CimpleKG: A Continuously Updated Knowledge Graph on Misinformation, Factors and Fact-Checks

Grégoire Burel, Martino Mensio, Youri Peskine, Raphael Troncy, Paolo Papotti, Harith Alani

CimpleKG: A Continuously Updated Knowledge Graph on Misinformation, Factors and Fact-Checks

Grégoire Burel[0000−0003−0029−5219][1], Martino Mensio[0000−0002−9875−6396][1], Youri Peskine[0009−0002−8160−019X][2], Raphael Troncy[0000−0003−0457−1436][2], Paolo Papotti[0000−0003−0651−4128][2], and Harith Alani[0000−0003−2784−349X][1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
{gregoire.burel,martino.mensio,harith.alani}@open.ac.uk
[2] EURECOM, Sophia Antipolis, France
{youri.peskine,raphael.troncy,paolo.papotti}@eurecom.fr

Abstract. Misinformation has a pervasive thread running through society, causing confusion, mistrust, and uncertainty. The detection, tracking, and countering of misinformation is a very active research area with an intense need for data about circulating claims and their attributes, fact-checks, and verification outcomes. Although various relevant datasets exist, they tend to be of limited scope in terms of time coverage, topics, country, language, and quantity. In this paper, we introduce CimpleKG as an open and continuously updated semantic resource. CimpleKG links daily updated data from 77 fact-checking organisations with over 217k documents from static misinformation datasets. The knowledge graph is also augmented with relevant textual features and entities extracted from the textual data integrated into the graph. At the time of writing, the knowledge graph contains more than 15m triples, including 263k+ distinct entities and 1m textual features with over 203k fact-checked claims, spanning 26 languages and 36 countries. CimpleKG is publicly available and has been used in various research studies and web applications.

Resource Type: Knowledge Graph.

SPARQL Endpoint: https://purl.org/net/cimplekg/sparql.
KG Releases: https://purl.org/net/cimplekg/knowledge-graph.
KG Explorer: https://purl.org/net/cimplekg/explorer.

Keywords: ClaimReview · Misinformation · Factcheck · Knowledge Graph.

1 Introduction

In response to the growing challenge of misinformation, hundreds of fact-checking organisations around the world have sprung up to investigate various claims and

debunk misinformation. Fact-checking organisations (or fact-checkers) are independent organisations that identify, contextualise, verify and rate the accuracy of public information and claims for fighting misinformation and supporting informed public discourse.

Understanding the complex dynamics between misinformation, fact-checking, and the broader information ecosystem is crucial for developing robust strategies to combat the harmful effects of misinformation in society. As a result, much misinformation research focuses on tracking its spread [28], identifying its sources [11], and building automated detection methods [25]. There's also a growing line of research that focuses on misinforming claims and their corresponding fact-checks, to better understand and track their interplay [2,3], and examine the impact of debunking efforts [17].

These lines of research require the availability of up-to-date data that contains the details of claims, their corresponding fact-checks and ratings as well as the entities involved in the claim verification processes. In this paper, we introduce a continuously updated public knowledge graph (KG) called CimpleKG[3] that can be used for supporting misinformation research. CimpleKG links various previously published static misinformation datasets with daily updated claims verification from vetted fact-checking organisations and augments them with additional information such as named entities and contextual factors (e.g., emotions, sentiment, political leanings, conspiracy theories, propaganda techniques).

Although our KG is not the first attempt at gathering and representing fact-checking data [27], CimpleKG is much larger than previous works in terms of time coverage, topics, country, language, quantity and freshness. It is also novel as it includes so-called factors extracted from the text to explain misinformation; it normalises the rating schemes used by fact-checking organisations and also resolves shortened URLs to their unshortened version. This is useful as fact-checkers tend to use archiving URL services when referring to misinforming URLs. Finally, contrary to previous work, CimpleKG is continuously updated making research more representative of the current misinformation landscape and near real-time integration into applications possible.

At the time of writing,[4] CimpleKG contains over 203k ClaimReview[5] spanning 26 languages, issued by 77 fact-checkers from over 36 countries. CimpleKG is updated daily as new claims are collected from fact-checkers. The KG has over 15m triples and also includes 217k documents from static datasets (news and well-known misinformation datasets of claims and tweets), 263k+ distinct entities and 1m+ textual features. Besides the aforementioned SPARQL endpoint, the daily collected fact-checks are also freely accessible as graph and non-graph serialised databases snapshots.

[3] CimpleKG SPARQL Endpoint, https://data.cimple.eu/sparql
[4] These statistics are based on the 11th of April 2024 snapshot.
[5] ClaimReview, https://schema.org/ClaimReview.

2 Related Work

Many datasets have been produced over the past years to support a wide range of research on misinformation. Examples include datasets of claims related to COVID19 (e.g. [18,12]), general medical topics (e.g. [26]), politics (e.g. [9]), or a mixture of topics (e.g. [28]). Such datasets often consist of various combinations of true or false claims or full articles, fact-check reports, un/reliable news sources, etc. However, there are very few resources that provide a general and up-to-date misinformation dataset, in a KG representation, consisting of detailed information on claims and their corresponding fact-checks.

