 2023. We first sampled around 97k articles relevant to the Digital Health topic by training and deploying a Deep Learning binary classifier by fine-tuning BERT. Successively, by deploying Natural Language Processing techniques, we extracted triples from the identified articles to form a Digital Health News Knowledge Graph, which consists of 431k distinct triples connecting 186k entities through 1866 relations. This graph provides insights into the evolution of Digital Health in news media and serves as a resource for further research in the field. Our analysis reveals significant trends in Digital Health as reflected in the news, with notable peaks coinciding with key events like the COVID-19 pandemic. We split the analysis geographically for the United States and European countries and tracked over time for each macro-region the predominant entities and relations. The classifier, the knowledge graph, and data analytics visualizations are made publicly available for future work.

**Keywords:** Information Extraction · Knowledge Graphs · Digital Health · Transformers · News Analysis · Named Entity Recognition.

## 1 Introduction

In recent years, the landscape of healthcare has undergone a profound transformation fueled by rapid advancements in digital technology. The integration of digital tools and innovative solutions into traditional healthcare systems has ushered in a new era characterized by unprecedented opportunities and challenges. This paradigm shift, often referred to as the digital transformation of healthcare, encompasses a wide array of technologies ranging from electronic health records and telemedicine to artificial intelligence (AI) and wearable devices.

News documents serve as a real-time reflection of societal attitudes, technological innovations, and policy developments, offering a unique lens through which to explore the multifaceted dimensions of digital transformation in health. By analyzing news articles, reports, and commentaries, we can glean insights into emerging trends, public perceptions, and the evolving discourse surrounding digital health initiatives.

In the present day, there exists a plethora of news monitoring tools offering diverse analytical capabilities for news analysis. Examples include Europe Media Monitor (EMM), Brandwatch, Brand24, Repustate, Cision Communication Cloud, SentiOne, and Meltwater[3]. However, these existing systems exhibit limitations in adequately capturing the nuanced dynamics of discourse and, while they can identify specific tags, keywords, or entities mentioned in news articles, generally they cannot extract the connections between them and support advanced queries about those. EMM represents a remarkable exception in this respect, as it extracts and maintains rich metadata aggregated over a few entity types [21]; however, only a limited number of target domains are covered by the system [2]. To address this constraint, researchers have proposed diverse methodologies to construct structured, interlinked, and machine-readable data frameworks tailored for news analysis [18,23]. A number of these frameworks utilize semantic technologies, including knowledge graphs (KGs).

KGs represent expansive networks comprising entities and relationships, providing machine-readable and interpretable information within a specific domain, structured according to formal semantics [10]. In recent years, KGs have gained increasing recognition for their capacity to organize structured data in a semantically meaningful manner, enabling effective support for various AI systems [19]. The relationship between two entities is typically expressed as a triple in the form of `<subject, predicate, object>`, for example: `<digital transformation, revolutionize, industry>`. KGs' structure is commonly defined by a domain ontology. Large-scale KGs are typically generated through a semi-automated process that incorporates both structured and unstructured data. Notable examples include DBpedia [12][4], Google Knowledge Graph[5], BabelNet[6], and YAGO[7]. Furthermore, KGs can undergo automatic refinement via link prediction techniques, aimed at identifying additional relationships among domain entities [11,17]. These approaches can, for instance, facilitate the generation of novel scientific hypotheses by establishing connections among known entities in innovative ways [5].

The creation of extensive and high-quality KGs from news data represents a contemporary open challenge that has garnered attention from only a hand-

[3] https://emm.newsbrief.eu, https://www.brandwatch.com/, https://brand24.com/, https://www.repustate.com/, https://www.cision.com/, https://sentione.com/, https://www.meltwater.com/

[4] https://www.dbpedia.org/

[5] https://developers.google.com/knowledge-graph

[6] https://babelnet.org/

[7] https://yago-knowledge.org/

ful of researchers [9]. Current solutions primarily rely on information extraction pipelines [3,9,1,15]. While information extraction techniques offer scalability potential, they often struggle to produce outputs of adequate quality for practical applications. Specifically, existing approaches for extracting entities and relationships from news articles tend to concentrate on specific domains [9], neglecting the significance of preprocessing, entity linking operations, and relational grounding. Consequently, applying existing methods for entity and relation extraction across extensive text collections would likely yield a notably noisy outcome [8]. Hence, several challenges must be addressed, including: a) synthesizing extracted information from diverse sources into a unified representation; b) assessing the accuracy of resulting triples; c) establishing a flexible ontological framework to formalize a broad spectrum of statements derived from news content.

