Web Semantics: Science, Services and Agents on the World Wide Web 37–38 (2016) 132–151

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

# Building event-centric knowledge graphs from news

Marco Rospocher [a], Marieke van Erp [b,*], Piek Vossen [b], Antske Fokkens [b], Itziar Aldabe [c], German Rigau [c], Aitor Soroa [c], Thomas Ploeger [d], Tessel Bogaard [d]

[a] Fondazione Bruno Kessler, Trento, Italy
[b] Vrije Universiteit Amsterdam, The Netherlands
[c] The University of the Basque Country, Donostia, Spain
[d] SynerScope B.V., Helvoirt, The Netherlands

ARTICLE INFO

*Article history:*
Received 7 April 2015
Received in revised form
21 October 2015
Accepted 22 December 2015
Available online 12 January 2016

*Keywords:*
Event-centric knowledge
Natural language processing
Event extraction
Information integration
Big data
Real world data

ABSTRACT

Knowledge graphs have gained increasing popularity in the past couple of years, thanks to their adoption in everyday search engines. Typically, they consist of fairly static and encyclopedic facts about persons and organizations – e.g. a celebrity's birth date, occupation and family members – obtained from large repositories such as Freebase or Wikipedia.

In this paper, we present a method and tools to automatically build knowledge graphs from news articles. As news articles describe changes in the world through the events they report, we present an approach to create Event-Centric Knowledge Graphs (ECKGs) using state-of-the-art natural language processing and semantic web techniques. Such ECKGs capture long-term developments and histories on hundreds of thousands of entities and are complementary to the static encyclopedic information in traditional knowledge graphs.

We describe our event-centric representation schema, the challenges in extracting event information from news, our open source pipeline, and the knowledge graphs we have extracted from four different news corpora: general news (Wikinews), the FIFA world cup, the Global Automotive Industry, and Airbus A380 airplanes. Furthermore, we present an assessment on the accuracy of the pipeline in extracting the triples of the knowledge graphs. Moreover, through an event-centered browser and visualization tool we show how approaching information from news in an event-centric manner can increase the user's understanding of the domain, facilitates the reconstruction of news story lines, and enable to perform exploratory investigation of news hidden facts.



## 1. Introduction

Knowledge graphs have gained increasing popularity in the last couple of years, thanks to their adoption in everyday search engines (e.g., Google, Bing). A knowledge graph is a knowledge-base of facts about entities (e.g., persons, organizations),[1] typically obtained from structured repositories such as Freebase and Wikidata, or extracted from encyclopedic knowledge such as Wikipedia. For instance, given a famous person, knowledge graphs typically cover information such as her birth date and birth place, her relatives and the major events and activities that made her famous. However, only a small part of what happens in the world actually makes it into these databases. There are many events that are not considered important enough to be included or may not directly involve famous people that have entries. Furthermore, current repositories tend to represent the actual state of the world and do not focus on the dynamics and the changes over time. More fluid information as reported in the growing stream of daily news tends to get lost in current knowledge graphs and our fading memories, but it can be of great importance to information professionals needing to reconstruct somebody's past or the massive history of complete industries, regions or organizations. There is thus a need for a different type of structured database constructed around events rather than entities and entity-focused actual facts. Capturing this dynamic knowledge requires to consider events as the unit for storing knowledge regardless of the fame of the people involved.

* Corresponding author.
*E-mail address:* marieke.van.erp@vu.nl (M. van Erp).

[1] The description of the latest release of DBpedia is an illustrative example as it states the following: "The English version of the DBpedia knowledge base currently describes 4.58 million things [...] including 1,445,000 persons, 735,000 places [...], 411,000 creative works [...], 241,000 organizations [...], 251,000 species and 6000 diseases". Events are not mentioned. http://blog.dbpedia.org/?p=77 *Last accessed: 7 April 2015.*

http://dx.doi.org/10.1016/j.websem.2015.12.004
1570-8268/

In this paper, we present a method and an open source toolkit to automatically build such *Event-Centric Knowledge Graphs (ECKGs)* from news articles in English and Spanish, Italian and Dutch. We define an Event-Centric Knowledge Graph as a Knowledge Graph in which all information is related to events through which the knowledge in the graph obtains a temporal dimension. In a traditional KG, information is often centered around entities. One can then find RDF triples (subject, predicate, object) where the subject and object are often entities, and any information about events is generally captured through the predicate. In ECKGs the subject of triples is typically the event related to entities and bound to time. This will allow specialists to reconstruct histories over time and networks across many different people and organizations through shared events. Dynamic trends and regional changes can be made visible abstracting from individuals and reasoning over the temporal aspects.