ClaimsKG [27] is one of the first KG datasets to provide a collection of fact-checked content. Their database relies on the ClaimReview data published by fact-checkers. ClaimReview is a structured data markup format that is part of the Schema.org vocabulary.[6] It is used by fact-checking organisations to publish specific details about the claim being examined, the fact-checking verdict (such as true, false, or misleading), the source of the claim, the date reviewed, and other relevant information. The ClaimReview format makes it possible to link fact-checking articles to fact-checked claims, commonly in the form of URL pairs, claim descriptions and ratings as well as information about what organisation verified the claims. The structured nature of ClaimReview makes it an ideal format for consuming fact-checks and it is used by search engines and social media platforms. The last release of ClaimsKG was in January 2023 and consisted of just under 75 thousand claims collected from 13 popular fact-checking websites.[7] The limitations of this resource are centred around the small number of fact-checking websites included in the ClaimReview crawl, infrequent updates at long intervals, and the narrow scope of the KG. Considering the rapid pace at which misinformation emerges and spreads, it is critical for any supporting dataset to include the most recent claims and their verification results and include a large variety of data sources.

Google is one of the main sponsors of the ClaimReview project and provides both a user interface and an API for searching and retrieving ClaimReview data. The interface enables users to search claims using keywords and to navigate to the full fact-check article. An API is also available which enables programs to search ClaimReview data.[8]

The Google Fact Check explorer[9] is designed for exploring ClaimReview data using query terms and often returns a subset of ClaimReview objects and values. For example, the numerical value of `reviewRating.ratingValue` attribute is not usually returned, and instead only the value of `ClaimReview.textualRating` is provided. The numerical value is useful for comparing the level of the factuality of claims, whereas the textual rating is sometimes filled with textual descriptions in various languages and hence is more difficult to parse and compare.

[6] Schema.org vocabulary, https://schema.org.
[7] ClaimsKG, https://data.gesis.org/claimskg.
[8] Google ClaimReview API, https://developers.google.com/fact-check/tools/api/reference/rest/v1alpha1/claims/search.
[9] Google Fact Check Tools, https://toolbox.google.com/factcheck.

Other ClaimReview fields not returned through the Google Fact Check API are `appearances` and `firstAppearance`, which are used by fact-checkers to indicate where the claim appeared. This information is valuable for propagating claim assessments to the URLs where they appeared which can help determine the credibility of the source as a whole. This enables us to establish, for example, how many misinforming claims appeared on a certain news source or by a specific social media account.

The Database of Known Fakes (DBKF)[10] is a more recent initiative aiming at enabling users to browse through previously fact-checked documents by known organisations. It collects new fact-checks and displays them in a web-based user interface, allowing to query the database with relevant filters, such as date, language, concepts, or authors. While DBKF shares daily-updated data, it still lacks a significant amount of fact-checks (136k vs 203k for CimpleKG) and does not allow search based on textual factors, or review label.

The CimpleKG described in this paper differs from the above in volume and velocity, by continuously collecting data from a larger amount of data sources, some changes in the data model that allow more flexible queries such as adding the mapping of the normalised labels and enriching the graph with entities and other textual factors that explain misinformation. It also includes various static datasets from many sources, making it a dense and rich resource. This additional information and the frequent KG updates allows to track the spread of misinformation with more precision and more reactivity compared to the other KGs and provide additional use cases for understanding misinformation ratings across fact-checkers.

3 The Misinformation Knowledge Graph (CimpleKG)

The misinformation knowledge graph (CimpleKG) is a KG that combines several static datasets with additional daily updated content collected from fact-checked claims from organisations based in 36 different countries.

3.1 Connecting Misinformation, Reviews, Factors and Entities

The ability to assign credibility ratings to a piece of information or claim is key for the development of research and tools that try to better understand or address the proliferation of misinformation (Section 2). In this context, since the 2000s, fact-checking organisations have been created to identify and verify claims that may be misleading, incorrect or harmful [10]. The types of fact-checked content can vary from political claims to health-related claims and often involve the creation of an article that discusses identified claims and assigns them a rating or label that typically goes from completely *misinforming* to *credible*. Although these ratings or labels are not always the same between fact-checkers,

[10] The DBKF, https://www.ontotext.com/company/news/the-database-of-known-fakes-a-valuable-eu-research-result/

the way they are structured has been standardised in the Schema.org vocabulary as ClaimReview (Section 2). In this paper, we use ClaimReview as the base of our KG and extend it with additional features such as textual *Factors* and named *Entities*. These features make it easier to discover how particular claims relate.

The textual content of a Claim associated with a ClaimReview typically involves some textual features or *Factors* such as emotion, sentiment, political leaning, propaganda techniques, and the mention of conspiracy theories that affect how specific claims are perceived. These factors can be extracted, to some extent, for a better understanding of how such features are associated with particular credibility labels. We extract these aforementioned factors automatically using the models developed in [20,19]. These models reported on average an $F1$ score of 0.71 ($\pm$0.09).

Claims typically mention named entities such as specific individuals or locations. Identifying such entities makes it possible to formulate more advanced questions about claims. For instance, we can search all the claims that mention *Ukraine* or *Donald Trump*. In this paper, we extract and disambiguate entities from the claims using DBpedia spotlight[11] [13] because of its simplicity and computational performance. It also identifies broader non-named entities (e.g. "vaccine"), and supports many languages.

Misinformation-related knowledge is not always completely captured by fact-checking organisations and some of such information may be available in manually annotated research datasets [22,16] or through specific social media verification programs [24]. These data sources may provide additional contextual information not directly found in fact-checks such as social media mentions or conspiracy theory annotations. In this paper, we integrate and link many of these static datasets to the ClaimReview data as they provide additional layers of information (Section 3.4).

