Context-Aware Knowledge Graphs Exploratory Search

Veronica dos Santos[1], **Edward Hermann Haeusler**[1], **Daniel Schwabe**[2], **Sergio Lifschitz**[1]

[1]DI (PUC-Rio) - Rio de Janeiro - RJ

[2]BioBD (PUC-Rio)

{vdsantos, hermann, Sergio}@inf.puc-rio.br, dschwabe@gmail.com

Abstract. *The exploratory search approach recognizes that user queries can be incomplete, inaccurate, and ambiguous. This occurs both because of incomplete domain knowledge by the user or due to implicit assumptions about the context. This ongoing research aims to enrich Knowledge Graphs (KG) to support context-aware exploration through expanded queries. We propose a Contextual KG (CKG) definition and schema that characterizes the necessary elements for modeling contextual information and a query-answering approach that retrieves all (contextualized) possible answers.*

1. Introduction

The Internet is currently one of the primary sources of information for users and programs to obtain data that will inform decisions and actions being carried out, for example, in the medical and political domains, to name a few. Among the sources used, directly or indirectly, we can identify Knowledge Graphs (KG) as one of the primary providers of structured data (e.g., Wikidata[1]). The major platforms and service providers on the Web (e.g. Google, Microsoft, Apple, Amazon, Facebook, Linked-In, Spotify, etc..) all use KGs as back-ends to provide their services. KGs are Knowledge Bases (KBs) modeled as a graph [Weikum 2021] since relationships are the focus of analysis.

The user information needs are frequently unclear and well-defined at the outset, and users often need to learn what is present in a publicly available KG on the Internet. This entails employing exploratory search to acquire the desired knowledge [Marchionini 2006] and discovering knowledge gaps relevant to the task at hand. KGs are suitable for such complex searches [Weikum 2021]. Exploratory search approaches over KGs ultimately result in graph sub-pattern queries [Lissandrini et al. 2020b].

KGs encompass different types of knowledge, including Factual knowledge (statements representing claims of truth) and Contextual knowledge (statements claimed to be true within specific contexts) [Groth et al. 2023]. A common way to contextualize claims is by adding property-value pairs as qualifiers. We must distinguish between *additive qualifiers*, which represent n-ary relationships and do not affect the assessment of the fact's truthfulness, and *contextual qualifiers*, which can restrict the contexts in which the underlying fact is considered true and may modify the fact itself [Patel-Schneider 2018].

Given the abundance of multiple, distributed, and potentially contradictory sources available on the Internet, their veracity becomes prominent. In this situation, we adopt the *Dual Open World Assumption* (DOWA), a variant of the traditional Open

[1]https://www.wikidata.org/wiki/Wikidata:Main_Page

World Assumption (OWA). Under DOWA, the presence of a claim in a KG does not automatically imply that it is true. Instead, the evaluation of truthfulness depends upon the contexts of claims, represented by their *contextual qualifiers*, and on the tasks being carried out or intended (purpose). Consequently, when addressing an information need, users should always be mindful of whether the context of the information retrieved is consistent with the one they are interested in.

This ongoing research aims to further structure and query KGs to support exploratory searches. To help explain the proposed approach, we present a use case focusing on the Brazilian geopolitical History domain.[2] Figure 1 shows a partial view. Notice it has two disconnected components but they share implicit relationships associated with the temporal context that will be revealed in the query answer, see section 3.