Similar challenges have previously been tackled within the scholarly domain by [8], wherein the authors introduced an information extraction methodology that integrated data from diverse tools based on a domain ontology, thereby facilitating the creation of expansive KGs. This innovative approach has served as a catalyst for subsequent research endeavors in the field [24,14,25,22,6]. However, this initial effort also encountered several limitations, including i) a restricted capacity to consolidate multiple instances of the same entity; ii) a superficial and manual strategy for mapping verbal predicates to semantic relations; iii) a limited methodology for assessing triple validity, reliant on a basic multi-layer perceptron classifier.

Therefore, in this paper, we present an enhanced information extraction architecture designed to extract and merge instances of open-domain entities from news articles and to identify and generalize various relationships among these entities by using hierarchical clustering, word embeddings, and dimensionality reduction techniques. The designed architecture is scalable and incorporates a module for unifying and grounding entity instances using an external resource, namely DBpedia. We apply the pipeline to a topic-specific news dataset reporting updates on different aspects of the Digital Health domain and generate a large KG of predictive triples concerning key entities in the domain. We present several data visualizations and show how these can potentially support insight gathering about trends and dynamics of digital transformation in the health sector over time and in different geographical regions.

The remainder of this paper is structured as follows. Section 2 presents the dataset we have used to extract triples. Section 3 discusses the classifier we have designed to identify health documents from the considered dataset. The pipeline of our approach is depicted in Section 4 whereas the KG we have created is illustrated in Section 5. Section 6 presents aggregated analysis and visualizations on the generated KG. Finally, Section 7 ends the paper with conclusions and future directions.

## 2 Dataset

The dataset under consideration encompasses around 7.8 million English-language news articles gathered from the Dow Jones Data, News, and Analytics (DNA) platform[8]. These articles cover a time frame spanning from September 1987 to December 2023 and originate from diverse global English-language outlets, such as The Wall Street Journal, the New York Times, and The Guardian.

DNA provides a range of general metadata for each article, including the publication date, the publisher, the title, and the full text. Furthermore, DNA provides a range of curated content-based descriptors that are useful for filtering data along specific dimensions. These descriptors include an 8-level taxonomy comprising approximately nine hundred Subject codes; a 7-level taxonomy encompassing nearly a thousand Industries codes, and a set of Region codes encompassing all countries and regions mentioned in the news items.

We started by discarding items with missing titles and short articles with text body character lengths lower than 300. Moreover, we filtered and merged Region codes to end up with a two-valued (Europe/US) attribute. We then tested for various combinations of these metadata tags as a means to sample news articles about Digital Health technologies. However, there were not enough documents in the domain by using such metadata tags only. On the one hand, health-related Subject tags fall short of retrieving financial/market news updates involving health tech key players; on the other, DNA's Industries classification schema is too coarse-grained to capture emerging technologies and products in this domain. Therefore, to identify a more representative sample of digital health articles, we opted for using a trained Deep Learning binary classifier, as described in the following section.

## 3 Classifier

We fine-tuned the BERT (Bidirectional Encoder Representations from Transformers)[9] language model using a near-balanced small set of 9097 news items sampled from DNA and several RSS feeds from specialized news outlets in health tech[10]. We will refer to negative instances as non-digital health-related documents, while positive instances will denote digital health-related documents.

Out of the 4602 negative instances, 3000 were DNA items and 1602 were articles scraped from 'negative' topic feeds of technology news outlets[11]. The DNA items were collected by concatenating title and full text of a sample of 500

[8] https://professional.dowjones.com/developer-platform/

[9] https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-2_H-128_A-2/1

[10] For example, https://www.healthtechdigital.com/, https://techcrunch.com/tag/healthtech/feed/ https://www.digitalhealth.net/news/.

[11] For example, https://techcrunch.com/tag/security/.

items for each set of 'negative' topic codes[12]. Analogously, we concatenated the title and full text for the 1602 tech news items.

Out of the 4495 positive instances, 4187 consisted of concatenated titles and full texts of articles from the health tech news outlets mentioned above. To select positive instances from DNA, we filtered for health-related Subject codes and manually verified a subset of 308 health tech items. The data text underwent pre-processing, which involved the removal of URLs, all-numeric tokens, and DNA and news outlet-specific tokens (e.g., 'Reuters', 'Reuters Limited', 'techcrunch'). Additionally, all texts were truncated to 1000 characters to eliminate any correlation between the topic and text length features of the article sources.

