Exploring Fact-checked Claims and their Descriptive Statistics

Malo Gasquet, Darlene Brechtel, Matthaus Zloch, Andon Tchechmedjiev, Katarina Boland, Pavlos Fafalios, Stefan Dietze, Konstantin Todorov

Exploring Fact-checked Claims and their Descriptive Statistics

Malo Gasquet[1], Darlène Brechtel[1], Matthäus Zloch[2], Andon Tchechmedjiev[3], Katarina Boland[2], Pavlos Fafalios[4], Stefan Dietze[2,5], Konstantin Todorov[1]

[1]LIRMM / University of Montpellier / CNRS, France
[2] GESIS - Leibniz Institute for the Social Sciences, Germany
[3] LGI2P, IMT Mines-Ales, France
[4] Institute of Computer Science, FORTH-ICS, Greece
[5] Heinrich-Heine-University Düsseldorf, Germany

Abstract. ClaimsKG is a RDF knowledge graph of fact-checked claims and related metadata, such as their truth values, authors or dates. It gathers information from popular fact-checking websites, annotates claims with related entities from DBpedia, and lifts the data into RDF by using a dedicated RDFS model. We present two open source, user-friendly Web-platforms operating on top of ClaimsKG: (1) the ClaimsKG Explorer – an engine to conduct ad-hoc/faceted search over the graph, and (2) the ClaimsKG Statistical Observatory – a tool allowing to extract and visualize detailed statistics of the ClaimsKG data.[1]

Keywords: Claims Search and Statistics; Fact-checking; Knowledge Graphs

Introduction. In times when we see misinformation spreading faster than truth [1], fact-checking organizations around the world, like Politifact or Snopes,[2] mobilize efforts to respond to this phenomenon. Large amounts of claims are processed weekly in order to manually assess their truthfulness, based on journalistic analysis of sources and context. However, metadata for fact-checked claims are spread across various distinct platforms on the Web where truth ratings are expressed in different ways and usually no explicit structured data are provided to facilitate search for claims and ratings meeting particular criteria.

In an attempt to provide support to scientific studies and facilitate search for claims metadata on the Web, we have created ClaimsKG, an RDF knowledge graph (KG) of fact-checked claims, enabling structured queries about their truth values, authors, dates, related entities and metadata [2]. ClaimsKG is generated through a pipeline, which periodically harvests data from popular fact-checking websites. The claims and their review articles are annotated with entities from DBpedia and described by a specific RDFS model based on established vocabularies such as schema.org and NIF. A normalised truth ratings scheme is introduced, containing four generic categories: *true*, *false*, *mixed* and *other*. Federated

[2] `https://www.politifact.com/`, `https://www.snopes.com/`

SPARQL queries enable advanced information discovery and exploitation of data from various sources. The resource currently contains 28,383 claims published since 1996 on six fact-checking portals. For more information, we refer to our resource paper [2] and to ClaimsKG's website (link given in Table 1).

Table 1: Links for access to ClaimsKG, the tools and their source codes.

ClaimsKG website	`https://data.gesis.org/claimskg/site`
SPARQL endpoint	`https://data.gesis.org/claimskg/sparql`
Explorer	`https://data.gesis.org/claimskg/explorer`
Explorer source code	`https://github.com/claimskg/claimskg-explorer`
Stat. Observatory	`https://data.gesis.org/claimskg/observatory`
Stat. Obs. source code	`https://github.com/claimskg/claimskg-statistical-observatory`

We provide a demonstration of two open source user-friendly web applications operating on top of the SPARQL endpoint of ClaimsKG [2]. *The ClaimsKG Explorer* is a web interface to conduct exploratory search over the graph. *The ClaimsKG Statistical Observatory* provides up-to-date statistics of the KG, allowing to conduct research into trends or particular events, as well as to monitor the overall "health" of ClaimsKG (e.g., revealing missing data after an update of the KG). Both applications aim to facilitate access to the data of ClaimsKG for non-computer science users, such as sociologists or journalists, who do not necessarily "speak" SPARQL. While most of the fact-checking web-portals provide (limited) search options, they are only restricted to their respective own content. Our applications provide a central entry point to a wealth of fact-checked claims and quality structured data in support of journalistic or social science research into the analysis of societal debates on various topics or events of interest. The links of the claims via their entities to DBpedia and the LOD cloud allow to discover more contextual information before resuming the search. Table 1 provides links for access to the online applications and their source code.