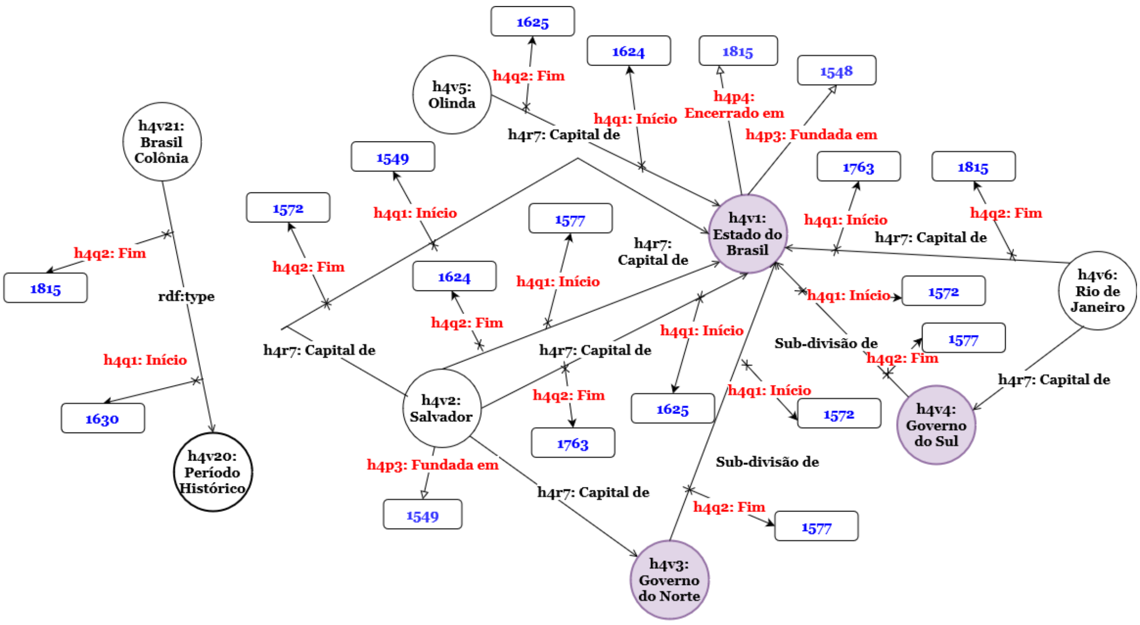

Figure 1. Temporal Contextualized Claims about Brazil

2. Contextual Knowledge Graphs

There are several proposals for KG data structures, some simpler, like RDF and LPG , and others more complex and abstract such as the multi-layer graph (graphs with higher-arity relationships and with identifiers on the edges) [Angles et al. 2022]. Table 1 shows a snippet of our KG modeled as a multi-layer graph H^3, with unique identified (column id) and qualified edges, using KGTK graph data model [Ilievski et al. 2020].

The model can be formally defined as follows (instances from the example follow each definition): V is a finite set of vertices (or nodes) that represents entities or concepts, $V = \{h4v1, h4v2, h4v3, h4v21\}$. R is a finite set of binary, directed or undirected relation types represented by their labels, $R = \{h4r6, h4r7\}$. E is a finite set of edges representing relationships based on a relation type R between two vertices from V, $E = \{e10, e20,$

[2]KG was constructed based on various websites that provide educational content for students.

[3]For the entire dataset, see `https://github.com/versant2612/CKG_UseCases/blob/53bb930d2a86d4a74f36cbb77c5c6c2bd7088aad/H4/CKG-H4.tsv`

$e30\}$. P is a finite set of property (or attribute) types, $P = \{h4q1, h4q2\}$, L is a finite set of vertices (or nodes) that represents literals corresponding to property values. Finally, pV is a finite set of edges representing relationships, based on a property P, between one vertice from V and another from L, $pV = \{p10, p20\}$ i.e. data values are in L. Key-value pairs can be attached to each E or pV in the form of Q (qualifier key) and V (vertex qualifier) or L (literal qualifier) as qualifications. These key-value qualifier pairs (qE and qP) allow differentiating instances of relationships. In table 1, $q301$, $q302$, $q201$ and $q202$ are examples of qualifications qE with $h4q1$ and $h4q2$ as Q and data values in L.

Such a definition can be used to represent any KG H which is composed by a KG schema, if it exists, and their instances. For this research, we refer to statements in a KG as claims instead of facts and adopt the DOWA. Based on the general definition of context stated in [Hogan et al. 2021], *"By context, we herein refer to the scope of truth, and thus talk about the context in which some data are held to be true"*, we define a CKG as:

Definition 2.1 (Contextual KG $\mathbb{H}$) $\mathbb{H} = H \cup C \cup I$, *that is, a multi-layer KG H with a set of entities, claims, and context information; the context mappings C between KG elements and context information; and interpretations I as rules to extract implicit context.*

Similarly to Contextualized Ontologies [Cafezeiro et al. 2008], in CKGs, mappings serve the purpose of establishing the role of each object, i.e., whether it functions as an entity in the "base KG" or as context information. The context specification C functions as a layer (in the sense of [Angles et al. 2022]) over the KG containing meta-information. All entities, claims, contexts, and mappings are components of the KG itself.