We then performed fine-tuning using 10-fold stratified cross-validation with 80-20% data splits and Binary Cross Entropy as Loss function, training for 10 epochs with Early Stopping on 1 epoch of non-increasing Accuracy score. To mitigate over-fitting on the relatively small training set we kept the model size small (4.3M trainable parameters) and added a dropout regularization layer in the training phase (0.2 dropout rate).

Table 1 displays the average cross-validation performance of the classifier on the provided dataset. Additionally, we collected a separate dataset comprising 100 negative instances sourced from DNA and 100 positive instances sourced from DNA using Subject codes. Subsequently, we trained the classifier on the entire aforementioned training dataset, and evaluated its performance on the new unseen test set of 200 documents. The classifier's performances on the test set are shown in the last row of Table 1.

**Table 1.** Evaluation of the binary Health Tech topic classifier.

| | **Precision** | **Recall** | **F-Score** | **Accuracy** |
|---|---|---|---|---|
| Cross-validation | 99.40 | 97.40 | 98.39 | 98.59 |
| 200 documents Test Set | 98.9 | 93.0 | 95.8 | 96.0 |

The results show that, although the model misses a higher number of positive instances, it is overall able to sample a consistent subset of relevant health tech articles from the DNA multi-domain corpus. Therefore, the satisfactory performance of the classifier allowed its deployment on the entire DNA dataset to achieve an overall of 97k health tech articles (i.e. the 1.2% from the entire DNA input dataset) for our further analysis.

The model, after training on the entire train set, has been made publicly available at the project repository[13].

[12] namely *gcat* (Political/General News), *mcat* (Commodity/Financial Market News), *ccat*(Corporate/Industrial News), *ecat*(Economic News), *gent*(Arts/Entertainment), *gcrim*(Crime/Legal Action)

[13] https://github.com/zavavan/dtm_kg/tree/master/data-collection/dna/bert_fine_tuned_healthTech

## 4 Pipeline

Our information extraction architecture for generating semantic triples consists of a customized spaCy NLP pipeline[14] coupled with a series of novel Entity and Relation processing modules, described in the following.

### 4.1 Data Preprocessing

Since the spaCy models that we apply for triple extraction[15] perform with high accuracy on benchmark corpora[16] comparable to our source news data, we were able to perform at this step only a minimal text normalization process, by simply removing URLs and a list of other news platform-specific token patterns (see Section 3) that typically disrupt the syntactic parsing of the sentence.

### 4.2 Triple Extraction

In the block responsible for triple extraction, news items that have been preprocessed undergo sentence segmentation, with each sentence being further processed through the spaCy pipeline, as indicated in the following. Drawing upon prior approaches [7], we establish a series of procedures for identifying candidate nominal entities and linking them through predicative relations, based on dependency parse trees.

**Entity extraction module:** This module identifies local nominal phrases with a defined range of syntactic variations (e.g., compound nouns and adjectives). These phrases are subsequently extended and connected by:

- Non-recursive attached prepositional phrases;
- Quantity-type entities (MONEY, PERCENT, QUANTITY, CARDINAL);
- Entity spans linked via pronominal anaphoras, resolved using the Spacy pipeline's coreference component[17].

Ultimately, the module yields a set $E = \{e_0, ..., e_n\}$ of non-unified, candidate entity phrases.

**Relation extraction module:** In this stage, for each sentence $s_i$, all the shortest paths of the dependency tree between every pair of entities $(e_m, e_n)$ containing a verb and matching any of a shortlist of expert-validated patterns are selected. The target pattern set has been determined through expert validation conducted over a sample of the most frequently matched patterns in an external,

[14] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.6.0.
[15] Including sentence splitting, Part-of-Speech (POS) tagging, Dependency Parsing
[16] https://catalog.ldc.upenn.edu/LDC2013T19
[17] https://github.com/richardpaulhudson/coreferee

open-domain text corpus[18]. This updated process generates a set of verbal relations $V = v_0, ..., v_k$ and a set of triples $S = s_0, ..., s_k$ in the form $< e_m, v, e_n >$ where $v \in V$ and $e \in E$.

The final objective of the pipeline is to enable the generalization from the surface form triples in set $S$ to a smaller set $T = t_0, ..., t_h$ of triples in the form of $< \epsilon_m, r, \epsilon_n >$, where each $\epsilon_i \in E$ represents a unified entity and $r$ is a label drawn from a generalized relation vocabulary $R$.

