# Creation, Enrichment and Application of Knowledge Graphs

Simon Gottschalk[1][0000−0003−2576−4640]

L3S Research Center, Leibniu Universität Hannover, Germany
gottschalk@l3s.de

## 1 Motivation, Objectives & Contributions

The world is in constant change, and so is the knowledge about it. Knowledge-based systems – for example, online encyclopedias, search engines and virtual assistants – are thus faced with the constant challenge of collecting this knowledge and beyond that, to understand it and make it accessible to their users. Only if a knowledge-based system is capable of this understanding – that is, it is capable of more than just reading a collection of words and numbers without grasping their semantics – it can recognise relevant information and make it understandable to its users. The dynamics of the world play a unique role in this context: Events of various kinds which are relevant to different communities are shaping the world, with examples ranging from the coronavirus pandemic to the matches of a local football team. Vital questions arise when dealing with such events: How to decide which events are relevant, and for whom? How to model these events, to make them understood by knowledge-based systems? How is the acquired knowledge returned to the users of these systems?

A well-established concept for making knowledge understandable by knowledge-based systems are *knowledge graphs* (KGs), which contain facts about entities (persons, objects, locations, ...) in the form of graphs, represent relationships between these entities and make the facts understandable by means of ontologies. This thesis considers KGs from three different perspectives: (i) *Creation of knowledge graphs*: Even though the Web offers a multitude of sources that provide knowledge about the events in the world, the creation of an event-centric KG requires recognition of such knowledge, its integration across sources and its representation. (ii) *Knowledge graph enrichment*: Knowledge of the world seems to be infinite, and it seems impossible to grasp it entirely at any time. Therefore, methods that autonomously infer new knowledge and enrich the KGs are of particular interest. (iii) *Knowledge graph interaction*: Even having all knowledge of the world available does not have any value in itself; in fact, there is a need to make it accessible to humans. Based on KGs, systems can provide their knowledge with their users, even without demanding any conceptual understanding of KGs from them. For this to succeed, means for interaction with the knowledge are required, hiding the KG below the surface.

Fig. 1 summarises the contributions reported in my thesis, divided into three steps: (i) KG creation, (ii) KG enrichment, and (iii) KG application. In combination, this pipeline follows a logical order – starting from input sources and

ending in interactive demonstrators which are based on the created and enriched KGs.

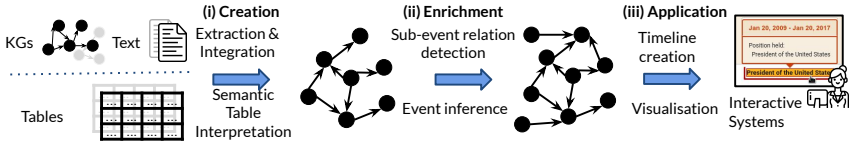


**Fig. 1.** Contributions in my PhD thesis.

## 2 Approaches, Results & Evaluation

### 2.1 Creation of an Event Knowledge Graph

Events come in many forms, including named events such as the FIFA World Cup 2018, textual events, event series and temporal relations such as a specific marriage. To integrate all such forms of events into a common resource, I have created *EventKG*.

**Approach** I have created *EventKG* [5,7], a temporal and event-centric knowledge graph, which integrates data from several sources: Wikidata, YAGO and several DBpedia language editions, as well as multilingual text data proceeded from Wikipedia and Wikipedia's Current Events Portal.

**Evaluation** A manual evaluation on a sample of events showed that the fusion of event times and event locations from different sources through rules and majority voting was successful in 75% and 94% of the cases with conflicting information from the sources (compared to 54% and 96% in the case of Wikidata, which has fewer event locations, though).

**Results** In its current version[1], *EventKG* V3.0 provides information for over 1.3 million events and over 4.5 million temporal relations, far more than any of its sources.