The KG engineer plays a crucial role to identify context information. When a KG schema is absent, KG profiling should be employed to extract latent structures from the KG instances. For each claim, he/she should identify if its R, P, and Q belong to any context C_i, specifying the corresponding mappings and adding them to the KG. The gray lines in table 1 correspond to a snippet of Temporal and Provenance mappings of qualifiers $h4q1$ (*inicio*), $h4q2$ (*fim*), and $h4q3$ (*fonte*) for relation type $h4r7$.

These mappings are specified as instances of the blue and green entities in the CKG conceptual schema (figure 2). The CKG $\mathbb{H}$ is enriched with explicit mappings that connect claims, entities, and their respective context information, facilitating effective contextualization and interpretation of the knowledge within the KG for decision making.

3. Possible Answers

User queries are typically incomplete, inaccurate, and/or ambiguous, often because crucial information, such as the context, is implicit or because users may not fully comprehend their underlying information needs within a given domain. Keyword-based search engines often assume that users are primarily interested in current or local information, implicitly adopting a specific default temporal or geographical context, which neglects so-called long tail scenarios. This highlights the importance of considering contextual information in exploratory search approaches to address such challenges.

Definition 3.1 (Possible Answer A) $A = \{S_1, S_2, \cdots, S_i\} | A \simeq \mathcal{Q}$ *, that is, A is composed of a set of zero or more fully contextualized claims S that potentially meet the user's information need. The possible answers are the result of a graph query $\mathcal{K}$ over $\mathbb{H}$ considering user's knowledge gaps and also KG and query incompleteness.*

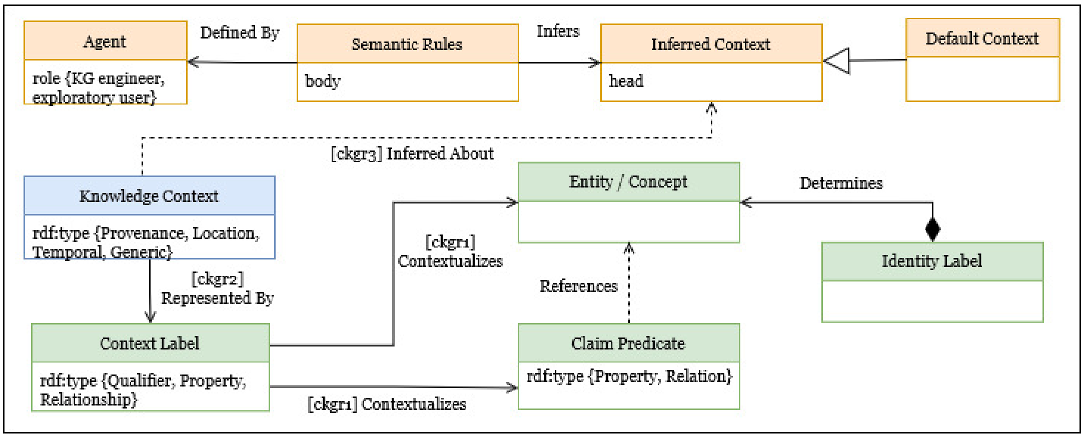

Figure 2. Contextual KG Schema

Table 1. Contextualized Claims & Context Mappings Sample

	id	node1	node1;label	label	label;label	node2	node2;label
	e1	h4v21	Brasil Colônia	is_a	rdf:type	h4v20	Período Histórico
	p10	h4v21	Brasil Colônia	h4q1	início	^1530	
	p11	h4v21	Brasil Colônia	h4q2	fim	^1815	
	e30	h4v2	Salvador	h4r7	capital de	h4v1	Estado do Brasil
H	q301	e30		h4q1	inicio	^1549	
	q302	e30		h4q2	fim	^1572	
	e20	h4v2	Salvador	h4r7	capital de	h4v1	Estado do Brasil
	q201	e20		h4q1	inicio	^1625	
	q202	e20		h4q2	fim	^1763	
	ckg:c1	ckgT1	Temporal	is_a	rdf:type	ckg:KnowledgeContext	
	ckg:c2	ckgT1	Temporal	ckgr1	ckg:Represented By	h4q1	início
C	ckg:c3	h4q1	início	ckgr2	ckg:Contextualizes	h4r7	capital de
	ckg:c4	ckgT1	Temporal	ckgr1	ckg:Represented By	h4q2	fim
	ckg:c5	h4q2	fim	ckgr2	ckg:Contextualizes	h4r7	capital de
	ckg:c29	ckgP1	Provenance	is_a	rdf:type	ckg:KnowledgeContext	
	ckg:c30	ckgP1	Provenance	ckgr1	ckg:Represented By	h4q3	fonte
	ckg:c34	h4q3	fonte	ckgr2	ckg:Contextualizes	h4r7	capital de

