Towards Knowledge Graph Based Services in Accounting Use Cases

Michael Schulze[1,2], Michelle Pelzer[3], Markus Schröder[1,2], Christian Jilek[1,2], Heiko Maus[2] and Andreas Dengel[1,2]

[1] *Computer Science Department, Technische Universität Kaiserslautern, Germany*
[2] *Smart Data & Knowledge Services Department, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)*
[3] *envia Mitteldeutsche Energie AG, Chemnitz, Germany*

Abstract

To assist knowledge work in accounting use cases such as bookkeeping, this paper presents a pipeline for constructing and enriching an accounting knowledge graph from heterogeneous accounting resources. To show the feasibility of the approach, we applied the pipeline in a multi-group energy provider by employing real company data. A set of prototypical knowledge services was realized with the accounting knowledge graph as the basis, for example, the suggestion of similar accounting cases to the accountant. For training decision trees to predict accounts, our results suggest that using semantically enriched data from the knowledge graph leads to better results compared to not using semantically enriched data.

1. Introduction

The work of accountants is characterized by searching relevant information which is distributed in different company sources. One typical task is bookkeeping where besides verifying an invoice, decisions are made about the set of accounts and VAT rates that need to be selected for a particular accounting case. For solving such tasks, accountants typically tap into several data sources, for example, chart of accounts, accounting manuals and handbooks, or accounting policies. In multi-group companies, accounting departments are usually responsible for all subsidiary groups which means that they have to consider such documents for each group separately. Further common sources are personal notes, referenced documents on an invoice, historical accounting cases as references, additional attachments, or the knowledge of a colleague.

Tapping into this many data sources can lead to cases where the time needed to find the relevant information is high, especially for non-standard cases. Jain and Woodcock [1] predict that also in future, 21 % of tasks that deal with invoice processing are not suited for automation. Therefore, we conclude that knowledge workers will be still important for such tasks. With our research, we aim to assist knowledge workers in such settings with an "Information Butler" [2] paradigm emphasizing the importance of context in which a knowledge worker is situated [3]. Also, by following this idea and adopting it to accounting scenarios, our goal is to embed knowledge graph-based services directly into the work context of an accountant, for example, when the accountant opens up an invoice for processing. Because we cooperated with a multi-group energy provider, we were able to base this research on real company data. There, the

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria

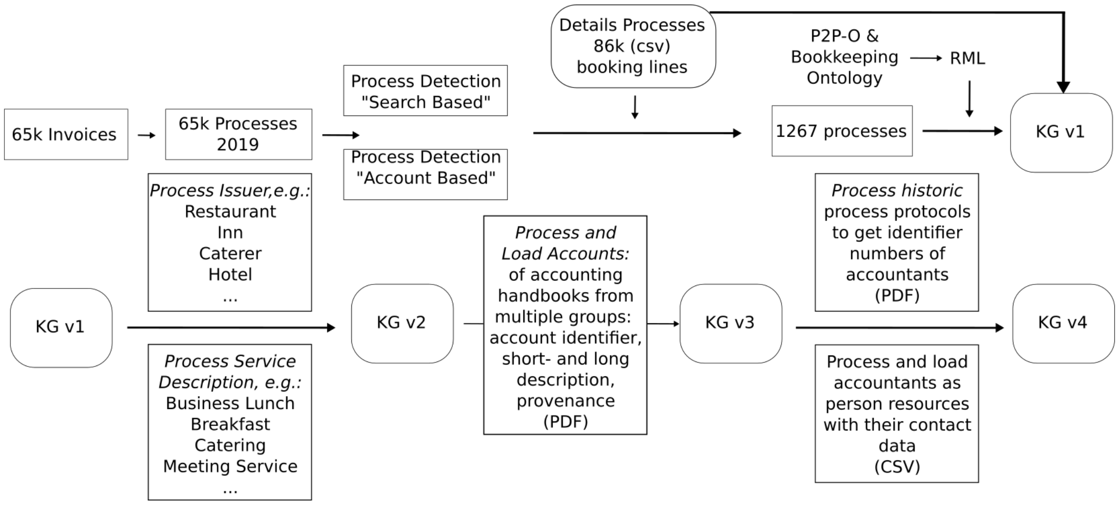

Figure 1: Pipeline of constructing and enriching an accounting knowledge graph for bookkeeping use cases. KG v1 to KG v4 depict the steps of semantic enrichment

accounting department was interviewed to derive requirements from a knowledge worker's view: a) a suitable approach should cover food and beverage scenarios in particular, such as eating at restaurants or catering, because such cases are, compared to other scenarios, particularly error-prone, b) historic accounting cases should be suggested as an assistance for using them as references; c) accounts and VAT rates that may fit to the current work context should be displayed with detailed information; and d) suitable co-workers that may have experience in the current accounting case should be suggested.