### 2.2 Creation of a Knowledge Graph from Tabular Data

Knowledge comes in many ways, not limited to existing knowledge graphs or encyclopedias. Another example are tabular datasets, which are structured into rows and columns but lack semantic annotations. With *Tab2KG*, I developed an approach for *semantic table interpretation* (STI) based on background knowledge which transforms tables into KGs.

[1] `http://eventkg.l3s.uni-hannover.de/`

**Approach** Typical approaches towards STI assume that most values in the table can be mapped to existing knowledge graphs. However, only about 3% of the tables contained in the 3.5 billion HTML pages of the Common Crawl Web Corpus[2] can be matched to DBpedia [12]. *Tab2KG*[3] performs STI purely based on numeric features describing the characteristics of semantic classes and table columns. I trained a Siamese neural network that rates the similarity between features and used a graph-based algorithm to identify the relations between those semantic classes.

**Evaluation** I have evaluated *Tab2KG*'s performance on STI on five different datasets against other STI baselines that do not require entity lookup in a KG. On average, *Tab2KG* increases accuracy, , i.e. the percentage of correctly identified data type relations and class relations, by more than 9% accuracy compared to the best-performing baseline.

**Results** *Tab2KG* allows to perform STI without relying on the ability to match table cells to entities or literal values in existing knowledge graphs. To make this possible and to integrate the *Tab2KG* approach with existing datasets, we have created a semantic description of domains and datasets following the DCAT vocabulary and the SEAS Statistics ontology.

### 2.3 Enrichment of an Event Knowledge Graph

After the creation of a KG, be it an event KG or any other kind of KG created from tabular data, there is no reason to assume it is complete. One reason being the open-world assumption under which there is no demand for a KG to be complete at any point in time, the other reason being the fact that the world is changing – and with it does the represented knowledge.

**Approach** I have introduced a novel approach called *HapPenIng*[4] [8] to enrich event KGs. In contrast to existing enrichment methods [11], *HapPenIng* does not only add new edges to the KG, but it does also create new nodes – without the use of any external knowledge. To do so, I rely on a specific type of nodes that lies in event KGs: event series such as the Wimbledon Championships or US presidential elections.

For event series completion in an event KG, I proposed two steps: First, I trained machine learning models to predict missing sub-event relations. Second, I ran a graph-based algorithm for the detection of missing event series editions under the assumption of similar patterns within event series.

[2] `http://commoncrawl.org/`
[3] `https://github.com/sgottsch/Tab2KG`
[4] `http://eventkg.l3s.uni-hannover.de/happening`

**Evaluation** For the first step of sub-event relation prediction, I trained a random forest classifier which has an accuracy of 0.98 in 10-fold cross validation. The second step of event inference was evaluated as follows: I could reconstruct close to half of randomly removed event nodes, with a precision of 0.70 – outperforming an embedding-based baseline with a precision of 0.26.

**Results** As a result of *HapPenIng*, new nodes are added to the KG and enriched with a label, locations and a time span. These event characteristics are inferred using a rule-based approach and a label generation algorithm based on edit distances. All together, I created a dataset of $90,000$ new sub-event relations and over $5,000$ events missing in Wikidata.

### 2.4 Application of an Event Knowledge Graph

Representing and storing knowledge alone does not imply access to it and understanding of it. Therefore, I took the last crucial step of creating applications which enable non-expert users to explore my event-centric knowledge graph.

**Approach** KGs can potentially overwhelm its users with the sheer amount on information they contain. To only show the most relevant events in the life of a person, I have first created a publicly available benchmark for training and evaluating biography timelines by mapping temporal relations found in *EventKG* to textual biographies. I trained a classifier on this benchmark and on features extracted from *EventKG*, to identify temporal relations relevant to a biography timeline.