Consider a user who is interested in the capital cities of Brazil during the colonial period. S/he issues a search query that is translated by the search interface into the graph query $K3$ as illustrated in table 2 using graph query language Kypher, based on Cypher[4] [Ilievski et al. 2020]. Graph queries formulated during exploration can be both complete and incomplete with respect to context[5]. The degree of incompleteness can be assessed by executing graph queries against the Context Layer C in the CKG.

In order to verify if K is incomplete the query engine (figure 3) must execute two types of graph queries. In step $B1$, the query engine evaluates $K3$ completeness using query $ck1$ over the predicate $h4r7$ *(capital de)* where $?C\_label$ can be Temporal, Location, Provenance, Generic or any other context type that the KG engineer added. Predicates $ckgr1$ and $ckgr2$ correspond to *Contextualizes* and *Represented By* relationships as in the CKG schema. The relation type $h4r7$ has three context qualifiers, two temporal, $h4q1$ and $h4q2$, and one for provenance, $h4q3$ *(fonte)*. Query $ck2$, executed in step $B2$ for entity type $h4v20$ *(Periodo Historico)*, retrieves two Temporal properties, $h4q1$ (*incio*) and $h4q2$

[4] Cypher uses ASCII-art stile to represent sub-graph patterns: $(node1) - [: connection] \rightarrow (node2)$

[5] More examples using Provenance and Location contexts, can be found at `https://github.com/versant2612/CKG_UseCases/blob/53bb930d2a86d4a74f36cbb77c5c6c2bd7088aad/H4/script_kgtk_H4.sh`

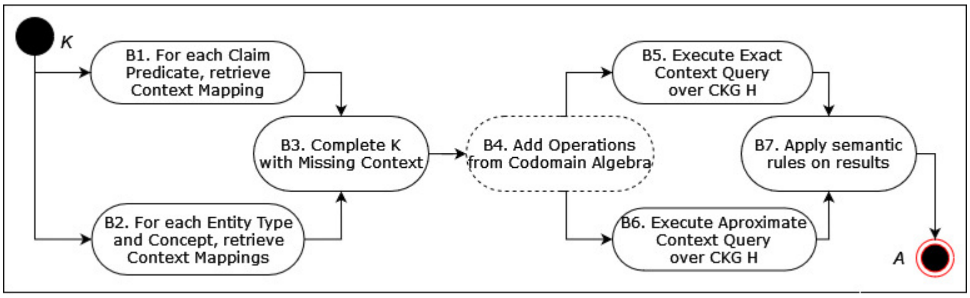

Figure 3. Query Engine

Table 2. Contextual query expansion for incomplete query K3

ID	Query {kgtk –debug query -i $GRAPH_CKG_H4 –as h4 \}
K3	–force –match 'h4: (v1)-[p1:h4r7]→(v2), (:h4v21)-[p2:is_a]→(:h4v20)'
cK1	–match 'h4: (C {label: C_label})-[p2:ckgr1]→(c1)-[p1:ckgr2]→(:h4r7)
cK2	–match 'h4: (C {label: C_label})-[p2:ckgr1]→(c1)-[p1:ckgr2]→(:h4v20)
K3e	–match 'h4: (v1)-[p1:h4r7]→(v2), (p1)-[q1:h4q1]→(v3), (p1)-[q2:h4q2]→(v4), (p1)-[q3:h4q3]→(v5), (:h4v21)-[p2:is_a]→(:h4v20), (:h4v21)-[p3:h4q1]→(v6), (:h4v21)-[p4:h4q2]→(v7), (p2)-[q4:h4q3]→(v8)' –where **'v3 <v7 AND v6 <v4'**

(*fim*), and a provenance qualifier, $h4q3$ *(fonte)*.

