ptpDG: A Purchase-To-Pay Dataset Generator for Evaluating Knowledge-Graph-Based Services

Michael Schulze[1,2(✉)], Markus Schröder[1,2], Christian Jilek[1,2], 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), Kaiserslautern, Germany
{firstname.lastname}@dfki.de

Abstract. This paper introduces ptpDG, a labeled-dataset generator that generates various data assets for evaluating knowledge graph construction approaches and downstream knowledge services in the purchase-to-pay domain: While organizations sell, purchase and complain about products in a multi-agent-system simulation, a ground truth knowledge graph emerges with different kinds of purchase-to-pay processes. Based on this knowledge graph, heterogeneous electronic purchase-to-pay documents such as e-invoices, credit notes and orders are generated. To those documents, noise patterns are added that we have frequently encountered in real industrial data. Finally, a provenance graph is generated which contains provenance information between document elements and ground truth triples. In this way, for such privacy sensitive scenarios, ptpDG enables data-driven evaluation and its publication.

Keywords: Knowledge Graph Construction · Evaluation · Simulation.

1 Introduction and Motivation

Purchase-to-pay processes are "knowledge-intensive processes" [2] consisting of heterogeneous documents such as orders, e-invoices and credit notes. To support knowledge workers in such work environments, our research is concerned with knowledge-graph-based services for users[3]. For such services, knowledge graphs have to be constructed in the first place which we also want to evaluate in a data-driven way. However, publication of real industrial data for scientific evaluation is rarely possible because this kind of data is often highly sensitive. This also holds for information contained in real purchase-to-pay documents because it consists of personal information and relates to third parties. In our experience, industry partners also have objections to anonymization techniques because risk of de-anonymization exists [4]. Even in the rare cases when data publishing may

[3] https://comem.ai/SensAI

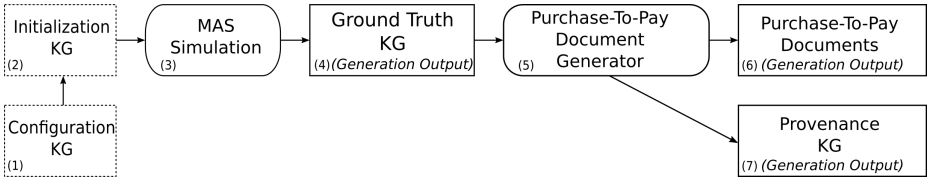

Fig. 1. General approach of ptpDG

be possible, or when it is not aspired at all, it is still a time consuming task to label such data [5].

Therefore, this paper introduces ptpDG, an approach that generates various data assets for evaluating knowledge graph construction approaches and downstream knowledge services: a), synthetic electronic purchase-to-pay documents such as e-invoices, credit notes or orders where noise is added (e.g. incomplete data), b), a ground truth knowledge graph which contains triples that can be constructed from such purchase-to-pay documents, and c), a provenance graph that contains relationships between information evidences in the documents and triples in the ground truth knowledge graph.

Besides enabling evaluation for such privacy sensitive cases, ptpDG can be used as a visualization and presentation tool for knowledge-graph-based services in the purchase-to-pay domain for stakeholders without the need to work on real sensitive data in the first place. Also, ptpDG may be leveraged for benchmarking knowledge graph construction techniques such as RDF mapping engines.

2 Approach

This section presents the general approach of ptpDG by means of Figure 1:

Initialization: With a configuration knowledge graph (1), it is possible to configure scenario related entities such as organizations, products or persons and their relations to each other (e.g. 1:1, 1:n, n:m). Based on this configuration, an initialization knowledge graph (2) for the next steps is generated. This contains the entities with their labels as well as required ontologies, such as P2P-O [6] for the purchase-to-pay domain.

Simulation: Because real purchase-to-pay processes emerge while people in organizations take decisions, we developed a multi-agent system (MAS) simulation (3) to realize the decentralized creation of such processes and their documents. In the simulation, organizations as agents purchase, sell and complain about products from which purchase-to-pay documents and their contents are created as triples in the ground truth knowledge graph (4). As a result, various types of processes emerge that are also specified in the invoicing norm EN16931 [3], for example, processes with sporadic purchase orders, with and without credit notes or with partial and final invoices. For knowledge workers in real purchase-to-pay processes, reconstructing such processes is a challenging

task which is why building knowledge graphs in such scenarios may be a promising approach in the first place [6]. Which particular processes are generated in the simulation depends on the randomized and individual decisions organizations take during the simulation, e.g., whether to complain about an invoice or not. For possibilities how to adjust parameters, we kindly refer the reader to `https://purl.org/ptp-dg#simulation`.

