Building A Knowledge Graph for Audit Information

Naser Ahmadi[1], Hansjorg Sand[2] and Paolo Papotti[1]

[1]*EURECOM, France*
[2]*KPMG, Germany*

Abstract
We present our insights from the experience of creating a knowledge graph (KG) for the auditing domain. We discuss the main challenges in building such KG starting from text and unstructured data and present an overview of our solution. The proposed approach follows a standard pipeline when it first extracts entities from auditing documents and then finds relationships among them. However, the process is especially challenging because auditing entities are in most cases non-named entities, which are hard to model in the graph and to identify in text. From our experience, we finally derive a set of observations on the limits of automatic methods for the construction of audit KGs and a possible direction to address them.

Keywords
knowledge graph, auditing, text, taxonomy, structured data

1. Introduction

A *Knowledge Graph* (KG) is a structured representation of information which stores real-world entities as nodes, and relationships between them as edges. KGs represent data with large collections of interconnected entities. Usually, types (classes) describe the entities (e.g., entity *Paris* is a *city*, *France* is a *country*), while predicates describe their relationships (a city *isCapital* of a country) and their properties (France has a *population:62M*). RDF KGs organize information in the form of triples with a *predicate* expressing a binary relation between a *subject* and an *object.* KGs store large amounts of triples, or *facts*, e.g., the English version of DBpedia stores 850 million facts. The syntactic and semantic structures of knowledge in KGs are useful in building applications, such as Question Answering [1, 2] and Semantic Search [3].

Manually building a KG is a very expensive process. For this reason, research has been conducted on KG creation both in academia [4, 5, 6, 7, 8] and in the industry [9, 10]. However, when applied on the textual documents in the financial domain, these methods fail short. Indeed the KGs for legal and audit enterprises are very different from Wikipedia pages. While most of the KGs in the literature are *encyclopedic*, covering objects and facts in the real world, some enterprises may have information which is mostly composed of non-named entities and abstract topics, making it close to a *commonsense* KG. See examples that highlight the difference in Figure 1. The latter category is much harder to build automatically, and most efforts rely on humans, usually in a crowdsourcing fashion, such as ConceptNet [11] and ATOMIC [12]. The specific and technical domain of an enterprise content is one of the biggest challenges in creating financial KGs [13], in general, and an audit KG in our setting.

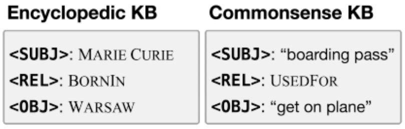


Figure 1: Examples of knowledge triples from encyclopedic and commonsense KGs [14].

External commonsense resources, such as ConceptNet, are used in some of the relevant methods, but they are not a direct solution to the KG construction problem. Many terms are domain-specific, so they are either missing from the existing resource or their modeling in the commonsense KGs does not match the level of details that is needed in the enterprise setting. For example, in an accounting dictionary *AIM* stands for *Alternative Investment Market* and *goodwill* is "a type of tangible assets that occurs when a buyer acquires an existing business", while these words have very different meanings in a general dictionary. We remark also the challenge in modeling the above definition of *goodwill* by using non-named entities in the KG, what are the right noun phrases to add? Can the properties expressed in the sentence be represented with binary relationships?

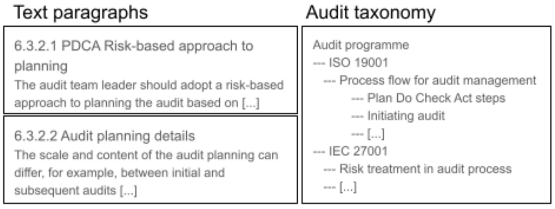
In our work, we are developing tools for automating different parts of a framework for continuous creation and curation of KGs. However, we face a lot of challenges that make the automatic creation of such data structures much harder than in other settings. We start with an example of a KG we are creating in our collaboration with KPMG and then explain the difficulties and the opportunities in building an audit KG.

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
naser.ahmadi@eurecom.fr (N. Ahmadi); hsand@kpmg.com (H. Sand); papotti@eurecom.fr (P. Papotti)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Audit Knowledge Graph

We introduce a very high-level KG based on node entities and only two kinds of relationships between entities. This KG is different from traditional entity-centric knowledge graphs and it is motivated by text data and taxonomies that are available in the KPMG corpus of textual documents. The design of the KG is done also according to target applications.



Text paragraphs

6.3.2.1 PDCA Risk-based approach to planning
The audit team leader should adopt a risk-based approach to planning the audit based on [...]

6.3.2.2 Audit planning details
The scale and content of the audit planning can differ, for example, between initial and subsequent audits [...]

Audit taxonomy

```
Audit programme
--- ISO 19001
   --- Process flow for audit management
          --- Plan Do Check Act steps
          --- Initiating audit
          --- [...]
--- IEC 27001
   --- Risk treatment in audit process
   --- [...]
```

Figure 2: An example of KPMG's documents (left) and an audit taxonomy (right).

