Christophe Debruyne, Katy Fokou, and Paul Stijfhals. Do you need a knowledge graph? helping organi- zations determine whether a knowledge graph is needed for their problems with the smals KG checklist. In Ilaria Tiddi, Maria Maleshkova, Tassilo Pellegrini, and Victor de Boer, editors, Joint Proceedings of the Semantics co-located events: Poster&Demo track and Workshop on Ontology-Driven Conceptual Modelling of Digital Twins co-located with Semantics 2021, Amsterdam and Online, September 6-9, 2021, volume 2941 of CEUR Workshop Proceedings. CEUR-WS.org, 2021

Do you need a knowledge graph? Helping organizations determine whether a knowledge graph is needed for their problems with the Smals KG Checklist

Christophe Debruyne, Katy Fokou, and Paul Stijfhals

Smals Research, Smals, Avenue Fonsny 20, 1060 Brussels, Belgium
first.last@smals.be

Abstract. A knowledge graph (KG) is a graph that needs to fulfill specific criteria and is key in unlocking data siloes and tacit information for innovative applications. However, organizations may find it hard to identify use cases for KGs and even assessing whether a KGs is a viable approach for a particular problem. Within Smals Research, we designed the Smals KG Checklist, which guides a group of stakeholders in determining whether KG technologies can be used to address a concrete problem. The checklist is to be used in a workshop environment where a facilitator guides the stakeholders in filling the checklist.

Keywords: knowledge graph engineering, project planning, project management

1 Introduction

Smals is a non-profit organization that realizes innovative ICT projects for mostly Belgian (semi-)government departments and other affiliated members. Within Smals, the Smals Research[†] department investigates opportunities for using, deploying, and developing skills and know-how for emerging and innovative technologies for its members. One of these technologies, amongst many others, is knowledge graphs (KGs) [1].

It has become apparent that business analysts[‡] and members (from now on called stakeholders) have difficulty understanding the concept of KGs. Stakeholders do have a good understanding of the problems they face, but do not know how these problems can be solved using KGs. Anecdotally, for instance, stakeholders have shown interest in recommender systems to propose related information, whereas it turned out to be "simply" information about the same entity from two different silos—i.e., a data integration problem. Stakeholders are familiar with graph databases as they have been adopted to solve specific problems such as graph analytics. Still, these graphs cannot

[†] https://www.smalsresearch.be/

[‡] Smals employees that are working closely with its members to comprehend their business needs, translating those into ICT projects, and follow up and manage the realization of these projects.

(yet) be regarded as KGs, which adds to the confusion.

To help stakeholders understand the concept of a KG and, more importantly, to assist them in assessing whether a KG is a viable approach in tackling concrete problems, we have designed the *Smals KG Checklist*[§]. The checklist is to be used in a collaborative setting such as a workshop in which stakeholders provide input. The checklist acts as a guide for the workshop facilitator, and it is up to the facilitator to capture and refine the stakeholder's input.

2 Graphs vs. Knowledge Graphs

In [1], the authors analyzed various definitions of the term KG. While there is no consensus, [1] allows us to summarize that a KG is a graph (representing entities and their relationships) that fulfills three criteria: **C1**) the KG has a non-trivial schema or *ontology*; **C2**) the KG integrates information from heterogeneous sources; and **C3**) the KG is used to gain insights by inferring implicit information from explicit information (either via the schema, machine learning, or tooling on top of the KG).

KGs are typically stored in graph databases, but graph databases are also used in other scenarios. Graph databases are necessary for graph analytics and may solve issues stemming from a relational database's computational limitations. An example of the latter is Neo4j's use case on access and identity management. Neo4j reported on the migration of a relational database to a graph database to avoid expensive recursive joins.[**] As the data was merely migrated and not all criteria were met, we argue that this project did not yield a KG.

How can we determine whether a project needs a graph or a KG? To determine whether a particular problem can be solved with a KG, we need to determine whether a solution requires the three criteria to be met.