3.2 The CimpleKG Data Model

As mentioned in the previous section, CimpleKG reuses the Schema.org ontology (denoted with the `sc` prefix in the rest of this document). An instance of a `sc:ClaimReview` is connected to a `sc:Claim` through `sc:itemReview`. It is also connected to the organisation that fact-checked the claim through `sc:author`, as well as the issued rating through `sc:reviewRating`. We have created `co:normalizedReviewRating`[12] to provide a normalised rating which is a controlled vocabulary represented in the Simple Knowledge Organization System (SKOS) [15]. An instance of *Rating* has a name (`sc:name`) and a rating value (`sc:ratingValue`). If it is an original rating, it is also connected to the organisation that used it through `sc:author` and is connected to the corresponding normalised rating through `sc:sameAs`. *SocialMediaPostings* are linked with *Claims* with `co:related` (based on some ground-truth from some datasets). We also provide the appearance of a *Claim* with `sc:appearance`. We use `sc:mentions` to link entities

[11] DBPedia Spotlight, https://www.dbpedia-spotlight.org.
[12] We prefix the newly defined properties and types in CimpleKG with the `co` prefix.

with any textual document (*ClaimReview*, *Review*, *Claim*, *SocialMediaPostings*, *NewsArticle*). Lastly, we extract textual features on the textual content and represent this information with the predicates `co:hasEmotion`, `co:hasSentiment`, `co:hasPoliticalLeaning`, `co:mentionsConspiracy`, `co:promotesConspiracy` and `co:usesPropagandatechnique`. An illustration of the data model is shown in Figure 1 and additional details about how to query the KG can be found on KG code and data repository[13].

The CimpleKG data can be accessed through a SPARQL endpoint and as RDF dump files.[14] All URIs are dereferenceable following the linked data principles. A RESTful API has also been deployed to access the KG.

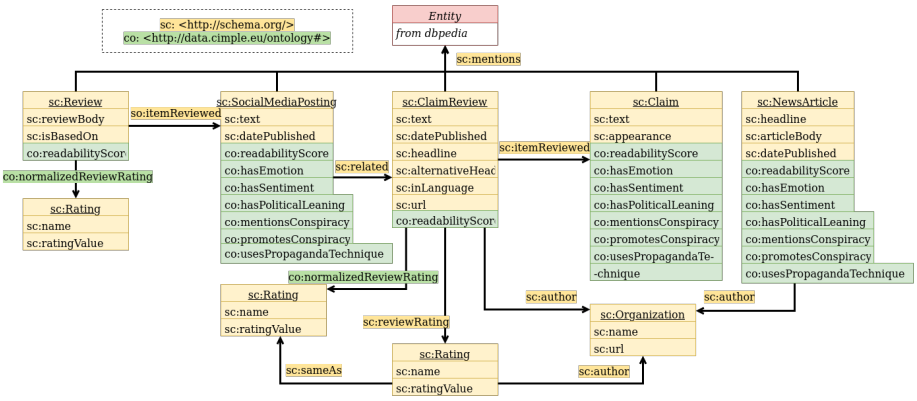

Fig. 1. Illustration of the CimpleKG data model.

3.3 Collecting and Integrating Newly Published Fact-Checks

The CimpleKG is generated using ClaimReview data collected from fact-checking organisations and various static datasets. Data from fact-checking organisations is continuously integrated whereas static datasets from static sources are added as relevant datasets once identified and published. New data is collected at 10 am UTC daily and takes 3 hours and 20 minutes to process on average.

To integrate newly published fact-checks into CimpleKG, we rely on a two-step process where: 1) data is continuously collected from fact-checking sources and, then; 2) the collected data is mapped to the CimpleKG graph structure presented in Section 3.2. During this step, both related entities and additional textual features are extracted to complement the KG with additional relevant knowledge. The various steps required for collecting and processing the data are displayed in Figure 2.

[13] CimpleKG repository, https://github.com/CIMPLE-project/knowledge-base.
[14] The RDF dumps and their automation are available as releases in https://github.com/CIMPLE-project/knowledge-base.

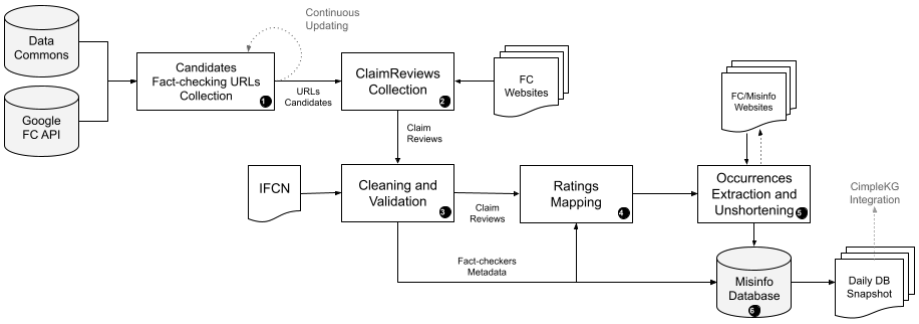

Fig. 2. Data collection and processing pipeline for gathering ClaimReviews.

The two-step process generates two different versions of the misinformation data. First, the semi-structured data created as part of the data collection step is made available daily as a set of files[15]. Second, the KG version of the data is integrated into CimpleKG and made available. The data collection and processing steps of the various fact-checks that are integrated into CimpleKG are shown in Figure 2 and can be divided into 6 primary steps.