2. Related Work

Use cases around the Financial Industry Business Ontology (FIBO) [4] deal with integrating financial information and providing knowledge services on top of it. However, reasonably, such services often have a strong and useful analytical rationale to help with decision making. Work and knowledge graph based services in the context of public procurement [5] aim to assist the knowledge worker. Compared to accounting, the procurement domain is more upstream but it may be useful to interlink such data for downstream tasks.

3. Approach

Because the envisioned knowledge services require upfront queries over information that resides in the different introduced data sources, which have different formats as well such as PDF or CSV, and because the work context of an accountant is important but nowhere modeled and displayed in the task process explicitly, we decided to first construct an accounting knowledge graph and then to implement the knowledge services on top of it.

3.1. Accounting Knowledge Graph Construction and Vocabulary

Figure 1 depicts the pipeline of constructing and enriching an accounting knowledge graph. The data basis was a CSV file containing 65k processes from the year 2019. It contained information such as the invoice issuer, invoice issuer number, booking area, the service description, or the amount of money. According to requirement a), we extracted at first all processes that deal with food and beverage scenarios. For this, two approaches were followed: On the one hand, with a keyword list, we mimicked the approach an accountant would use by searching the service descriptions of invoices for keywords such as "secretariat service" or "restaurant". On the other hand, we filtered processes which were booked on standard food and beverage accounts where food and beverage cases are booked exclusively. By merging the results, 1267 food and beverage processes have been detected which represent the first input for the knowledge graph.

To construct the first version of the knowledge graph (KG v1 in Figure 1), another CSV file containing a set of booking statements for each process has been considered. Such booking statements contain, among others, an account, a VAT rate identifier and the accordingly split amount of money. Technology-wise, we used RML [6] with rmlmapper[1] in version 5.0.0. Ontology-wise, we used the P2P-O ontology [7] as a basis where vocabulary is provided for describing electronic invoices, such as the invoice issuer (seller). To account for the need to semantically describe and connect booking statements and the content of accounting handbooks, we developed a bookkeeping ontology which aims to extend P2P-O. To the best of our knowledge, there is no ontology available covering such bookkeeping vocabulary, and thus fitting our exact requirements[2]. However, we could reuse vocabulary from the FIBO Ontology [4], such as for the ledger account and account identifier. The bookkeeping-ontology is documented with WIDOCO [8] and published at https://purl.org/p2p-o/bkk. It further has a business friendly license to enable reuse. For validation, we checked the consistency and iteratively evaluated the ontology with OOPS! (OntOlogy Pitfall Scanner!) [9].

3.2. Semantic Enrichment

For the second version of the knowledge graph (KG v2 in Figure 1), the type of invoice issuer was refined as well as the service description. By means of keyword matching, triples to the invoice issuer have been automatically added which state whether the issuer is, for example, a restaurant, a hotel, or a caterer. In the same way, triples have been added to the service description to specify whether the invoice is, for example, about a breakfast, a business lunch, or a meeting service. For a full list of enrichments, we would like to refer to the ontology page. To reach KG v3 in Figure 1, we incorporated information about accounts which resides in multiple PDF handbooks for each company group separately. Accordingly, per account, information about the account ID, short- and long descriptions as well as the provenance in form of the accounting handbook were incorporated into the knowledge graph. Because the accounting handbooks were all structured identically starting with the account identifier per account, and by exploiting this structure with regular expressions, the handbook text could be processed automatically. To enable the suggestions of colleagues (Figure 1, KG v4), we firstly included all

[1] https://github.com/RMLio/rmlmapper-java
[2] By searching in ontology repositories such as https://lov.linkeddata.es/dataset/lov/

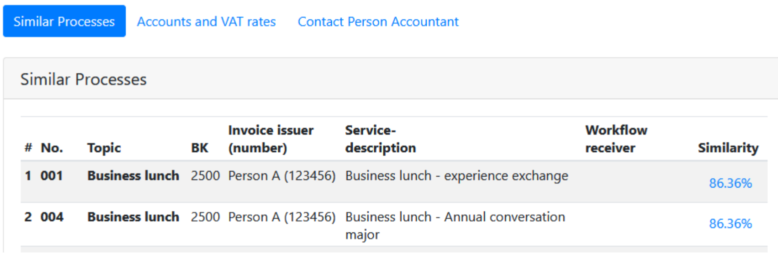

Figure 2: Excerpt of suggesting similar processes (synthetic data for the screenshot).

Table 1
Results of learning decision trees to predict accounts by using different degrees of semantically enriched data. I: with type of invoice issuer, II: with type of service, III: with the combination of both types. Numbers are weighted averages over all predicted classes obtained from WEKA [10].