3 Related Work

KG technology vendors often publish whitepapers and blog posts on the successful application of KG. Even organizations report on their use cases and lessons learned (e.g., [2]). While valuable, not all organizations are faced with the same problems and organizations do not necessarily have the expertise to extrapolate those examples to similar cases. There are also resources aimed at both academia and industry (e.g., [3]). While they also provide valuable information on everything involved in a KG project (activities, methodologies, techniques, etc.) and examples, they do not provide a tool such as our checklist to determine the applicability of KG technologies for a problem. The Smals KG Checklist, which we present in this paper, thus addressed this critical gap.

We designed a tool that can be used in a workshop setting, much like a business model canvas or an ethics canvas [4]. As we start from concrete problems and the questions we needed to be answered are well-scoped, we did not pattern our checklist after these canvases.

[§] Made available with a CC BY-NC-SA 4.0 via https://www.smalsresearch.be/wp-content/uploads/2021/06/smals-kg-checklist.pdf

[**] https://neo4j.com/blog/enterprise-identity-access-management/, last accessed May 25th, 2021

4 The Smals KG Checklist

The checklist consists of two parts (see Fig. 1 and Fig. 2). In Part I, we first identify the problem, stakeholders, and core concepts. Then we aim to answer three questions by filling in Part II (Fig. 2). These three questions are related to the criteria of a KG mentioned in the previous section. The final question on Part I is used to identify future opportunities for the KG, or its applicability in the longer term. Once filled in, and refined over time, the checklist can be used to determine whether KG technologies are needed to solve a particular problem. The questions on KG criteria are given a color, and these colors reappear in the sections of the second part: purple corresponds with C1, green with C2, and orange with C3. These sections are key in determining whether all criteria are (or have to be) met:

- Section I is used to list the sources that will inform our schema. Section I has two colors as the integration of an existing database could appear in both the bottom-up integration of structured data (Section II in Fig. 2) and as input for the KG's schema by lifting its database schema.
- Sections II and IV are concerned with integrating structured and unstructured data, respectively. The integration of provenance information, metadata, annotations,…of data mentioned in Sections II and IV is captured in Section III, and therefore placed between Sections II and IV. We have noticed that it helps to ask this question explicitly.
- Sections V and VI are respectively concerned with symbolic reasoning (e.g., OWL reasoning) and machine learning. As reasoning over the KG requires an ontology, there is an arrow from Section II to Section I. Reasoning and AI techniques are clear indications of gathering insights from the KG and are therefore orange. We can argue that retrieving information according to Linked Data principles does not necessarily help one gain insights, but visualization tools (e.g., [5, 6]) and faceted browsing (e.g., [7, 8]), amongst others, may. This is why Section VII, concerned with applications on top of the KG, has a gradient fill instead of a complete fill. We deem the criteria w.r.t. to Section VII fulfilled when the need for such tools to gather insights is mentioned.

We use one of Smals Research's KG projects with the Belgian Social Security to illustrate the checklist by filling in the forms. This project aims to integrate the data of three databases to which inspectors have access but are unable to query as a whole. Integrating the data into a KG requires an ontology and would greatly facilitate their work by allowing them to gain insights in, for instance, one's employment histories.