1. *Collection of ClaimReviews URLs Candidates:* The first step required for collecting the fact-checks is to identify the URLs that contain them. We collect this data from DataCommons[16] irrespective of their publication language using their public data feed and we use the Google Fact-checking API for obtaining additional URLs. We use these two aggregators because they contain the largest quantity of fact-checks, and they are updated very frequently. Going manually to all the IFCN signatories would require additional custom collection logic, while these aggregators can already provide the data together. For both data sources, we collect the URLs of the reviews. The other fields, especially from Google Fact-checking API, tend to be incomplete in the `appearance` and `firstAppearance` attributes. Instead, when scraping from the Google Fact-checking search interface background data, we are able to retrieve URLs of appearance, but they are frequently mixed with URLs whose stance does not support the claim. Since these fields are critical for understanding *where* misinformation happens, we find it best to recollect them directly from the fact-checkers.
2. *Collection of ClaimReview from fact-checkers:* The second step involves the retrieval of the ClaimReview data associated with the previously identified URLs directly from the fact-checkers' websites. This step is needed because the data collected from the previous step may be incomplete (the issue of missing `appearance`). For each URL collected during the first step, we obtain the page content where the corresponding ClaimReview appears. For some

[15] ClaimReview data, https://github.com/MartinoMensio/claimreview-data/releases.

[16] DataCommons, https://www.datacommons.org/factcheck/download#fcmt-data.

fact-checkers, the ClaimReview is not embedded in the source of the page, because the submission to Google may be performed on a private channel. As we see in Table 1, for most of the fact-checkers, we can collect the data with complete attributes, while with some fact-checkers the recollection fails (total recollection percentage: 71.07%, average recollection percentage: 50.87%). In Section 6, we discuss how recollection can be improved in future updates.

3. *Validation and Cleaning:* The third step is designed for cleaning and validating the data collected in the previous step as some of the data may be wrong or incomplete. To make the collected data usable, we try to fix and normalise it with several processes (e.g. dirty-json[17] to fix common JSON errors with strings or use multiple parsers to allow parsing JSON-LD transformed with different specifications). We discard items that are not easily fixable and, for the remaining ClaimReview, we only keep the ones that are from International Fact-Checking Network (IFCN) signatories[18] in order to ensure that the collected data is trustworthy (we discard 63,955 ClaimReview that cannot be verified). The list of IFCN signatories is updated every time new data is collected and this data is used for adding information about fact-checking organisations such as their country of origin and language.
4. *Ratings Mapping:* Since each fact-checker uses a different type of rating, we need to map them to a common value (step 4 in figure 2). Similar to our previous work [14], we first try to use the numerical ratings provided by fact-checkers. We use the following mappings: *credible* when the rating is greater than 0.8, *mostly_ credible* when the rating is between 0.6 and 0.8, *uncertain* when the rating is between 0.4 and 0.6, *not_ credible* when the rating is less than 0.4, and *not_ verifiable* when the numerical value is missing. For the textual labels, mappings are created for each fact-checker based on their textual labels so they map to the 5 aforementioned labels.
5. *Occurrences Extraction and Unshortening:* The next step (step 5 in the figure) is focused on extracting the `appearance` and `firstAppearance` fields from the collected ClaimReview that have them. The extracted URLs are then unshortened since many fact-checkers use URL shorteners or archiving websites to capture snapshots of the page for the content that then gets deleted. URL unshortening allow us to know the real URL where it appeared, so it can be used for tracking their appearance online rather than the more rarely used shortened version of the URLs.
6. *Misinformation Database and Snapshot:* The final step is to store the data in a database and export it in a format that can be easily processed for integration in CimpleKG. A snapshot is created daily based on the collected data and made available publicly. The data comprises both statistical information about the collected data and various subsets of the data.[19]

[17] Dirty-JSON, https://github.com/RyanMarcus/dirty-json.
[18] IFCN, https://ifcncodeofprinciples.poynter.org/signatories.
[19] The details of the daily snapshot and the description of each exported file can be found at https://github.com/MartinoMensio/claimreview-data.

Table 1. Recollected percentages from the top 30 fact-checkers. Total recollection percentage: 71.07%, average recollection percentage: 50.87%

Web Domain	Recollected	Total	Web Domain	Recollected	Total
afp.com	86.87%	33,727	youturn.in	0.00%	4,835
snopes.com	99.98%	16,321	dpa-factchecking.com	0.00%	4,822
vishvasnews.com	99.99%	13,417	indiatoday.in	99.71%	4,498
politifact.com	51.29%	12,718	newtral.es	0.00%	4,249
newschecker.in	99.95%	11,694	newsmeter.in	99.98%	4,238
boomlive.in	99.97%	10,270	fullfact.org	100.00%	4,118
factly.in	0.10%	8,394	thequint.com	99.97%	3,969
checkyourfact.com	99.91%	8,093	usatoday.com	0.00%	3,787
leadstories.com	100.00%	7,719	aosfatos.org	99.97%	3,559
altnews.in	99.96%	7,270	maldita.es	0.00%	3,440
factcrescendo.com	0.06%	6,992	dogrulukpayi.com	99.91%	3,422
uol.com.br	35.86%	6,926	correctiv.org	100.00%	3,331
demagog.org.pl	92.08%	6,088	factcheck.org	41.44%	2,985
sapo.pt	100.00%	6,020	observador.pt	100.00%	2,908
teyit.org	93.01%	4,953	tfc-taiwan.org.tw	0.00%	2,871

The integration of the daily collected data into CimpleKG follows also six primary steps. The code used for converting the daily snapshots is available at https://github.com/CIMPLE-project/knowledge-base.