**Results** *EventKG+BT*[5] [7,9] let's a user explore the lives of any persons in *EventKG*. Instead of making users read a whole biography text about a person of interest, they can easily interact with the generated timeline and follow what was really important in that person's life. I have also created *EventKG+TL*[6] [6], a system that let's a user explore a topic of interest by showing related events and their relevance from different language point of views. Both systems take a step towards the reduction of the workload that is necessary when closely reading encyclopedic articles.

**Evaluation** My evaluation showed that users prefer timelines created with my approach over the timelines created by the state-of-the-art Time Machine approach [2] in close to 70% of the cases. I also identified which features are particularly important for timeline creation. Finally, I have shown that *EventKG* serves as the best source for biography-relevant facts: *EventKG* contains 55% of the facts extracted from my encyclopedic benchmark, in contrast to other knowledge graphs such as YAGO and Wikidata, which only cover less than 40% of these facts.

[5] `http://eventkg.l3s.uni-hannover.de/timelines.html`
[6] `http://eventkg.l3s.uni-hannover.de/eventkg_tl.html`

## 3 Discussion and Future Work

I have presented *EventKG* – a knowledge graph that represents the happenings in the world in 15 languages – as well as *Tab2KG* – a method for understanding tabular data. For the enrichment of KGs without any background knowledge, I propose *HapPenIng*, which infers missing events from the descriptions of related events. I demonstrate means for interaction with KGs at the example of two web-based systems (*EventKG+TL* and *EventKG+BT*) that enable users to explore the happenings in the world as well as the most relevant events in the lives of well-known personalities.

*EventKG* is reusable for a variety of novel algorithms and real-world applications, including Question Answering [3], image classification [10], recommendation [1]. *EventKG* is at the core of the Open Event Knowledge Graph [4] and has been cited more than 80 times until now [5,7].

*Tab2KG* allows the integration of semantic table interpretation task with semantic data catalog descriptions and is flexible towards previously unseen data. With *HapPenIng*, I have shown that under the given circumstances in specific scenarios (e.g., event series), it is possible to go beyond traditional approaches towards knowledge inference and generate new KG nodes.

In the future work, I envision the extension of *EventKG*, e.g., by doing live updates of the KG and by involving more sources such as news articles. Finally, I would like to further contribute to the challenge of making KGs accessible to the public, with a specific focus on applications which are not focused on single tasks but provide a holistic view on world knowledge.

## References

1. Abdollahi, S., Gottschalk, S., Demidova, E.: EventKG+Click: A Dataset of Language-specific Event-centric User Interaction Traces. In: CLEOPATRA Workshop @ESWC (2020)
2. Althoff, T., Dong, X.L., Murphy, K., Alai, S., Dang, V., Zhang, W.: TimeMachine: Timeline Generation for Knowledge-base Entities. In: SIGKDD. ACM (2015)
3. Costa, T.S., Gottschalk, S., Demidova, E.: Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. In: CIKM (2020)
4. Gottschalk, S., et al.: OEKG: The Open Event Knowledge Graph. In: CLEOPATRA Workshop @ESWC (2021)
5. Gottschalk, S., Demidova, E.: EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In: ESWC. Springer (2018)
6. Gottschalk, S., Demidova, E.: EventKG+TL: Creating Cross-Lingual Timelines from an Event-Centric Knowledge Graph. In: ESWC (2018)
7. Gottschalk, S., Demidova, E.: EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. Semantic Web (2019)
8. Gottschalk, S., Demidova, E.: HapPenIng: Happen, Predict, Infer—Event Series Completion in a Knowledge Graph. In: ISWC. Springer (2019)
9. Gottschalk, S., Demidova, E.: EventKG+BT: Generation of Interactive Biography Timelines from a Knowledge Graph. In: ESWC (2020)

10. Müller-Budack, E., Springstein, M., Hakimov, S., Mrutzek, K., Ewerth, R.: Ontology-driven Event Type Classification in Images. In: WACV (2021)
11. Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web **8**(3) (2017)
12. Ritze, D., et al.: Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases. In: WWW. pp. 251–261 (2016)