Overview of the Tools. The *ClaimsKG Explorer* allows to conduct search over ClaimsKG based on various filters, such as the claim author, truth value or time of utterance and to navigate through the content of the graph. It is developed within the Angular framework and is entirely dynamic: all data are collected and processed in real time via HTTP requests to the ClaimsKG SPARQL endpoint through the Virtuoso API. The results are retrieved as JSON files and the information is provided through a Web user interface. Each query is processed with respect to the different filters in order to reduce the response time. For example, auto-completion (of entities or authors) is handled by sending queries as soon as a user types the third letter, in order to reduce the number of results. Given the dynamic nature of all operations, a possible update of the KG will not impact the web application. However, if the structure of the graph or the use of vocabularies is changed, the Explorer should be updated accordingly. The tool can be easily set up on another endpoint containing similarly structured data.

The *ClaimsKG Statistical Observatory* is a web application that allows to extract and visualize quantitative descriptors pertaining to ClaimsKG's content and to monitor the graph's quality after possible updates. The application is developed through the Python Flask framework but interacts with the KG through

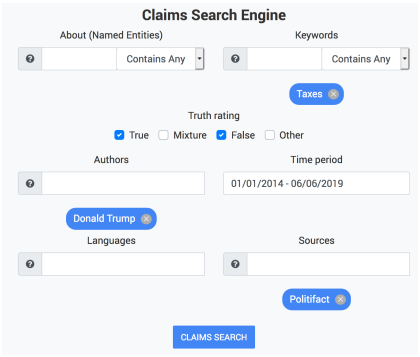
its SPARQL endpoint. Given the query-intensive nature of the operation, the Observatory generates the statistics through an update step so as not to make large separate queries for each user but rather to compute the statistics once when the KG is updated. A REST API allows for an easy update of the statistics, alongside with the application interface.

Demonstration Scenarios. Imagine a journalist who would like to analyze true and false claims by D. Trump regarding taxes since 2014. After opening the *ClaimsKG Explorer*, she can access the search engine by clicking on "Explore" or on "Search" in the menu, where she can also obtain information about the project, its contributors or get an overview of the data statistics. The Search engine (Fig. 1a) provides the possibility to filter the data based on several criteria: 1) a set of named entities contained in the body of the claim only or both in the claim and the text of its review; 2) a set of keywords ("Taxes" in our example); 3) the truth rating of the claims ("True" and "False" here); 4) the author(s) of the claims (Trump in our case); 5) the time frame (since 2014); 6) the language and 7) the fact-checking portal that has reviewed the claims. After clicking on "Claims search", she is led to the results page (Fig. 1b) with the claims corresponding to the selected criteria, which can be exported by selecting a number of attributes (Fig. 2a) either as a sub-graph of the RDF KG or as a CSV file. Clicking on a specific claim from the search results list allows to obtain detailed information, such as its date of publication, its source as well as a list of references, keywords and entities contained in the claim or its review (Fig. 2b). From there, one can access the web page of the claim in its corresponding fact-checking website by clicking on the document icon. Note that all entity mentions are clickable, allowing to navigate through the content of the graph.