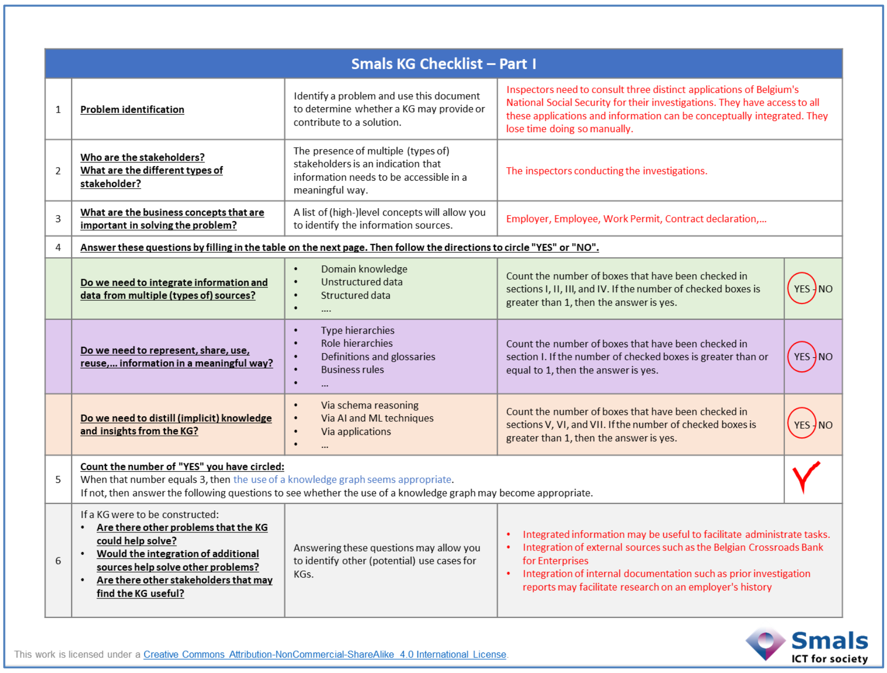
5 Summary

It is difficult for organizations to identify when KGs technologies are a viable means to an end. There are many valuable resources on KGs available, but they either focus on constructing KGs or report on KG use cases that organizations need to extrapolate. To help organizations determine whether a KG will help tackle a concrete problem, we designed the Smals KG Checklist. The tool is meant to be filled by a group of stakeholders, with a facilitator taking the lead and guiding the discussion.

Disclaimer. The views and opinions expressed in this paper are those of the authors and do not express the views or opinions of Smals. *Acknowledgements.* We thank Gunther Hellebaut, Karel Van Eeckhoutte, and Catherine Vanden Daelen for their feedback.

References

1. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C.N., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J.F., Staab, S., Zimmermann, A.: Knowledge Graphs. CoRR. abs/2003.0, (2020).
2. Hubauer, T., Lamparter, S., Haase, P., Herzig, D.M.: Use Cases of the Industrial Knowledge Graph at Siemens. In: van Erp, M., Atre, M., López, V., Srinivas, K., and Fortuna, C. (eds.) Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018. CEUR-WS.org (2018).
3. Pan, J.Z., Vetere, G., Gómez-Pérez, J.M., Wu, H. eds: Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer (2017). https://doi.org/10.1007/978-3-319-45654-6.
4. Reijers, W., Koidl, K., Lewis, D., Pandit, H.J., Gordijn, B.: Discussing Ethical Impacts in Research and Innovation: The Ethics Canvas. In: Kreps, D., Ess, C., Leenen, L., and Kimppa, K. (eds.) This Changes Everything - ICT and Climate Change: What Can We Do? - 13th IFIP TC 9 International Conference on Human Choice and Computers, HCC13 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 19-21, 2018, Proceedi. pp. 299–313. Springer (2018). https://doi.org/10.1007/978-3-319-99605-9_23.
5. Mouromtsev, D., Pavlov, D., Emelyanov, Y., Morozov, A., Razdyakonov, D., Galkin, M.: The Simple Web-based Tool for Visualization and Sharing of Semantic Data and Ontologies. In: Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015. CEUR-WS.org (2015).
6. Debruyne, C., O'Sullivan, D.: Visually Exploring SPARQL Endpoints with Murmuration. In: Garijo, D. and Lawrynowicz, A. (eds.) Proceedings of the EKAW 2020 Posters and Demonstrations Session co-located with 22nd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2020), Globally online & Bozen-Bolzano, Italy, September 17, 2020. pp. 17–21. CEUR-WS.org (2020).
7. Kharlamov, E., Giacomelli, L., Sherkhonov, E., Grau, B.C., Kostylev, E. V., Horrocks, I.: SemFacet. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17. pp. 2475–2478. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3132847.3133192.
8. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter - Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L.J.B., Kurz, T., Schlegel, K., and Sande, M. Vander (eds.) Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop co-located with the 13th Extended Semantic Web Conference ESWC 2016, Heraklion, Crete, Greece, May 30, 2016. CEUR-WS.org (2016).