Consider the following example on the company Porsche. In DBpedia, the entry for the company Porsche provides triples that state what type of companies it is, what cars it makes, what management it has, etc. It does not list the history of deals, the market events, the changes in managements, nor the successes and failures over a longer period of time. On 15 October 2015, the Wikipedia entry of the same company does give a brief history in natural language, including how it was fully acquired by Volkswagen in 2009 but obtained 100% voting rights within the Volkswagen group in 2013 by buying back 10% stake from Qatar Holding. This history is not represented as structured data in DBpedia. If we next look at the Wikipedia page for Qatar Holding, we also find a brief history in natural language text that is not represented as structured data in the corresponding DBpedia entry. Interestingly, the history of Qatar Holding mentions that it currently still holds about 17% stake in the Volkswagen Group and Porsche. It does not mention that 10% of this stake was sold back to the Porsche family in 2013. Apparently, this event was important for the Porsche SE history but not for the Qatar Holding history. As events are first class citizens in our ECKGs (similar to entities in many other KGs), these selling and buying events are represented as a single event in which Porsche loses an asset and Qatar Holding acquires one, regardless of the perspective of the two companies and their relevance for either one. We leave it up to the user to order events in time, place and around participants to reconstruct storylines or histories from a complete representation of all events reported in the news.

From a representational point of view, in our ECKGs every event is a node of our knowledge graph and is uniquely identified by an URI, on which various properties can be asserted via triples. This provides a homogeneous representation of events, differently from what happen in other resources: e.g., in DBpedia, an analogous representation is applied only for *named* events[2] such as http://dbpedia.org/resource/2009_Japanese_Grand_Prix, while a minimal number of *smaller* events without established name are captured by properties such as http://dbpedia.org/property/acquired.

By exploiting state-of-the-art Natural Language Processing (NLP) techniques, we automatically extract information about the *events* mentioned in millions of news articles, together with the information on the event participants, time and location. All the extracted content is organized in an ECKG in a structured representation grounded in Semantic Web best practices. Moreover, these pieces of information are linked to available linked data resources (e.g. whenever possible, entities participating in events are linked to their DBpedia referent, otherwise an entity instance in our knowledge base is created) as well as to the actual textual occurrences from which they were extracted. Determining event identity and anchoring events to time eventually results in the representation of long-term developments and story-lines, where events are connected through bridging relation such as cause or co-participation. These "histories" reconstructed from news capture changes in the world instead of static properties and facts in traditional knowledge graphs.

To construct an ECKG, we have identified four main information extraction challenges: (i) proper modeling of the expression of information in text and the referential value of the expression in the formal semantic ECKG model; (ii) correctly extracting and interpreting the information contained in a news article, according to the ECKG data model; (iii) linking the extracted information to established linked data repositories (e.g., DBpedia); (iv) establishing referential identity for entities and events across different expressions and mentions within and across different sources (e.g., same entity or event mentioned in different news articles), potentially in different languages.

Our approach tackles all four challenges, as demonstrated in the four knowledge graphs that we built in several distinct domains. The text corpora from which we have constructed our ECKGs range from a few hundred to millions of news articles. The individual modules in our processing pipeline all perform at the level of or exceed the current state-of-the-art in natural language processing technology. Our ECKGs can then be used to answer queries that are difficult to answer using traditional KGs or the unprocessed source documents, as is the current de facto standard for information professionals. To the best of our knowledge, we are the first to automatically build ECKGs from large, unstructured news article text collections. Furthermore, our method also works cross-lingually, enabling integration of ECKGs extracted from different languages.

In this paper, we combine the contributions reported in several publications on the NewsReader project from the perspective of ECKGs. These contributions cover:

1. a definition of Event-Centric Knowledge Graphs (Section 1)
2. a formal semantic representation for ECKGs that includes reference to the original source (Section 3)
3. a method and open source tools for the extraction of Event-Centric Knowledge Graphs in four languages (Section 4)
4. four openly available ECKGs (Section 5)
5. a first assessment of the quality of automatically created ECKGs (Section 6).

The paper is organized as follows. In Section 2, we describe the background and related work. In Section 3, we describe how we modeled the information we extracted. In Section 4, we describe our processing pipeline. In Section 5, we describe our four use cases, namely general news, the FIFA world cup, and the global automotive industry, and news articles about the Airbus A380 in different languages. In Section 6, we report a first assessment of the accuracy of the ECKGs automatically created with our approach. In Section 7, we describe event-centric information access using the SynerScope tool, and report on additional applications and investigations enabled by our ECKGs. In Section 8, we discuss our approach and conclusions.

## 2. Background and related work

Knowledge Graphs (KGs) are used extensively to enhance the results provided by popular search engines (e.g. Google Knowledge Graph,[3] Microsoft's Satori[4]). These KGs are typically powered by

[2] Entities of type http://dbpedia.org/ontology/Event, in many cases corresponding to sports events or military conflicts.

[3] http://www.google.com/insidesearch/features/search/knowledge.html *Last accessed: 7 April 2015.*

[4] http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/ *Last accessed: 7 April 2015.*