### 4.3 Entity Refining

Initially, entities undergo a cleanup process where leading/trailing punctuation marks and stopwords are removed. Subsequently, all entity tokens whose POS tag is neither Verb nor Proper Noun are lemmatized and lower-cased.

To merge normalized candidate entities, we leverage their linking to DBpedia entries by applying the DBpedia Spotlight module. To achieve this, we run the module over modified article sentences where the original subjects and object entity spans are replaced with their normalized forms. Entities that are linked to the same DBpedia entries are merged. For instance, the entities *'Gartner'* and *'@Gartner_inc'* are merged as they are linked to the DBpedia entry of the Gartner consulting firm: http://dbpedia.org/resource/Gartner. This merging process is formalized with a relation `owl:sameAs` in the output KG. If only the first condition is satisfied, a semantic 'relatedness' link (formalised with a `skos:related` relation) is assigned between the candidate entity and the DBpedia entry, indicating that the former is 'related to' the latter, but not an instance of it. For example, the span *'@gartner_survey'* is considered only `skos:related` to the DBpedia entry for Gartner.

### 4.4 Relation Refining

To identify the most suitable predicate label $r$ for each relation verb $v$ in a triple $< e_m, v, e_n >$ and to establish the mapping from $v$ to $r$ in the resulting triple, we follow these steps. First, we generate a word embedding representation of the verb predicates using a pre-trained model. Next, we perform optimized clustering of the relation vectors. Finally, we employ a representative instance from each cluster to map the verb predicates.

*Relation Embeddings:* For each single or multi-token relation predicate, we used the static, 300-dimensional word embeddings learned with GloVe [20] and made available for text Span objects in the Spacy *en_core_web_lg-3.6.0* pipeline[19].

[18] The resource can be found at the URL https://github.com/zavavan/dtm_kg/tree/master/data-collection/dna/bert_fine_tuned_healthTech.

[19] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.6.0. We tested using various contextual embedding vectors however it turned out that these representations were not suitable for generalizing enough over relations, probably due to the context-specific information they are encoding.

*Dimensionality Reduction and Clustering:* We employed the HDBSCAN clustering algorithm, complemented by prior application of the UMAP dimension reduction technique on the word embeddings vectors[20]. HDBSCAN, a hierarchical variant of the density-based DBSCAN algorithm, is characterized by labelling as outliers and leaving unclustered data points situated in low-density regions [13]. As a result, high-dimensional data need a greater number of observed samples to achieve the appropriate density level for HDBSCAN to function effectively. Nonetheless, leveraging UMAP for non-linear, manifold-aware dimension reduction [16] has proved effective in reducing datasets to the proper number of dimensions for HDBSCAN to cluster the vast majority of instances.

To optimize the UMAP and HDBSCAN combination, we conducted a grid search across the hyperparameters of both algorithms, evaluating clustering using the score:

$$S = silhouette_X \cdot clustered_X. \tag{1}$$

Here, $silhouette_X$ represents the mean silhouette coefficient over all clustered instances of dataset $X$ by HDBSCAN [4], and $clustered_X$ denotes the fraction of instances of $X$ clustered. In practice, we optimized for cluster cohesion and separation, while penalizing configurations with low dataset coverage. Subsequently, we selected a subset of top-scoring hyper-parameter configurations and plotted their $S$ score against the number of output clusters. This enables us to select a sub-optimal configuration that strikes a balance between generalization (fewer clusters) and accuracy (cluster number closer to the dataset size). In our experiments, the chosen configuration reached a $S$ score of 0.62, with $silhouette_X = 0.65$, $clustered_X = 0.92$ and 2 UMAP components.

*Relation Mapping:* Finally, for each relation verb $v$ in the dataset, we replaced it with a predicate label $r$ consisting of the most frequent lemma among the 'exemplars' relations returned by HDBSCAN for the cluster of $v$[21]. Otherwise, we map it to itself if $v$ is an outlier.

## 5 Knowledge Graph Analysis

The described pipeline has been deployed to generate a Digital Health News Knowledge Graph (referred to as DHNEWS_KG), comprising roughly 431k distinct (non-reified) triples, connecting 186k unique entities via a total of 1866 generalized relations. In the corresponding ontology, designed to describe DHNEWS_KG (*dhnewskg-ont* namespace prefix), each extracted claim is successively reified into instances of the `dhnewskg-ont:Statement` class, with `dhnewskg-ont:Statement` representing a specific assertion derived from a collection of news items.