1. *Claim Review text scrapping:* First, we extract the textual data of the new ClaimReview documents. We use the `trafilatura` python package [1] to retrieve the body of the Claim Review from the specified URL.
2. *Entity extraction:* We use DBpedia spotlight [13] to extract relevant entities in the text of the claims, and ClaimReview. This results in 192,183 distinct entities extracted. We also experimented with the latest spaCy models leveraging on LLMs that also extract non-named entities.
3. *Factors extraction:* We also extract *factors* from the textual content of the claim (Section 3.1). This results in 497,182 *factors* extracted.
4. *Conversion of objects to RDF triples:* Then, each `Claim`, `ClaimReview`, `Organization` and `Rating`[20] are converted to RDF triples. They are associated with their respective types and properties (e.g. name, datePublished, URL, etc). For each resource, we generate a unique URI identifier using the SHA224 cryptographic hash function over a unique string identifier[21]. This way ClaimReviews fact-checking the same claim will point to the same document in the KG.
5. *Connection of the objects:* We connect resources through the following Schema.org properties: author, mentions, reviewRating, itemReviewed and appear-

[20] Both original and normalised ratings are accessible

[21] The CimpleKG URI patterns are specified at: https://github.com/CIMPLE-project/converter/blob/main/URI-patterns.md.

ance. We also define our own set of properties for the tracking of *factors*. This results in a graph totalling 8,454,322 RDF triples.

6. *Mapping of the KG and serialisation:* Lastly, to map the collected data to the CimpleKG model, we use the RDFLib python library, and serialise it using the TTL file format. The data is then integrated into CimpleKG.

3.4 Integrating Static Datasets with the Fact-checks

Integrating previously published misinformation datasets into the KG makes it possible to link existing fact-checked claims with related data such as social media posts (`sc:SocialMediaPost`) and news articles (`sc:NewsArticle`). Table 2 shows the statistics of these static datasets. In this work, we have specifically integrated datasets of tweets and claims labelled as misinformation related to COVID-19. As with the ClaimReview data integrated into CimpleKG (Section 3.3), we extract the entities and textual factors from the text of these documents.

1. *Community Notes (BirdWatch):* Community Notes (CN) is a program where users can identify and review potential misleading tweets/X posts. In our work, we add relevant posts with their CN review to CimpleKG, as well as their rating. We also use the dataset provided in [24] that links some tweets with the ClaimReview data integrated into CimpleKG (Section 3.3). This does not require any disambiguation as the ground truth data provides the link between the tweet and the ClaimReview.
2. *CLEF CheckThat! 2022:* The CLEF CheckThat! 2022 dataset [16] contains social media posts linked to claim reviews. The data is ingested into CimpleKG by disambiguating the claim reviews in the mentioned dataset and the daily collected ClaimReview data. The disambiguation is done by generating a unique URI from the claim text, the fact-check url and the review rating. If a document in the ClaimReview data already exists with these values, it is reused, else we create a new document.
3. *MediaEval-FND 2022:* The MediaEval-FND 2022 dataset [22] focuses on conspiracy theories mentioned on Twitter/X during the coronavirus outbreak. This dataset provides Tweets and ground truth for Covid-related conspiracy theories *factors*.
4. *AFP:* The AFP dataset contains news articles collected through Agence France Presse (AFP). This data showcases the journalists' online discourse and adds a different context to social media posts and fact-checking articles. Note that this dataset is not directly linked to claims or reviews. However articles are linked to other relevant concepts like entities and factors which are, in turn, connected to claims, ClaimReviews, and tweets within CimpleKG.
5. *Propaganda corpus:* The Propaganda corpus [5] focuses on propaganda detection in news articles. It originally contained more than 451 articles annotated with 18 propaganda techniques[22]. We simplified the data by splitting

[22] QCRI propaganda techniques, https://propaganda.qcri.org/annotations/definitions.html.

the articles into sentences and only keeping the most prevalent techniques[23]. This results in 1,908 claims.

Table 2. Statistics of the static datasets integrated into CimpleKG.

Dataset	Document Types	Nb. of Documents
AFP	News Article.	193,933 news articles.
Birdwatch	Social Media Posts, Reviews.	6,563 tweets, 1,983 reviews, 1,112 links to ClaimReview.
CLEF CheckThat!	Social Media Posts, Claim Reviews.	1,196 tweets, 1,198 links to ClaimReview.
MediaEval 2022	Social Media Posts.	2,702 tweets.
Propaganda Corpus	Claims.	1,908 claims.

Extraction of factors and entities is also performed on the static datasets[24], and then all objects are converted to RDF triples and integrated into the CimpleKG, along with the ClaimReview data. The static datasets represent 6,782,846 triples, totalling around 45% of Cimple KG, and include 624,402 textual factors.

4 CimpleKG Statistics

This section provides statistics about the misinformation data integrated into CimpleKG. These statistics are based on the 11$^{\text{th}}$ April 2024 database snapshot.

4.1 Fact-checkers and Language Statistics

The current fact-checked data integrated into CimpleKG contains ClaimReview from 77 different fact-checking agencies based in 36 different countries and publishing fact-checks in 26 different languages. As shown in Table 3, most fact-checks are published as English (37.7%), followed equally by French and Portuguese (respectively representing 9.1% and 7.8% of the languages found in the data). However, as displayed in Table 4, the country with the most IFCN-registered fact-checking organisations is India (18.2%) followed by France (10.4%) and the USA (9.1%).

[23] Name Calling/Labelling, Repetition, Slogans, Appeal to fear/prejudice, Doubt, Exaggeration/Minimisation, Flag-Waving, Loaded Language, Causal Oversimplification, Appeal to Authority, Black-and-White/Fallacy

[24] For news articles, factors are only computed on headline and first paragraph, as those sections contain the most important information per journalistic practice

Table 3. Distribution of ClaimReview languages for the fact-checkers found in continuously updated fact-checkers data.