Figure 2 shows a sample with a few sentences from two documents (left) and a fragment of a taxonomy for the audit process (right). In our corpus there are thousands of documents with variable size, from very short (only one sentence) to quite large documents with dozens of paragraphs. For the taxonomies, they can vary in size but are in the order of hundred nodes, each composed of a short sentence. These can be considered the starting point of the KG construction and from those several other nodes are derived.

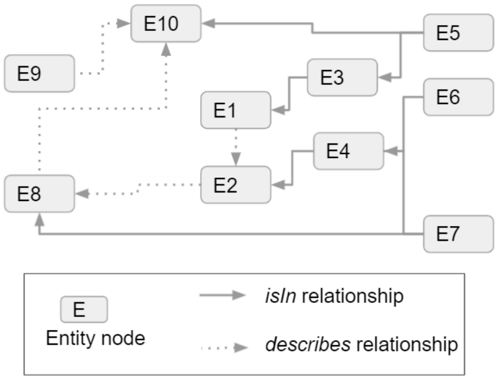


Figure 3: Generic KG with one node type and two kinds of relationships.

In Figure 3, there is only one kind of node, representing entities. Those are very generic texts, they can be single words, paragraphs or long documents. The relationships across them are represented by directed edges and the nodes are connected in many to many relationships. We consider two kinds of relationships. The first one is the containment, in the example **E6** *is contained in* **E4**. This could be a word contained in a document, for example, or a sub-element in a hierarchy (e.g., the relation between *IEC 27001* and *Audit process* in the hierarchy in Figure 2). Also, **E2** could be a topic that *describes* document **E8**. We remark that all manually defined edges are given the same weight with value 1, but in the KG edges can be weighted with a value between 0 and 1 for uncertain relationships (according to the confidence given by an automatic tool, for example).

The above example representation is very generic and simplified, we introduce it to give a feeling of the kind of graph that we are interested in. However, in our deployed KG, the nodes are of six different types:

- **Documents** nodes are (possibly long) texts containing one to multiple paragraphs. For example, in Figure 2 two paragraphs are shown on the left side; those correspond to two D nodes.

- **Taxonomy** nodes are auditing concepts following a hierarchical structure. For example, every process step can be represented as a path from the root node to the leaf, e.g., *Audit programme* → *ISO 19001* → *Initial audit*.

- **Caption** nodes are client-specific short documents that are described by taxonomy nodes, i.e., a *describes* edge goes from a *taxonomy* node to a *caption* node.

- **Topics** nodes are terms with one or multiple related entities; e.g., "risk treatment" and "audit process" are topics in the *describes* relationship with the *Risk treatment in audit process* step. Entities are associated in an *isIn* relationship with a topic.

- **Entities** nodes contain n-gram terms that are representative of relevant items, names and concepts in the audit domain. Every entity is the representative for a *family* of words, where a family includes (with *isIn* relationships) synonyms and abbreviations that can be used to express such entity in documents.

- **Word** nodes are words in an entity, their synonyms or other variations. E.g., *auditing*, *adt* and *prc* are words for entity *audit process*.

There are two main design choices behind our representation.

First, we use several node types and very few relationship types, as the latter are harder to extract automatically from text. We found that NLP analysis of the text can identify the two (relatively simple from a semantic viewpoint) relationships, while for the entity types the task is simplified by the awareness of their provenance, i.e., some types that can be mostly derived from the source of extraction. However, obtaining such types and relationships automatically from text documents is a difficult task, as we discuss in the next section.

Second, some node types are inspired by the target users. The proposed representation has been validated by experts and it is used for one text matching application at the firm. This application exploits the rich granularity of the text representation in the KG. Indeed, the different types enable the immediate characterization of a new text, say a customer document, in terms of entities (with entity and word nodes) and more abstract concepts (set of entities). We found this freedom crucial given the challenge of fixing the right abstraction for the expression of non-named entities in the KG.

3. Limits and Opportunities of Automatic Methods

Given the nature of the auditing content, automatic methods for encyclopedic KG construction are not very effective [15, 16, 17]. We experimented largely with such methods, but with results that were far away from the required quality [18]. We list five main challenges. (1) Auditing entities are not standard named entities, such as *France* and *IBM*. (2) Non-named entities are expressed as noun phrases that can be recognized as subject in sentences but are hard to organize in a structured graph. For example, “tangible asset" should be modeled with one or two entities? (3) Most of these entities are oftentimes used in the form of acronyms or abbreviations. (4) Taking in account the richness of human language, there are many variations of noun phrases in expressing the same concept. (5) There is no training data in this domain, and general corpora miss the subtle differences in the audit domain [19, 15]. While some of these challenges apply in general for KG construction, we found that these problems are especially hard for existing tools in this setting.

As the project moved forward, different parts of the KG have been manually defined by the domain experts at KPMG. For example, a list of potential entities has been identified with NLP traditional tools and then manually revised by a human team. This process had identified some of the opportunities to introduce automatic methods to help in the KG construction. Moreover, the manually crafted portions of the KG offered us some ground truth for the evaluation of the proposed algorithms [20].