The original $K3$ and its contextually expanded version specifies a pattern with two disconnected sub-graphs. In such cases, any Codomain Algebra can be applied ($B4$) to context values (e.g., Dates, Geometries, Integers, etc) to infer relationships not directly materialized in the KG, such as claims co-occurence in time or entities overlapping in space or the ordering of information sources based on ranking. Considering that the two parts of the sub-graph pattern have temporal context, the query expansion added the time-overlap operation in the –WHERE clause. This enables additional insights and relationships in the analysis of context values, enriching the answer with implicit knowledge and providing further context-aware capabilities for exploratory search.

The query engine generates Exact and Approximate answers. A new query $K3e$ is formulated in step $B5$ to retrieve connected and fully qualified claims S for an exact answer. Another query can be formulated in step $B6$ to retrieve incomplete qualified claims by using the –OPTIONAL parameter for context added. These queries aim to provide flexibility in retrieving answers based on varying levels of query and KG completeness, being more cooperative and less interactive and iterative.

4. Final Remarks

Regarding to exploratory search applications various solutions have been proposed, focusing on approximate methods, query suggestion, and query refinement techniques. TriniT is a exploratory querying system [Yahya et al. 2016] that addressed vocabulary mismatch using rules for query relaxation and KG incompleteness treated through triples addition in query time (eXtended KG). Another approach found in the literature involves interactive query expansion [Lissandrini et al. 2020a]. This approach utilizes sample query results as input and generates the *k* most relevant expansions to complement the original query, based on element labels, similar to how language models expand keywords. The system collects user feedback to improve the accuracy of the expansions over time.

The main novelties of our approach are: first, we assume the KG supports a decision process by the user where s/he will decide which information should be used for its intended purposes. From this point of view, we seek to provide *Possible Answers* considering the available context information as a way to more fully support this process. A second aspect is the handling of the incompleteness of the KG itself. We apply an answer expansion approach taking into account all relevant context information to support the interpretation of the claims. Thus, the answers provided are comprehensive and contextually relevant. Lastly, our approach does not rely on an interactive flow with the user. Instead, it is designed as a stateless approach where the *The Best Possible Answer* is determined based on the available context and query flexibility options. This approach allows for efficient and flexible exploration without requiring constant user input.

Currently, we are evaluating our approach using Wikidata datasets and developing use cases of Context Interpretations I. And as future work, in addition to Codomain Algebra, we will evaluate how Context Algebra [Cafezeiro et al. 2008] can be used for Knowledge Graph Engineering.

References

Angles, R., Hogan, A., Lassila, O., et al. (2022). Multilayer graphs: a unified data model for graph databases. In *GRADES-NDA*, pages 11:1–11:6.

Cafezeiro, I., Haeusler, E. H., and Rademaker, A. (2008). Ontology and context. In *IEEE Int. Conf. on Pervasive Computing and Communications (PerCom)*, pages 417–422.

Groth, P., Simperl, E., et al. (2023). Knowledge Graphs and their Role in the Knowledge Engineering of the 21st Century. *Dagstuhl Reports*, 12(9):60–120.

Hogan, A., Blomqvist, E., et al. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4).

Ilievski, F., Garijo, D., et al. (2020). KGTK: A toolkit for large knowledge graph manipulation and analysis. In *ISWC*, pages 278–293. Springer.

Lissandrini, M., Mottin, D., Palpanas, T., and Velegrakis, Y. (2020a). *Graph-Query Suggestions for Knowledge Graph Exploration*, page 2549–2555. ACM.

Lissandrini, M., Pedersen, T. B., Hose, K., and Mottin, D. (2020b). Knowledge graph exploration: Where are we and where are we going? *SIGWEB Newsl.*

Marchionini, G. (2006). Exploratory search: From finding to understanding. *CACM*, 49(4):41–46.

Patel-Schneider, P. F. (2018). Contextualization via qualifiers. In *Joint Proc. of the Int. Workshops on Contextualized Knowledge Graphs, and Semantic Statistics co-located with (ISWC 2018)*, CEUR Workshop Proceedings. CEUR-WS.org.

Weikum, G. (2021). Knowledge graphs 2021: A data odyssey. *PVLDB*, 14(12):3233–3238.

Yahya, M., Berberich, K., Ramanath, M., and Weikum, G. (2016). Exploratory querying of extended knowledge graphs. *Proc. VLDB Endow.*, 9(13):1521–1524.