Language	Amount	Proportion	Language	Amount	Proportion
English	29	37.7%	Croatian	1	1.3%
French	7	9.1%	Danish	1	1.3%
Portuguese	6	7.8%	Dutch	1	1.3%
Spanish	6	7.8%	Filipino	1	1.3%
Hindi	3	3.9%	German	1	1.3%
Italian	3	3.9%	Greek	1	1.3%
Polish	3	3.9%	Indonesian	1	1.3%
Turkish	2	2.6%	Nepali	1	1.3%
Albanian	1	1.3%	Norwegian	1	1.3%
Arabic	1	1.3%	Russian	1	1.3%
Bangla	1	1.3%	Serbian	1	1.3%
Bulgarian	1	1.3%	Serbo-Croatian	1	1.3%
Catalan	1	1.3%	Telugu	1	1.3%

Table 4. Top 10 countries with the most fact-checkers.

Country	Amount	Proportion
India	14	18.2%
France	8	10.4%
USA	7	9.1%
Brazil	4	5.2%
Italy	4	5.2%
Poland	3	3.9%
Turkey	3	3.9%
United Kingdom	3	3.9%
Australia	2	2.6%
Portugal	2	2.6%

Table 5. Top 10 countries with the most fact-checks.

Country	Amount	Proportion
India	58,49	28.6%
USA	40,468	19.9%
France	31,605	15.6%
Brazil	9,302	4.6%
Portugal	8,928	4.4%
Turkey	8,767	4.3%
Poland	7,244	3.6%
United Kingdom	5,878	2.9%
Italy	3,840	1.9%
Germany	3,342	1.6%

4.2 Fact-checks Statistics

Currently, the fact-checked data integrated into CimpleKG contains 203, 209 fact checks with most of the reviewed claims identified as *Not Credible* (69.8%) or *Not Verifiable* (15.6%). The remaining claims are identified as *Credible* (6%), *Uncertain* (7.2%) or *Mostly Credible* (1.4%). As displayed in Table 5 and Figure 3, most of the fact-checks are produced by India (28.6%), followed by the USA (19.9%) and France (15.6%) with AFP fact checking from France producing the most fact-checks (14.4%) followed by Snopes.com from the USA (8%).

4.3 Entities and Factors Statistics

CimpleKG currently contains 15, 237, 168 triples that describe 203, 209 fact-checked claims and 217, 616 static documents obtained from the static datasets

Table 6. Top 15 fact-checking organisations with the most fact-checks.

Organisation	Country	Amount	Proportion
AFP fact checking	France	29,300	14.4%
Snopes.com	United States of America	16,318	8.0%
MMI Online Limited	India	13,416	6.6%
Newschecker	India	11,746	5.8%
BOOM	India	10,267	5.1%
Check Your Fact	United States of America	8,086	4.0%
Lead Stories	United States of America	7,719	3.8%
Pravda Media Foundation	India	7,268	3.6%
Demagog Association	Poland	6,718	3.3%
PolitiFact	United States of America	6,523	3.2%
Polígrafo	Portugal	6,020	3.0%
Full Fact	United Kingdom	5,656	2.8%
Teyit	Turkey	4,607	2.3%
TV Today Network Limited	India	4,485	2.2%
Newsmeter	India	4,237	2.1%

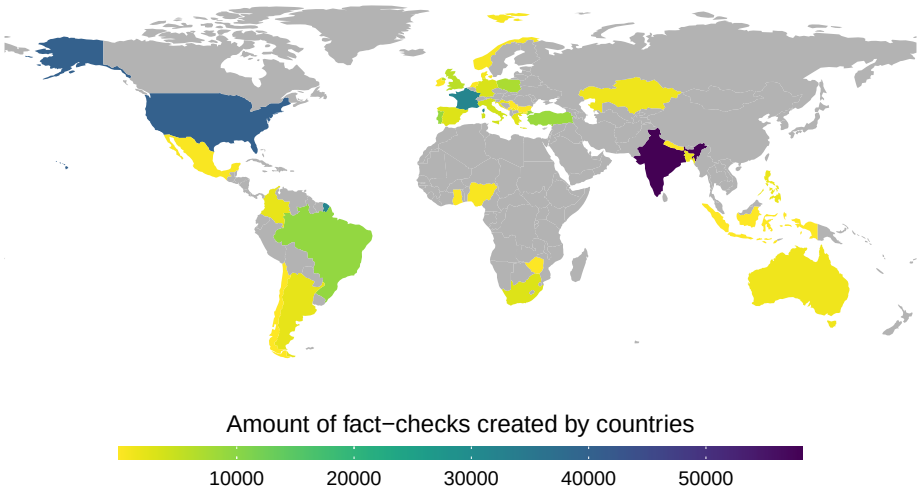

Fig. 3. Amount of fact-checks created for each country.

discussed in Section 3.4. Using DBpedia spotlight [13], we extracted 263,243 distinct entities while our various BERT-based supervised models have extracted 1,121,584 textual factors [20,19]. As mentioned in Section 2, our KG differs from previous efforts such as ClaimsKG in a few distinctive ways. CimpleKG spans more fact-checking organisations, and more different languages than previous work and is updated daily with the most recent fact-checks. Our graph also provides normalised ratings between the original ratings used by the different fact-

checking organisations, we also unshorten the URLs mentioned in fact-checks and extract textual factors from documents and the text of the claims.