[20] https://umap-learn.readthedocs.io/en/latest/parameters.html

[21] https://hdbscan.readthedocs.io/en/latest/api.html. Notice that as HDBSCAN can generate clusters of arbitrary forms, it does not hold a notion of cluster centroid and there are typically multiple 'most representative' data points in a cluster.

**Table 2.** Sample statements extracted by the pipeline, with their support.

| Subject Entity | Relation | Object Entity | Support |
|---|---|---|---|
| Italy | report | coronavirus death | 374 |
| clinical trial | involve | patient | 128 |
| interactive graphic | track | global spread | 90 |
| fitch | undertake | sensitivity analysis | 75 |
| administration | approve | drug | 47 |
| meningitis immunity | fight | endometrial cancer | 35 |
| dow chemical | develop | drug | 28 |
| drug tamoxifen | reduce | risk | 17 |
| fibrocell | announce | fda acceptance | 2 |
| zealand pharma | announce | fda acceptance | 1 |

Figure 1 provides an example of a claim reification, with the ontology instance `dhnewskg:drug_tamoxifen` serving as `rdf:subject` and the `dhnewskg-ont:hasSupport` data property reporting the number of news articles supporting the claim.

```
dhnewskg-ont:statement_90694 a rdf:Statement ;
dhnewskg-ont:comesfromNewsArticle dhnewskg:lba0000020030305dz34000cl ;
dhnewskg-ont:hasSupport 1 ;
dhnewskg-ont:negation false ;
rdf:object dhnewskg:receptor ;
rdf:predicate dhnewskg-ont:block ;
rdf:subject dhnewskg:drug_tamoxifen .
```

**Fig. 1.** A sample reification for a statement concerning the ontology instances `dhnewskg:drug_tamoxifen` and `dhnewskg:receptor`.

A sample of generated (un-reified) statements is illustrated in Table 2, together with their supports. The support distribution of the triples has a marked long-tail pattern, with a few key statements occurring frequently and a vast majority matched only a few times.

DHNEWS_KG inherits DBpedia entity typization of `owl:sameAs` linked entities. Table 3 lists the 20 predominant DBpedia-inherited types within the graph. All not-linked entities are classified into the generic `dhnewskg:Entity` type. Out of the overall set of unique DHNEWS_KG entities, around 8% have been linked to DBpedia entries using 14975 `owl:sameAs` and 33345 `skos:related` predicates, indicating entity equality and relatedness, respectively. Overall, 23.8% of all triples had either subject or object entities linked to DBpedia.

As the deployed methodology has been manually evaluated and proved to have reliable Precision[26], we made publicly available the automatically gener-

**Table 3.** Number of matches and unique matches of the 20 most represented DBpedia entity types in DHNEWS_KG.

| DBpedia Entity Type | #Matched Entities | #Unique Entities |
|---|---|---|
| DBpedia:Organisation | 20050 | 2640 |
| DBpedia:Company | 15667 | 1736 |
| DBpedia:Country | 12124 | 324 |
| DBpedia:Disease | 7881 | 918 |
| DBpedia:Person | 6594 | 2611 |
| DBpedia:ChemicalSubstance | 6583 | 1378 |
| DBpedia:Drug | 5927 | 1180 |
| DBpedia:Politician | 4069 | 1187 |
| DBpedia:Work | 1872 | 567 |
| DBpedia:MonoclonalAntibody | 1258 | 128 |
| DBpedia:GovernmentAgency | 1123 | 107 |
| DBpedia:AdministrativeRegion | 1088 | 185 |
| DBpedia:City | 971 | 205 |
| DBpedia:Bank | 789 | 128 |
| DBpedia:Biomolecule | 754 | 165 |
| DBpedia:Group | 729 | 109 |
| DBpedia:AnatomicalStructure | 656 | 151 |
| DBpedia:ChemicalCompound | 631 | 191 |
| DBpedia:ArchitecturalStructure | 593 | 183 |
| DBpedia:Gene | 586 | 124 |
| dhnewskg-ont:Entity | 800527 | 185653 |

ated DHNEWS_KG graph via data access endpoints. We have set up a Virtuoso SPARQL endpoint for this purpose where DHNEWS_KG can be queried, and analytical information on target entities, attributes, and relations can be retrieved in user-specified data formats[22]. As an example, a SPARQL query like the one in Figure 2 can be run to return all the 480 statements from the 'DHNEWS_KG' graph having the target entity `dhnewskg:biogen` as subject, where `dhnewskg:biogen` is a graph resource `owl:sameAs`-linked to the DBpedia entry for the American multinational biotechnology company Biogen Inc.

```
PREFIX dhnewskg: <http://dhnewskg.org/dhnewskg/resource/>
PREFIX dhnewskg-ont: <http://dhnewskg.org/dhnewskg/ontology#>
SELECT ?statement
FROM <DHNEWS_KG>
WHERE { ?statement a rdf:Statement .
 ?statement rdf:subject dhnewskg:biogen . }
```