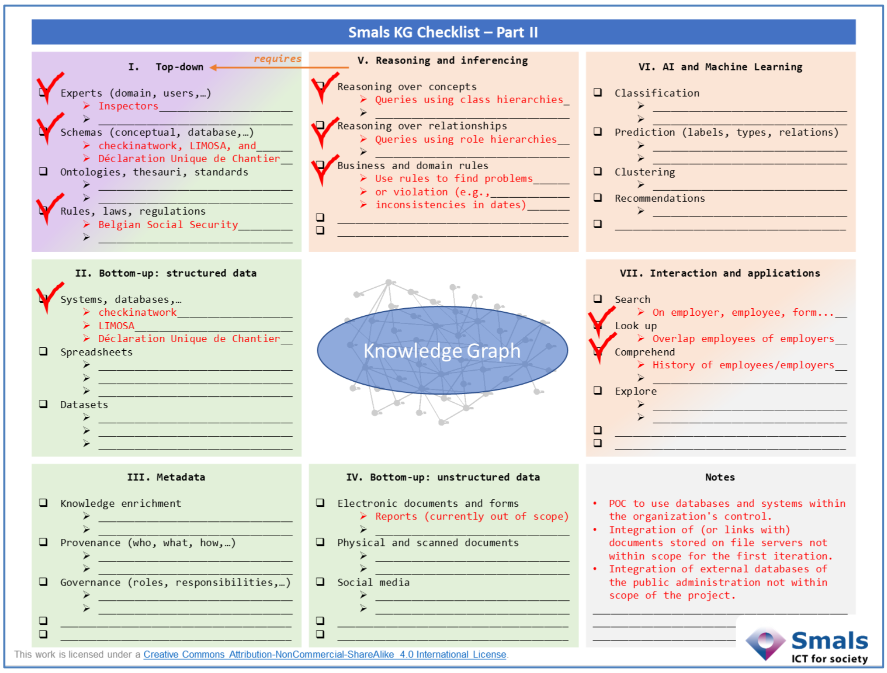


Smals KG Checklist – Part I

1	**Problem identification**	Identify a problem and use this document to determine whether a KG may provide or contribute to a solution.	Inspectors need to consult three distinct applications of Belgium's National Social Security for their investigations. They have access to all these applications and information can be conceptually integrated. They lose time doing so manually.	
2	**Who are the stakeholders?** **What are the different types of stakeholder?**	The presence of multiple (types of) stakeholders is an indication that information needs to be accessible in a meaningful way.	The inspectors conducting the investigations.	
3	**What are the business concepts that are important in solving the problem?**	A list of (high-)level concepts will allow you to identify the information sources.	Employer, Employee, Work Permit, Contract declaration,…	
4	**Answer these questions by filling in the table on the next page. Then follow the directions to circle "YES" or "NO".**			
	Do we need to integrate information and data from multiple (types of) sources?	• Domain knowledge • Unstructured data • Structured data • ….	Count the number of boxes that have been checked in sections I, II, III, and IV. If the number of checked boxes is greater than 1, then the answer is yes.	[✓ YES] - NO
	Do we need to represent, share, use, reuse,… information in a meaningful way?	• Type hierarchies • Role hierarchies • Definitions and glossaries • Business rules • …	Count the number of boxes that have been checked in section I. If the number of checked boxes is greater than or equal to 1, then the answer is yes.	[✓ YES] - NO
	Do we need to distill (implicit) knowledge and insights from the KG?	• Via schema reasoning • Via AI and ML techniques • Via applications • …	Count the number of boxes that have been checked in sections V, VI, and VII. If the number of checked boxes is greater than 1, then the answer is yes.	[✓ YES] - NO
5	**Count the number of "YES" you have circled:** When that number equals 3, then the use of a knowledge graph seems appropriate. If not, then answer the following questions to see whether the use of a knowledge graph may become appropriate.			✓
6	If a KG were to be constructed: • **Are there other problems that the KG could help solve?** • **Would the integration of additional sources help solve other problems?** • **Are there other stakeholders that may find the KG useful?**	Answering these questions may allow you to identify other (potential) use cases for KGs.	• Integrated information may be useful to facilitate administrate tasks. • Integration of external sources such as the Belgian Crossroads Bank for Enterprises • Integration of internal documentation such as prior investigation reports may facilitate research on an employer's history	