Data	TP Rate	FP Rate	Precision	Recall	F-Measure
without semantically enriched data	.630	.271	.719	.630	**.587**
semantically enriched data I	.634	.225	.631	.634	**.626**
semantically enriched data II	.668	.224	.672	.668	**.649**
semantically enriched data III	.748	.148	.757	.748	**.739**

accountants as person resources with their contact data. For linking accountants to historical processes, historic process protocols have been analyzed to find out which accountants were involved in a process (also by employing regular expressions).

4. Knowledge Graph Based Services and Prediction of Accounts

Because the goal was to embed knowledge services directly into the workspace of an accountant, information is shown in form of a sidebar on the desktop. The screenshot in Figure 2 shows an excerpt of the suggestions of similar accounts. In this way, an accountant can look into such processes to get clues for processing the current accounting case. For calculating the similarity between processes, we leveraged the semantically enriched information about types of invoice issuer and services. An insight from interviews was that the similarity indicators needed to be weighted differently so that the booking area is more important than the invoice issuer and the invoice issuer is more important than the type of service. Therefore, initially, a ratio of 4:2:1:1 was adopted (where the additional "1" represents the type of invoice issuer). Because during enrichment accountants have been linked to historic processes, the similarity matrix could also be used for suggesting colleagues. For account and VAT rate prediction, two approaches have been followed: First, based on the similarity matrix, we suggest the accounts and VAT rates of similar processes. Second, Table 1 summaries results for learning decision trees by using different degrees of semantically enriched data: using non-enriched data and using semantically enriched data with (I) the type of the invoice issuer, (II) the type of the service, and (III) with the combination of both. WEKA [10] was used because it supports categorical data [10], and results are obtained by employing 10-fold cross validation. First results in Table 1 suggest that adding and in particular combining a few features obtained from a semantically enriched knowledge graph can lead to better results.

5. Conclusion

In this paper, we presented a pipeline for constructing and enriching an accounting knowledge graph for bookkeeping use cases. To enable reuse, we further proposed required bookkeeping vocabulary. Besides showing the feasibility of realizing knowledge services on top of such an accounting knowledge graph, first results indicate that using semantically enriched data leads to better results when learning decision trees to predict accounts. One lane for future work is the offering of explanations in form of derived rules.

Acknowledgements: This work was funded by BMBF project SensAI (grantno. 01IW20007).

References

[1] K. Jain, E. Woodcock, A road map for digitizing source-to-pay, 2017. URL: https://www.mckinsey.com/business-functions/operations/our-insights/a-road-map-for-digitizing-source-to-pay, Last accessed 12 Jul 2022.

[2] A. Dengel, H. Maus, Personalisierte Wissensdienste: Das Unternehmen denkt mit, IM+io Fachmagazin 3 (2018) 46–49.

[3] C. Jilek, M. Schröder, S. Schwarz, H. Maus, A. Dengel, Context spaces as the cornerstone of a near-transparent and self-reorganizing semantic desktop, in: The Semantic Web: ESWC 2018 Satellite Events, volume 11155 of *LNCS*, Springer, 2018, pp. 89–94.

[4] M. Bennett, The financial industry business ontology: Best practice for big data, Journal of Banking Regulation 14 (2013) 255–268.

[5] A. Soylu, Ó. Corcho, B. Elvesæter, C. Badenes-Olmedo, T. Blount, F. Y. Martínez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, T. C. Lech, D. Roman, Theybuyforyou platform and knowledge graph: Expanding horizons in public procurement with open linked data, Semantic Web 13 (2022) 265–291.

[6] A. Dimou, M. V. Sande, P. Colpaert, R. Verborgh, E. Mannens, R. V. de Walle, Rml: A generic language for integrated rdf mappings of heterogeneous data, in: Proc. of the Workshop on Linked Data on the Web, volume 1184 of *CEUR Workshop Proc.*, CEUR-WS.org, 2014.

[7] M. Schulze, M. Schröder, C. Jilek, T. Albers, H. Maus, A. Dengel, P2P-O: A purchase-to-pay ontology for enabling semantic invoices, in: The Semantic Web - 18th International Conference, ESWC 2021, volume 12731 of *LNCS*, Springer, 2021, pp. 647–663.

[8] D. Garijo, Widoco: a wizard for documenting ontologies, in: International Semantic Web Conference, Springer, Cham, 2017, pp. 94–102.

[9] M. Poveda-Villalón, A. Gómez-Pérez, M. C. Suárez-Figueroa, Oops! (ontology pitfall scanner!): An on-line tool for ontology evaluation, International Journal on Semantic Web and Information Systems (IJSWIS) 10 (2014) 7–34.

[10] E. Frank, M. A. Hall, I. H. Witten, The WEKA workbench., in: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition, Morgan Kaufmann, 2016.