Purchase-To-Pay Documents: Electronic purchase-to-pay documents, which are now as triples in the ground truth knowledge graph, are generated with the Purchase-To-Pay Document Generator (5) in configured standards, formats and syntaxes. Because those documents are still too perfect compared to real-world documents, based on the idea in [5], noise is added with patterns found in real invoices, credit notes etc. (6). The current set of patterns have been derived from interviewing invoice processing industry experts in the TRAFFIQX network[4] and from analyzing real documents of this network. For example, regarding patterns how purchase-to-pay documents are referred to each other (or not), only last digits of invoice- or order-references are displayed, or such references are left out completely. Another common pattern is that the person who is responsible for the order or invoice – or her/his name abbreviation – is entered in the field that is actually preserved for the document reference.

Provenance Graph: To enable data-driven evaluation of knowledge graphs constructed from such documents, a provenance knowledge graph (7) is generated which contains relationships between particular information evidences in the documents (e.g. the name abbrevation in the order reference field) and the correct triples that may be constructed from this information. In the current case of XML-documents, the concrete location within a document is represented in an XPath query. Finally, for the generated dataset and knowledge graphs, metadata such as configuration parameters is generated.

3 Application of ptpDG

On ptpDG's project site `https://purl.org/ptp-dg`, a tutorial shows how a dataset with 105k triples was generated in which six organizations trade 30 products over 60 rounds of simulation. In this dataset, 1328 different processes are generated with 2277 documents in total. To ensure that resulting documents comply with given standards, they have been validated against respective XSD specifications. Consistency of the resulting knowledge graphs have been evaluated with OWL reasoner, which also means that the knowledge graphs comply with OWL restrictions specified in P2P-O [6]. As presented on the project site, further plausibility checks with SPARQL queries and expected results have been conducted, for example, to ensure that the number of final invoices and number of partial-final invoicing processes is equal.

[4] `https://www.traffiqx.net/en/about-us`

4 Related Work

Different invoice generators exist for presentation purposes and use cases, for example, for entity extraction from paper-based invoices that have been scanned [1]. However, such approaches do not provide labeled data to evaluate knowledge graph construction approaches. Also, to the best of our knowledge, there is no approach that considers the process context of invoices. ptpDG is inspired by a previous approach called Data Sprout [5]. It also generates labeled data and ground truth triples from a given content knowledge graph in the context of heterogeneous spreadsheet generation. However, besides generating other type of data for purchase-to-pay processes, ptpDG extends this approach by introducing a MAS simulation and, as a result, by dispensing with the content knowledge graph as an input.

5 Conclusion and Outlook

This paper introduced ptpDG, a labeled-dataset generator for the sensitive purchase-to-pay domain based on a MAS simulation. In this way, ptpDG moves towards enabling data-driven evaluation of knowledge-graph-based services: A knowledge graph construction approach can now take the generated documents as an input, and the resulting knowledge graph can be evaluated against the provided provenance and ground truth knowledge graph.

For future work, we plan to extend ptpDG with more heterogeneous documents, for example, with synthetic emails as purchase orders and other documents such as dispatch advice- and service provision-documents. This way, it will be possible to cover more kinds of processes specified in EN16931 [3]. Also, we plan to include more patterns (as in [5]) regarding organization names, product descriptions, and in general regarding those fields where users can insert text freely to better align the generated documents with real ones. To further evaluate the generated data beyond the presented plausibility checks, we work on the structural comparison between synthetic and real data. First results indicate that with the current version of ptpDG it is easier to find a configuration that generates correct ratios of different kinds of processes and documents than it is to find a configuration that at the same time generates correct time intervals. Further, the support of more different standards and syntaxes such as EDIFACT[5] is planned.

Acknowledgements This work was funded by the Investitions- und Strukturbank Rheinland-Pfalz (ISB) (project InnoProm) and the BMBF project SensAI (grantno. 01IW20007).

[5] `https://unece.org/trade/uncefact/introducing-unedifact`

References

1. Blanchard, J., Belaïd, Y., Belaïd, A.: Automatic generation of a custom corpora for invoice analysis and recognition. In: Workshop on Industrial Applications of Document Analysis and Recognition, WIADAR@ICDAR 2019, Sydney, Australia, September 22-25, 2019. IEEE (2019)
2. Ciccio, C.D., Marrella, A., Russo, A.: Knowledge-intensive processes: Characteristics, requirements and analysis of contemporary approaches. J. Data Semant. **4**(1), 29–57 (2015)
3. EN 16931-1:2017: Electronic invoicing - part 1: Semantic data model of the core elements of an electronic invoice. Standard, CEN (2017)
4. Ji, S., Mittal, P., Beyah, R.A.: Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. IEEE Commun. Surv. Tutorials **19**(2), 1305–1326 (2017)
5. Schröder, M., Jilek, C., Dengel, A.: Dataset generation patterns for evaluating knowledge graph construction. In: The Semantic Web: ESWC 2021 Satellite Events - Virtual Event, June 6-10, 2021, Revised Selected Papers. Lecture Notes in Computer Science, vol. 12739, pp. 27–32. Springer (2021)
6. Schulze, M., Schröder, M., Jilek, C., Albers, T., Maus, H., Dengel, A.: P2P-O: A purchase-to-pay ontology for enabling semantic invoices. In: The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings. LNCS, vol. 12731, pp. 647–663. Springer (2021)