Smals ICT for society

Fig. 1. Part I of the Smals KG Checklist. Red denotes the answers for one of our pilot projects.



Smals KG Checklist – Part II

I. Top-down

- ☑ Experts (domain, users,…)
  - ➢ Inspectors____________________
  - ➢ ______________________________
- ☑ Schemas (conceptual, database,…)
  - ➢ checkinatwork, LIMOSA, and______
  - ➢ Déclaration Unique de Chantier__
- ❑ Ontologies, thesauri, standards
  - ➢ ______________________________
  - ➢ ______________________________
- ☑ Rules, laws, regulations
  - ➢ Belgian Social Security________
  - ➢ ______________________________

II. Bottom-up: structured data

- ☑ Systems, databases,…
  - ➢ checkinatwork__________________
  - ➢ LIMOSA_________________________
  - ➢ Déclaration Unique de Chantier__
- ❑ Spreadsheets
  - ➢ ______________________________
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Datasets
  - ➢ ______________________________
  - ➢ ______________________________
  - ➢ ______________________________

III. Metadata

- ❑ Knowledge enrichment
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Provenance (who, what, how,…)
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Governance (roles, responsibilities,…)
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ ________________________________
- ❑ ________________________________

V. Reasoning and inferencing — requires → I. Top-down

- ☑ Reasoning over concepts
  - ➢ Queries using class hierarchies_
  - ➢ ______________________________
- ☑ Reasoning over relationships
  - ➢ Queries using role hierarchies__
  - ➢ ______________________________
- ☑ Business and domain rules
  - ➢ Use rules to find problems_____
  - ➢ or violation (e.g.,___________
  - ➢ inconsistencies in dates)______
- ❑ ________________________________
- ❑ ________________________________

Knowledge Graph

IV. Bottom-up: unstructured data

- ❑ Electronic documents and forms
  - ➢ Reports (currently out of scope)
  - ➢ ______________________________
- ❑ Physical and scanned documents
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Social media
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ ________________________________
- ❑ ________________________________

VI. AI and Machine Learning

- ❑ Classification
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Prediction (labels, types, relations)
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ Clustering
  - ➢ ______________________________
- ❑ Recommendations
  - ➢ ______________________________
- ❑ ________________________________

VII. Interaction and applications

- ❑ Search
  - ➢ On employer, employee, form...__
- ☑ Look up
  - ➢ Overlap employees of employers__
- ☑ Comprehend
  - ➢ History of employees/employers__
  - ➢ ______________________________
- ❑ Explore
  - ➢ ______________________________
  - ➢ ______________________________
- ❑ ________________________________
- ❑ ________________________________

Notes

- POC to use databases and systems within the organization's control.
- Integration of (or links with) documents stored on file servers not within scope for the first iteration.
- Integration of external databases of the public administration not within scope of the project.

Smals ICT for society

Fig. 2. Part II of the Smals KG Checklist. Red denotes the answers for one of our pilot projects.