Operational Research Literature as a Use Case for the Open Research Knowledge Graph

Mila Runnwerth[1], Markus Stocker[1][0000−0001−5492−3212], and Sren Auer[1][0000−0002−0698−2864]

TIB - Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hannover, Germany
http://www.tib.eu
{firstname.surname}@tib.eu

Abstract. The Open Research Knowledge Graph (ORKG) provides machine-actionable access to scholarly literature that habitually is written in prose. Following the FAIR principles, the ORKG makes traditional, human-coded knowledge findable, accessible, interoperable, and reusable in a structured manner in accordance with the Linked Open Data paradigm. At the moment, in ORKG papers are described manually, but in the long run the semantic depth of the literature at scale needs automation. Operational Research is a suitable test case for this vision because the mathematical field and, hence, its publication habits are highly structured: A mundane problem is formulated as a mathematical model, solved or approximated numerically, and evaluated systematically. We study the existing literature with respect to the Assembly Line Balancing Problem and derive a semantic description in accordance with the ORKG. Eventually, selected papers are ingested to test the semantic description and refine it further.

Keywords: Knowledge graph · Mathematical knowledge management · Operational research literature · Operations research literature

1 Introduction

Today's scholarly communication behaviour and logistics is still defined by centuries of printed document culture. Although there is progress by transforming journals into digital article repositories that, in principle, provide access to the content at all times and irrespective of a researcher's location, the nature of an article itself has not changed: The investigated hypothesis, the used methodology, the experiment, and the outcome are written in prosaic form; the final document is usually published for no other purposes than reading, seemingly optimised for human cognition.

The Open Research Knowledge Graph (ORKG) [8] questions the paradigm of document-centric scholarly information communication [2]. It aims at transforming research literature into structured, machine-actionable data in order to represent and express information through semantically rich, interlinked knowledge graphs. Similarly to DBpedia, a prosaic knowledge source is transformed

according to Linked Open Data standards [1]. Users are enabled to compare papers, discover patterns across methods or disciplines, or get a structured overview in a chosen context.

The main use cases of the ORKG's beta version[1] are article search, a machine-actionable, semantic representation, and especially paper comparison as introduced in [10]. To date, it indexes about 400 research articles. More than half are assigned to the subject cluster *Physical Sciences & Mathematics*.

1.1 Operational Research as a Use Case from Mathematics

The structural science mathematics provides particularly suitable content for the ORKG: Its published prose is clear and dense from a linguistic point of view. However, as [4] have shown the high degree of abstraction in mathematics makes a conceptualisation consisting of the categories *process*, *method*, *material*, and *data*, which have been adapted to empirical sciences, inexpedient. The ORKG is not limited to this model but its feature, the abstract annotator, has been shown to be out of its depth with regard to mathematics. The applied mathematical science of operational research (OR) combines the rather abstract fields of combinatorics and numerical analysis with mundane research questions from economics. In favour of this study, we narrowed the topic down to the optimisation problem of *Assembly Line Balancing*. Its name derives from mass production where the intricate logistics for paced manufacturing assembly lines have to be organised efficiently, i. e. optimally. The Assembly Line Balancing Problem (ALBP) and its variations are not only well-covered in scholarly literature but also provide an abundance of structured overviews of exact and heuristic algorithms or benchmarks thereof. Thus, the research literature about ALBPs is an appropriate use case for the ORKG. In a first step, we choose literature reviews and articles that suggest minor optimisations to existing methods, which are compared to each other. Then, we suggest a semantic description that covers the content of the collection. It will serve as a prospective template for the ORKG. Furthermore, we will look for elements and patterns in the papers that are suited for automatic extraction in the future. Third, we ingest those literature reviews or articles that are published under an eligible licence, i. e. a CC-BY, CC-BY-SA, or arXiv's *Non-exclusive licence to distribute* into the ORKG. During this intellectual step, we will test and refine the proposed data model. Finally, we consider the representations and comparisons of the scholarly contributions in the ORKG and discuss its added value for researchers.