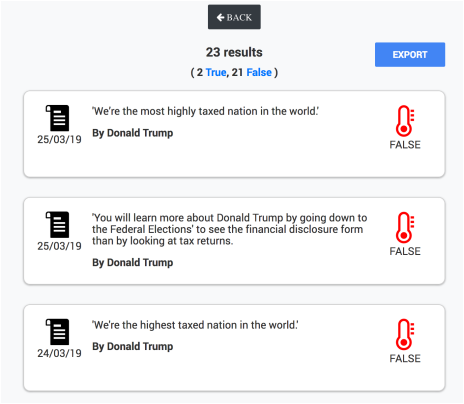
Now imagine a social scientist who is interested in studying the evolution of the political discourse about immigration over the last eight years. They would open the *Statistical Observatory* and select the "By theme" menu item at the center top of the page. They would then scroll down until reaching the "Number of claims by theme by year" chart, then find "immigration" in the legend of the chart and double-click it to single it out. The chart would now only show the selected topic, revealing that fact-checked claims about immigration were almost non-existent in 2012, but as the 2016 U.S. election drew close, the number of claims pertaining to immigration jumped from 1 in 2012 to 66 in 2016 and continued growing: as of the end of 2018 the number of claims about immigration was multiplied by almost four (221 in Dec 2018) reflecting the increased attention to that topic (cf. the bottom chart in Fig. 3). Analysing the content of the claims, the scientist would get further information on emerging topics.

References

1. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
2. A. Tchechmedjiev, P. Fafalios, K. Boland, S. Dietze, B. Zapilko, and K. Todorov, "Claimskg - a knowledge graph of fact-checked claims," in *ISWC*, 2019, to appear.

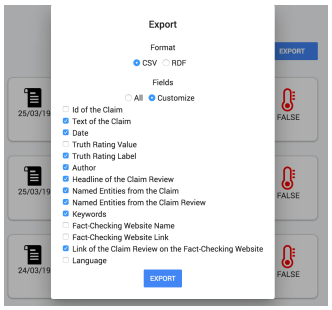


(a) The search engine.

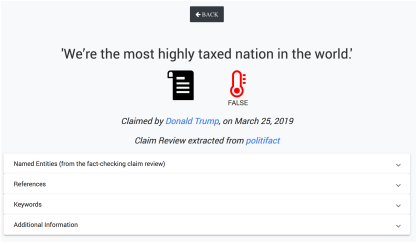


(b) Search results.

Fig. 1: The Claims Explorer user interface - I



(a) Data export options.



(b) Claim information page.

Fig. 2: The Claims Explorer user interface - II

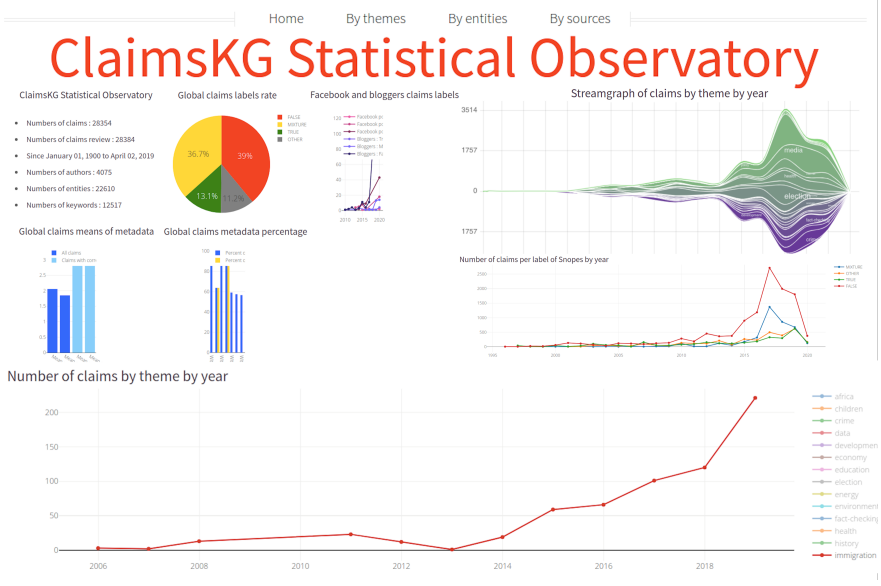


Fig. 3: Various descriptive statistics charts in the Observatory.