In our pipeline, the first task is the automatic identification of nodes and the second task is the identification of relationships across the different nodes. We first tackle the task of generating the *entity* nodes, or key short phrases, that act as subjects and objects. Starting from those, we generate *families* of words for each entity node. The goal is to find a group of semantically equivalent *words*, including abbreviations and acronyms, and to associated them to the representative entity given only the documents [20]. Words and representative entities are related with *isIn* relationships. When evaluated against the ground truth written by the experts, we found that the proposed unsupervised technique for mapping words and entities can achieve high precision, but only limited recall, with the latter varying between 0.55 and 0.4 depending on the language at hand, i.e., English is easier than German [20].

We then propose a method to identify relationships of type *describes* between nodes, and we conduct experimental campaigns on the discovery of relations between documents and taxonomy nodes [21]. Our method exploits a deep learning approach for the unsupervised modeling of the entities as vectors in the presence of free text and structured data [22]. Such vectors are then used in the unsupervised matching step. In particular, we report promising results in matching documents and taxonomy nodes, which is a challenging task for existing methods because of the long textual content in our entities. Compared to the manually created relationships, the unsupervised method obtains 0.6 F-measure when looking at top-3 matches [21].

While our initial results are promising, we need better methods that involve the experts in the KG building process with simple interfaces [23, 24]. The design of human-in-the-loop solutions is at the core of our current efforts. The knowledge graphs with the human-in-the-loop solutions we work on will support a broad range of scenarios in financial and economic settings:

- Automated classification of financial records in data ingestion and analysis pipelines.
- Automated classification of financial transaction documents to support automated transaction processing.
- Automated metadata tagging for documents and sub-documents in legal and accounting corpora to improve the reliability of semantics search engines.

References

[1] C. Unger, A. Freitas, P. Cimiano, An introduction to question answering over linked data, in: Reasoning Web International Summer School, Springer, 2014, pp. 100–140.

[2] D. Diefenbach, V. Lopez, K. Singh, P. Maret, Core techniques of question answering systems over knowledge bases: a survey, Knowledge and Information systems 55 (2018) 529–569.

[3] H. Bast, B. Björn, E. Haussmann, Semantic search on text and knowledge bases, Foundations and Trends in Information Retrieval 10 (2016) 119–271.

[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, T. M. Mitchell, Toward an architecture for never-ending language learning., in: AAAI, 2010, pp. 1306–1313.

[5] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia-A crystallization point for the web of data, Web Semantics 7 (2009) 154–165.

[6] F. M. Suchanek, G. Kasneci, G. Weikum, YAGO: A core of semantic knowledge unifying wordnet and wikipedia, in: WWW, 2007, pp. 697–706.

[7] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Comm. of the ACM 57 (2014) 78–85.

[8] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, Y. Ye, KATARA: a data cleaning system powered by knowledge bases and crowdsourcing, in: SIGMOD, 2015.

[9] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, W. Zhang, From data fusion to knowledge fusion, PVLDB 7 (2014) 881–892.

[10] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, A. Doan, Building, maintaining, and using knowledge bases: a report from the trenches, in: SIGMOD, 2013, pp. 1209–1220.

[11] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[12] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, ATOMIC: an atlas of machine commonsense for if-then reasoning, in: AAAI, AAAI Press, 2019, pp. 3027–3035.

[13] S. Elhammadi, L. V.S. Lakshmanan, R. Ng, M. Simpson, B. Huai, Z. Wang, L. Wang, A high precision pipeline for financial knowledge graph construction, in: COLING, 2020, pp. 967–977.

[14] T. Safavi, D. Koutra, Relational world knowledge representation in contextual language models: A review, arXiv preprint arXiv:2104.05837 (2021).

[15] M. Kejriwal, Domain-specific knowledge graph construction, Springer, 2019.

[16] M. Kejriwal, R. Shao, P. Szekely, Expert-guided entity extraction using expressive rules, in: SIGIR, 2019, pp. 1353–1356.

[17] B. Abu-Salih, Domain-specific knowledge graphs: A survey, Journal of Network and Computer Applications 185 (2021) 103076.

[18] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, P. Levis, C. Ré, Fonduer: Knowledge base construction from richly formatted data, in: SIGMOD, ACM, 2018, pp. 1301–1316.

[19] N. Jain, Domain-specific knowledge graph construction for semantic analysis, in: European Semantic Web Conference, Springer, 2020, pp. 250–260.

[20] N. Ahmadi, A framework for the continuous curation of a knowledge base system, Ph.D. thesis, 2021. EURECOM.

[21] N. Ahmadi, H. Sand, P. Papotti, Unsupervised matching of data and text, in: ICDE, IEEE, 2022.

[22] R. Cappuzzo, P. Papotti, S. Thirumuruganathan, Creating embeddings of heterogeneous relational datasets for data integration tasks, in: SIGMOD, 2020.

[23] S. Zhang, L. He, E. C. Dragut, S. Vucetic, How to invest my time: Lessons from human-in-the-loop entity extraction, in: SIGKDD, ACM, 2019, pp. 2305–2313.

[24] P. Ristoski, A. L. Gentile, A. Alba, D. Gruhl, S. Welch, Large-scale relation extraction from web documents and knowledge graphs with human-in-the-loop, J. Web Semant. 60 (2020).