5 CimpleKG Use Cases and Usage

The CimpleKG dataset has been used in multiple research studies and is integrated into multiple applications:

- *Misinfome Bot* (https://twitter.com/MisinfomeB) is a social media bot that automatically corrects misinformation spreaders by posting fact-checks to known misinformation sharers. The bot uses CimpleKG to identify recent misinformation and fact-checks URLs (Listing 1.1). It was used for understanding the impact of automated misinformation corrections in social media [4].
- *Fact-Checking Observatory* (FCO, https://fcobservatory.org/): The FCO monitored the spread of misinformation and corresponding fact-checks during the COVID-19 pandemic, taking into account their topics, language, and geographic location of fact-checkers. FCO used the pairs of misinformation links and their fact-checks to track their spread on Twitter/X. The FCO data was used for studying the co-spreading relationships between misinformation and fact-checks during the COVID-19 pandemic [2,3].
- *Iffy Index* (https://iffy.news/index/) is an external website that collects source-credibility assessments from multiple sources, including our Misinformation dataset. Iffy has been used in 24 research papers and several tools.
- *Exploratory Search Engine* (https://explorer.cimple.eu/) enables searching and browsing for the KG. Claims and tweets can be discovered based on filters corresponding to the language, rating, entities being mentioned or the different factors (sentiment, emotion, political leaning, conspiracy theory, propaganda technique, etc.) that have been computed and are related to ClaimReview that have been harvested. For example, https://explorer.cimple.eu/reviews/claim/fd86a971142de3b2f5705eeaf95ade6c8a900c87227b3799a1bc2f86 shows the view of a particular Claim and its annotations.
- CimpleKG was used by a large-scale study that compared fact-checking by experts (ClaimReview) against those done by the crowd (Twitter BirdWatch/Community Notes) [24,23]. This study relied on the data contained in CimpleKG, and discovered that, in some settings, crowdsourced fact-checks are comparable to those performed by expert fact-checking organisations.
- CimpleKG data was used by the *Linked Credibility Reviews* system [6] and for performing explainable misinformation detection [8,7]. The authors used CimpleKG data to run their experiments and evaluations.

The KG can be queried using the ontologies described in Section 3.2. For example, using the DBPedia ontology and resources, we can obtain the individuals (`dbo:Person`) that are the most associated with Donald Trump (`dbr:Donald_Trump`). We can also easily obtain information about how the original fact-checker ratings are mapped to normalised ratings using the `schema:Rating` type and

`schema:sameAs` property. Finding recent misinformation and fact-checks URLs pairs can be performed using the SPARQL query in Listing 1.1. Such a query is used by the *Misinfome Bot* when looking for misinformation spreaders. Additional query examples can be found on the KG data repository.

```
PREFIX sc: <http://schema.org/>
PREFIX co: <http://data.cimple.eu/ontology#>
SELECT DISTINCT ?fc_url ?misinfo_url
WHERE {
    ?rev a sc:ClaimReview ;
         sc:url ?fc_url ;
         sc:datePublished ?date_published ;
         co:normalizedReviewRating ?rating ;
         sc:itemReviewed ?claim .
    ?claim a sc:Claim ;
           sc:appearance ?misinfo_url .
    ?rating sc:ratingValue "not_credible" .
    FILTER (?date_published >= xsd:date("2024-03-11")) .
}
ORDER BY DESC(?date_published)
LIMIT 10
```

Listing 1.1. SPARQL Query used by the Misinfome Bot for retrieving the 10 most recent *not_credible* fact-checks and misinformation URL pairs published since the 11$^{\text{th}}$ of March 2024.

6 Maintenance, Limitations and Future Work

Our data collection approach has continuously been refined since we started collecting ClaimReviews in 2019. We plan to add more static datasets and calculate additional factors as we support more tools and websites that use CimpleKG (Section 5). We made the code of CimpleKG available (see Section 3.3) so anyone can run their instance of CimpleKG, report issues and improve the KG.

While fact-checkers are developing MediaReview,[25] a system similar to ClaimReview, to describe different media types, it's still being finalised. Once this format is widely adopted, we will incorporate MediaReviews into CimpleKG.

We have noticed that the ClaimReviews collected through DataCommons and the Google Fact-checking API are sometimes malformed or refer to URLs that do not always contain the reviewed claims. For example, some fact-checkers may use a domain (e.g. `instagram.com`) rather than a specific URL (e.g. a specific Instagram post) or inverse the claim URL with the fact-check URL. Although we try to fix these errors automatically (Section 3.3), this is not always possible. For example, the JSON-LD of the ClaimReviews found on fact-checking

[25] MediaReview, https://www.claimreviewproject.com/mediareview.

websites may not be parsable. To ensure the accuracy of our data, we opt to recollect the information directly from the fact-checkers. As a consequence, our collection contains fewer fact-checks as 28.93% of the DataCommons fact-checks are dropped during the recollection process due to the quality assurance issues we encounter. In CimpleKG, we prioritise quality over quantity and will continue to fine-tune our data collectors to accommodate the fact-checks that cannot be accurately recollected with the generic ClaimReview collection methods.

One unique aspect of CimpleKG is the addition of textual factors. These factors are detected using multiple BERT-based models trained mostly on short textual documents (i.e., tweets). This makes these models less reliable when used on longer text (e.g., news articles). In the future, we plan to improve the accuracy of these models by training them on more diverse data and to integrate newer factor detection models, such as persuasive techniques and narratives [21].

7 Conclusions

In this paper, we described a new semantic resource called CimpleKG, which uses a knowledge graph to store and represent an ever-growing dataset of misinformation. The data consists of continuously collected data from 77 fact-checking organisations and data from several static datasets. Currently, CimpleKG contains over 15 million RDF triples that describe 203,209 fact-checked claims and 217,616 documents from static misinformation datasets. It also contains 263,243 distinct entities and 1,121,584 textual features to further describe the ingested documents and claims. CimpleKG is freely available and has already been used by numerous studies and tools and is continuously updated daily.