**Fig. 2.** Query returning all DHNEWS_KG statements with the graph entity `dhnewskg:biogen` as `rdf:subject`.

[22] https://api-vast.jrc.service.ec.europa.eu/sparql/. Currently, the access is password protected, with credentials available upon request to authors. Provisional credentials for the reviewing process: (dtsmm_user, dtsmm_user_2024).

For each result statement, a link to its corresponding URL in a Virtuoso Faceted Browser endpoint[23] is returned, allowing further navigation. The Faceted Browser allows the exploration of the KG by querying the endpoint with free text search patterns. This returns a list of literal property values or labels that match the text pattern specified as input. Then, it is possible to narrow down the result set by filtering for specific entity types or property values, allowing for navigation of the target sub-graph structure.

## 6 Data Analytics Dashboards

We offer aggregated analyses of the evolving landscape of Digital Health dynamics over the years through a collection of interactive visualization dashboards accessible on the page http://192.167.149.18:5006/dashboard.

In the Health Tech News Ratio panel (also illustrated in Figure 3), we present the month-sampled time series depicting the proportion of 97k news articles concerning Digital Health, as identified by the classifier mentioned in Section 3, out of the total number of English language DNA news articles pertaining to Europe and the US, respectively. Although it underestimates the absolute count (attributable to the classifier's lower recall), the plot highlights significant trends in the emergence and establishment of digital health technologies, revealing notable distinctions between European and US contexts.

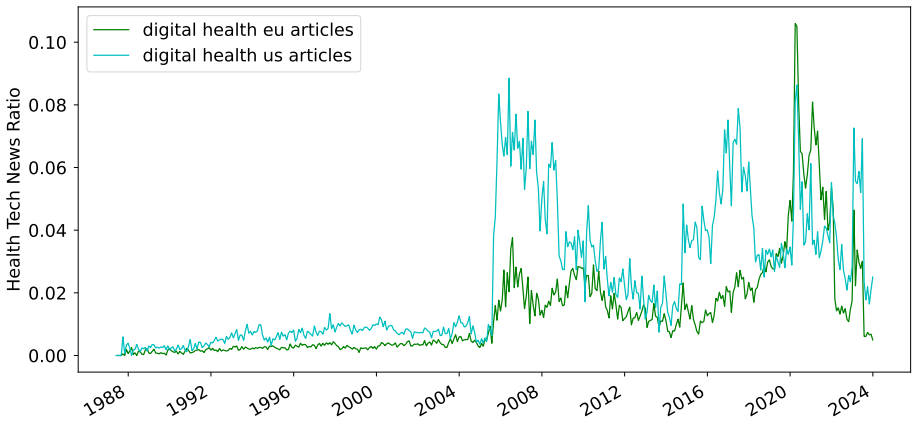


**Fig. 3.** Monthly time series of the ratio of news articles about Digital Health over the total of DNA news about Europe and the US, respectively.

The coverage of the digital health topic remained marginal in the news, accounting for approximately 1%, until the first significant breakthrough was

[23] https://virtuoso.diglife.eu/fct/

recorded in 2005 in both regions. This surge in exposure likely stemmed from the implementation and gradual adoption of Health Information Exchange (HIE) networks. A second peak occurred in 2017 in the US, while European news exhibited relatively steady growth from 2015 onward. A third notable spike occurred globally in 2020, likely attributable to the widespread efforts to leverage innovative technologies in combating COVID-19.

The Top Entity Types bar plots in the dashboard show the predominant DBpedia-inherited entity types within the graph for triples tagged with Europe, US, and EU-US region codes via their article support.

For a subset of predominant types, the Top Key Entities plots track the occurrence of several key entities per year, where occurrence means the entity is either the Subject or Object of an extracted triple in the KG. One can point out here how some major pharmaceutical industry corporations seem to have an impact on a global scale (both for Europe and the US), while major information technology giants show different impacts in the Digital Health industry in the two contexts.