2 A Template for the Assembly Line Balancing Problem

Scholarly research of the ALBP can be traced back to the 1960s when it was shown to be a NP-hard combinatorial optimisation problem [7]. Since then scientists work on the sophistication of the mathematical model, exact algorithms

[1] `https://orkg.org/`

for defined special cases or heuristic algorithms in order to find optimal solutions in adequate time. Recently, several reviews have been published to benchmark the stated mathematical model, exemplary data scenarios [5] or performances of the selected methods. We chose to set a focus on these reviews at first, but most publications were not openly accessible or free to be reused in the ORKG according to their respective licences. Eventually, articles introducing a new model statement for the ALBP or a heuristic to solve it were also considered. The collection comprised 28 topically relevant papers of which eight provide openly accessible preprints on *arXiv*[2]. These were manually ingested into the ORKG with varying degrees of thoroughness (cf. Section 2.2): From the single statement of the research problem to detailed descriptions of the algorithms and data sets that were applied[3].

The collection of research articles was organised in the open source reference management system Zotero[4], also including documented experiences of the whole process.

2.1 A Semantic Model to Reflect ALBP Research

The ORKG's performance depends on a data model that is well-tuned to the content it is supposed to represent. That means expert knowledge in both the considered field and data modelling is required. Authors who possess the domain knowledge may not be able to squeeze it into the RDF scheme of the ORKG because there is no or little expertise in knowledge engineering. Data curators on the other hand may struggle with the proper in-depth indexing of the latest research knowledge. The ORKG's flexibility is an advantage because it allows almost limitless adaptions to describing papers by reusing existing concepts (mostly entries by former contributors) and relations but also by introducing new ones. The default schema stems from the comprehension of empirical sciences: A method is applied to a defined research question. This application causes a process that involves material to be observed or changed. Meanwhile observational data is collected and eventually evaluated in order to prove or disprove a hypothesis constructed prior to the experiment.

In operational research in general and with respect to the ALBP in particular, there is also a rather standardised development that can be represented by a data model: The practical problem is formulated as a mathematical model or programme. Depending on the choice of the model, there is a toolbox of direct or heuristic algorithms to yield an exact (or approximate) solution to the model. Usually, in scholarly literature either a new variant of the ALBP is stated and the derived model is traced back to established methods or a new or rather slightly modified method is tested against known methods to solve the same problem. Thus, we conclude that most research papers about the ALBP are comprised of the elements listed in Table 1.

[2] `arxiv.org`

[3] An exemplary comparison of three selected papers can be found at `https://www.orkg.org/orkg/comparison?contributions=R12018,R12059,R12193`.

[4] `https://www.zotero.org/`

Table 1. A semantic model translating the OR research process into the ORKG scheme. The arrows connote a hierarchical descent, the asterisks connote a newly introduced property.

OR Term	**ORKG Properties**	**Example**
Name of the optimisation problem	Has research	ALBP
Model / programme	Has approach → Has model*	MIP
Exact method	Has method → Has exact solution method*	Branch & bound
Heuristic	Has method → Has heuristic*	Tabu search
Instance data set	Has instance*	Roszieg
Programming language	Has implementation → Has programming language*	C, GCC 3.4.0
System specifications	Has implementation → Has system specification*	Athlon 64 X2 4400
Performance	Has performance*	$\mathcal{O}(n \log n)$, 0,2 ms

Has Performance contains the results that depend on the method that is applied, the graph the algorithm is applied on, and the specification of the implementation and system. Thus, it is semantically interlinked with other elements.

If the suggested structure in Table 1 proves valid, it can be cast into a topic-specific template on its own in order to facilitate highly consistent knowledge graphs of further relevant papers independent of the curator.

2.2 Entering ALBP Literature into the ORKG

After careful study and annotation of the eight papers from arXiv, we entered the data into the ORKG. In the first of three steps of the procedure, the formal metadata can be automatically ingested via DOI[5] or a BibTeX entry. There is an additional fallback option to enter the formal metadata manually. Since the preprint repository arXiv does not provide a DOI for its documents, we chose BibTeX entries for the import.