Resource Availability Statement: Besides the URLs mentioned in the paper, the resources can also be resolved using the following persistent URLs: 1) the snapshots for the dynamic data collected for creating CimpleKG are available at https://purl.org/net/cimplekg/claimreviews; 2) the CimpleKG SPARQL endpoint can be found at https://purl.org/net/cimplekg/sparql, and; 3) The code used for mapping the collected data to CimpleKG and the daily KG snapshots can be downloaded from https://purl.org/net/cimplekg/knowledge-graph. The data collected for creating CimpleKG and CimpleKG are both licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0) license.[26]

Acknowledgement: This work was supported by the European CHIST-ERA program within the CIMPLE project (grant agreement CHIST-ERA-19-XAI-003).

[26] CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Barbaresi, A.: Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In: 59th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations. pp. 122–131. Association for Computational Linguistics (2021)
2. Burel, G., Farrell, T., Alani, H.: Demographics and topics impact on the co-spread of covid-19 misinformation and fact-checks on twitter. Information Processing & Management **58**(6) (2021)
3. Burel, G., Farrell, T., Mensio, M., Khare, P., Alani, H.: Co-spread of misinformation and fact-checking content during the covid-19 pandemic. In: 12th International Conference on Social Informatics (SocInfo). pp. 28–42. Springer (2020)
4. Burel, G., Tavakoli, M., Alani, H.: Exploring the impact of automated correction of misinformation in social media. AI Magazine **45**(2), 227–245 (2024)
5. Da San Martino, G., Barrón-Cedeño, A., Nakov, P.: Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In: Feldman, A., Da San Martino, G., Barrón-Cedeño, A., Brew, C., Leberknight, C., Nakov, P. (eds.) Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 162–170. Association for Computational Linguistics, Hong Kong, China (2019)
6. Denaux, R., Gomez-Perez, J.M.: Linked credibility reviews for explainable misinformation detection. In: 19th International Semantic Web Conference (ISWC). pp. 147–163. Springer (2020)
7. Denaux, R., Gómez-Pérez, J.M.: Sharing retrieved information using linked credibility reviews. In: ROMCIR@ ECIR. pp. 59–65 (2021)
8. Denaux, R., Mensio, M., Gomez-Perez, J.M., Alani, H.: Weaving a semantic web of credibility reviews for explainable misinformation detection. In: 13th International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization (2021)
9. D'Ulizia, A., Caschera, M.C., Ferri, F., Grifoni, P.: Repository of fake news detection datasets. 4TU.ResearchData. Dataset (2021)
10. Graves, L., Cherubini, F.: The rise of fact-checking sites in europe. Digital News Project Report (2016)
11. Lewandowsky, S., Ecker, U.K., Cook, J.: Beyond misinformation: Understanding and coping with the "post-truth" era. Journal of Applied Research in Memory and Cognition **6**(4), 353–369 (2012)
12. Memon, S.A., Carley, K.M.: Characterizing covid-19 misinformation communities using a novel twitter dataset. In: 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN 2020) (2020)
13. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: 7th International Conference on Semantic Systems. p. 1–8. Association for Computing Machinery, New York, NY, USA (2011)
14. Mensio, M., Alani, H.: News source credibility in the eyes of different assessors. In: International Conference for Truth and Trust Online (2019)
15. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. Working draft, W3C (2008)
16. Nakov, P., Barrón-Cedeño, A., da San Martino, G., Alam, F., Struß, J.M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G.K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y.S., Wiegand, M.,

Siegel, M., Köhler, J.: Overview of the CLEF–2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In: Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 495–520. Springer International Publishing (2022)
17. Nyhan, B., Porter, E., Reifler, J., Wood, T.J.: Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. Political Behavior **41**(1) (2019)
18. Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Fighting an infodemic: Covid-19 fake news dataset. In: First International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT). pp. 21–29. Springer (2021)
19. Peskine, Y., Papotti, P., Troncy, R.: Detection of COVID-19-Related Conpiracy Theories in Tweets using Transformer-Based Models and Node Embedding Techniques. In: Multimedia Benchmark Workshop (2022)
20. Peskine, Y., Troncy, R., Papotti, P.: Analyzing COVID-Related Social Discourse on Twitter using Emotion, Sentiment, Political Bias, Stance, Veracity and Conspiracy Theories. In: 3rd International Workshop on Knowledge Graphs for Online Discourse Analysis (BeyondFacts) (2023)
21. Peskine, Y., Troncy, R., Papotti, P.: Eurecom at semeval-2024 task 4: Hierarchical loss and model ensembling in detecting persuasion techniques. In: Proceedings of the 18th International Workshop on Semantic Evaluation. SemEval 2024, Mexico City, Mexico (June 2024)
22. Pogorelov, K., Schroeder, D.T., Brenner, S., , Maulana, A., Langguth, J.: Combining Tweets and Connections Graph for FakeNews Detection at MediaEval 2022. In: Multimedia Benchmark Workshop (2022)
23. Saeed, M.: Employing Transformers and Humans for Textual-Claim Verification. Ph.D. thesis, Sorbonne Université (2022)
24. Saeed, M., Traub, N., Nicolas, M., Demartini, G., Papotti, P.: Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In: 31st ACM International Conference on Information & Knowledge Management. pp. 1736–1746 (2022)
25. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
26. Srba, I., Pecher, B., Tomlein, M., Moro, R., Stefancova, E., Simko, J., Bielikova, M.: Monant medical misinformation dataset: Mapping articles to fact-checked claims. In: 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2949–2959 (2022)
27. Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., Todorov, K.: Claimskg: A knowledge graph of fact-checked claims. In: 18th International Semantic Web Conference (ISWC). pp. 309–324. Springer (2019)
28. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)