Lastly, within the Entity Chord Diagrams panel (with a sample for US reproduced in Figure 4), we present the most frequently connected entity pairs within the KG through chord illustrations, serving as both Subjects and Objects of predicative triples. The size of the chords corresponds to the support of the depicted relation[24].

## 7 Conclusions

We described the development of a Digital Health News Knowledge Graph (DHNEWS_KG) using a dataset of 7.8 million English news articles from 1987 to 2023. The pipeline involved training a deep learning classifier to identify relevant articles and employing NLP techniques to extract and generalize knowledge triples.

The resulting, large-scale DHNEWS_KG offers insights into digital health trends. The analysis we provided allows for comparing digital health news coverage in the US and Europe over a large period, tracking the relative impact of key players. While the method has been proven to generate accurate triples, we plan to further investigate its recall performance, particularly compared to recent zero-shot learning approaches supported by Large Language Models.

DHNEWS_KG, along with its visualization and analytics tools, have been made publicly available for stimulating further research. We are currently working on expanding the supporting tools and services to further help exploiting DHNEWS_KG and extracting useful hidden insights, focusing mainly on detection of trending technologies.

[24] For the sake of visualization, we pre-filtered for relations with a minimum number of 20 occurrences in the dataset.

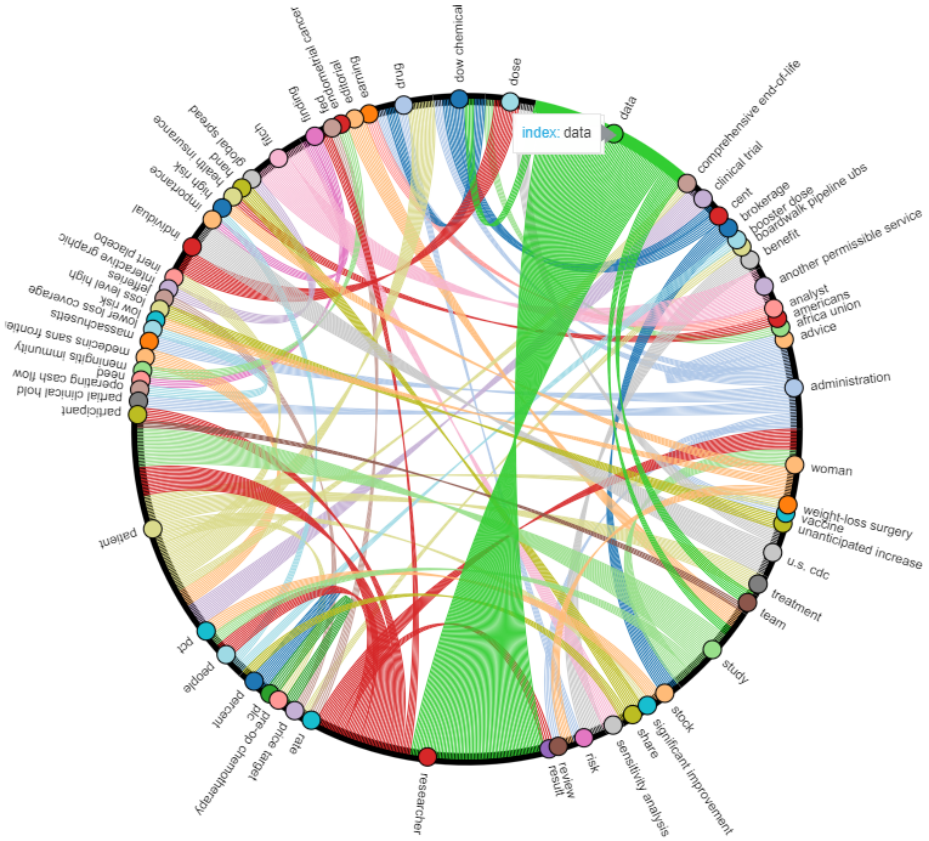


**Fig. 4.** Chord diagram illustrating the strength (statement support) of relationships among significant entities within the dataset.