In the second step, the document is classified by subject. The ORKG's specified, hierarchical classification does currently not allow for several attributions. Hence, when assigning a single subject, a multidisciplinary field such as operational research is prone to inconsistencies with respect to its main focus in the respective paper or the curator. We chose to consistently assign the collection to Applied Mathematics → Numerical Analysis & Computation, although several other closely related fields would have been adequate as well, for example Engineering → Operations Research (and more). However, OR being predominantly a multidisciplinary subject involving mathematics, computer science, and economics, engineering seemed too misleading for a semantically sound assignment.

[5] `https://www.doi.org/`

The curator may choose between several templates; the template called *Research Problem* is closest to the data model as suggested in Section 2.1. The template provides the field *Has research* where keywords can be chosen from the suggested list or entered manually. Each entry is added to a bag-of-words and, thus, will be provided for autocompletion further on. This unrestricted freedom leads to a number of challenges. We struggled with typos (e. g. 'optimsation') as it was not immediately obvious how to correct these. Moreover, the same word was included in its American and British form, respectively, i. e. 'optimization' and 'optimisation'. We plan to add some functionality to ORKG to semi-automatically interlink such surface forms as they are describing the same concept. An underlying controlled vocabulary with an additional feature to enter free text would avoid wreaking havoc in the bag-of-words. A user entering 'optimisation problem' may thus be faced with four versions of which one contains a typo and two are identical.

Further predefined fields are *Has evaluation*, *Has approach*, *Has method*, *Has implementation*, *Has result*, *Has value*, and *Has metric*. Not all of these semantic relations make sense for describing a mathematical paper, or rather, they lack distinctive accuracy, e. g. when does an approach become a method; or do we mean the outcome of the algorithm or its performance when stating the result? Yet, the relevant semantic units of an OR paper can be transferred and amended easily.

Each field contains further fields in turn that may be annotated and indexed. And as a last resort, new relations can be introduced on every hierarchical level. The OR terms introduced in Table 1 were mapped by employing existing relations and introducing new ones (marked by an asterisk in the table). After leaving the hierarchical top level which is edited in the main browser window, every edit thereafter is conducted in a small overlay window. So while modelling, there is no visual aid where the description process is hierarchically taking place at the moment. However, we made it a habit to describe the top level first, save the description, such that the visualisation of the graph is available. From there, refinement is more accessible.

The eight papers are not consistently described in this fashion because each paper gave reason to a refinement iteration of previous graphs. Thus, after each paper, there are (or should be) well-documented, retroactive modifications to each graph representing a paper. Again, this inevitably leads to inconsistencies even among papers that are ingested by the same curator. Another critical observation is our choice of terms: General denominations such as *Has model*, *Has instance*, or *Has performance* could mean completely different things in another context. Even between OR researchers these terms might not be semantically tight enough to guarantee frictionless communication. Hence, the relations are prone to cross-contextual use that might make the otherwise carefully created model fuzzy.

In theory, a paper can be thoroughly represented by modelling each sentence as Linked Data, at arbitrary granularity. Again, in agreement with another field expert from biochemistry, we concluded: when to finish the indexing procedure

is at the margin of discretion. Of course, the ORKG's crowd-sourcing philosophy allows and even demands for further refinement by others or at a later stage. Thus, a knowledge graph is never truly complete, especially if dynamic data such as citations will be taken into account in the future. An exemplary paper description is shown in Figure 1.