## Acknowledgements

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

## References

1. Alani, H., Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic Web Machine Reading with FRED. Semantic Web **8**(6), 873–893 (2017)
2. Atkinson, M., Piskorski, J., Tanev, H., Zavarella, V.: On the creation of a security-related event corpus. In: Caselli, T., Miller, B., van Erp, M., Vossen, P., Palmer, M., Hovy, E., Mitamura, T., Caswell, D. (eds.) Proceedings of the Events and Stories in the News Workshop. pp. 59–65. Association for Computational Linguistics, Vancouver, Canada (Aug 2017)
3. Barbosa, C., Félix, L., Vieira, V., Xavier, C.: Sara - A Semi-Automatic Framework for Social Network Analysis. In: Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web. pp. 59–62. SBC, Porto Alegre, RS, Brasil (2019)
4. Batool, F., Hennig, C.: Clustering with the Average Silhouette Width. Computational Statistics and Data Analysis **158** (2021)
5. Borrego, A., Dessi, D., Hernández, I., Osborne, F., Recupero, D.R., Ruiz, D., Buscaldi, D., Motta, E.: Completing scientific facts in knowledge graphs of research concepts. IEEE Access **10**, 125867–125880 (2022)
6. Chessa, A., Fenu, G., Motta, E., Osborne, F., Reforgiato Recupero, D., Salatino, A., Secchi, L.: Data-driven methodology for knowledge graph generation within the tourism domain. IEEE Access **11**, 67567–67599 (2023)
7. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E.: CS-KG: A large-scale knowledge graph of research entities and claims in computer science. In: Sattler, U., Hogan, A., Keet, C.M., Presutti, V., Almeida, J.P.A., Takeda, H., Monnin, P., Pirrò, G., d'Amato, C. (eds.) The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13489, pp. 678–696. Springer (2022)
8. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E.: Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Generation Computer Systems **116**, 253–264 (2021)
9. Dörpinghaus, J., Klante, S., Christian, M., Meigen, C., Düing, C.: From social networks to knowledge graphs: A plea for interdisciplinary approaches. Social Sciences & Humanities Open **6**(1), 100337 (2022)
10. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: Martin, M., Cuquet, M., Folmer, E. (eds.) Joint Proceedings of the Posters and Demos Track

of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016. CEUR Workshop Proceedings, vol. 1695, pp. 1–4. CEUR-WS.org (2016)
11. Kumar, A., Singh, S.S., Singh, K., Biswas, B.: Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications **553**, 124289 (2020)
12. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**, 167–195 (2015)
13. Malzer, C., Baum, M.: A hybrid approach to hierarchical density-based cluster selection. In: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems. vol. 2020-September, p. 223 – 228 (2020)
14. Man, T., Vodyaho, A., Ignatov, D., Kulikov, I., Zhukova, N.: Synthesis of multilevel knowledge graphs: Methods and technologies for dynamic networks. Engineering Applications of Artificial Intelligence **123** (2023)
15. Martinez-Rodriguez, J.L., Lopez-Arevalo, I., Rios-Alvarado, A.B.: OpenIE-based approach for Knowledge Graph construction from text. Expert Systems with Applications **113**, 339–355 (2018)
16. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction (2020)
17. Nayyeri, M., Cil, G.M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., Salatino, A., Recupero, D.R., Vassilyeva, N., Motta, E., et al.: Trans4e: Link prediction on scholarly knowledge graphs. Neurocomputing **461**, 530–542 (2021)
18. Opdahl, A.L., Al-Moslmi, T., Dang-Nguyen, D.T., Gallofré Ocaña, M., Tessem, B., Veres, C.: Semantic knowledge graphs for the news: A review. ACM Comput. Surv. **55**(7) (dec 2022)
19. Peng, C., Xia, F., Naseriparsa, M., Osborne, F.: Knowledge graphs: opportunities and challenges. Artificial Intelligence Review pp. 1–32 (2023)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014)
21. Steinberger, R., Pouliquen, B., van der Goot, E.: An introduction to the europe media monitor family of applications (2013)
22. Tamašauskaite, G., Groth, P.: Defining a knowledge graph development process through a systematic review. ACM Transactions on Software Engineering and Methodology **32**(1) (2023)
23. Tan, F.A., Paul, D., Yamaura, S., Koji, M., Ng, S.K.: Constructing and interpreting causal knowledge graphs from news (2023)
24. Xiao, Y., Li, C., Thürer, M.: A patent recommendation method based on kg representation learning. Engineering Applications of Artificial Intelligence **126** (2023)
25. Yu, S., Peng, C., Xu, C., Zhang, C., Xia, F.: Web of conferences: A conference knowledge graph. In: WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining. pp. 1172–1175 (2023)
26. Zavarella, V., Consoli, S., Reforgiato Recupero, D., Fenu, G., Angioni, S., Buscaldi, D., Dessí, D., Osborne, F.: Triplétoile: Extraction of knowledge from microblogging text. Heliyon **10**(12), e32479 (2024). https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e32479