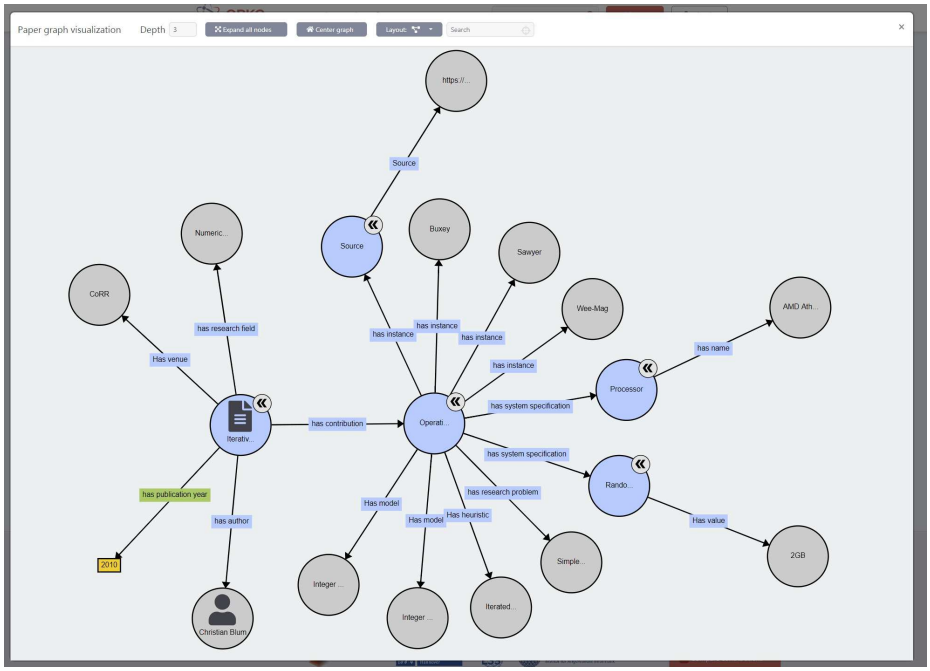

Fig. 1. Visualisation of a paper's knowledge graph.

3 Conclusion and Further Work

OR is a suitable test case for the ORKG, because the topic itself and the appropriate publication habits are highly structured and can easily be mapped to the default data schema already provided by the ORKG. However, on the basis of this study we tackle several general and subject-specific further improvements in the future:

- Creating checklists and guidelines to define both minimum requirements and a gold standard for a paper's knowledge graph.
- Underlying a general, and for templates a subject-specific, vocabulary with moderated editing workflows. Also, the resources should be displayed alphabetically or by assigned relevance instead of a last-in-first-out fashion in the tabular view.

- Linear user guidance for generating a skeleton data set and visual support for the refinement.
- Similarly described papers with identical or semantically close descriptions should yield similarity.
- The selected papers met our expectations of being highly structured and easy to parse for the defined information patterns. They are well suited for a pilot study of automated extraction for information framing a basic knowledge graph. Automated indexing where scientific literature is indexed with terms of well-maintained thesauri like the German Authority File or automatically classified with the Mathematics Subject Classification (MSC)[6] may provide a first draft to be ingested into the ORKG[9].
- The aforementioned MSC would provide the obvious classification backbone for contributions from the mathematical sphere. The extremely confined example of the ALBP suggests that this would not only call for 63 template schemas for each top level class but at least 5.000 refinements accounting for MSC's subclasses. However, with this first experiment we cannot estimate the structural synergies between classes. We rather expect, given a wisely chosen sample that future work might result in a manageable number of mathematical templates with few extensions for the subclass topics. An example are the MSC classes 44 and 45 covering ordinary and partial differential equations, respectively. Even if they turn out to differ minutely in their ORKG template, these differences will be provided for in other templates, e. g. (numerical) analysis.
- Since the ORKG follows a crowdsourcing philosophy, seeking support from and collaborate with further projects in the field of mathematical knowledge engineering guarantees high quality and integrity of the data and its community-curated modelling. Critical exchange with the researchers of *MathDataHub* is established[3], but projects like swMATH, a database for mathematical software, should be considered more closely[6].

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. International Semantic Web Conference Asian Semantic Web Conference 2007, (2007). https://doi.org/10.1007/978-3-540-76298-0_52
2. Auer, S.; Kovtun, V.; Prinz, M.; Kasprzik, A.; Stocker, M.: Towards a Knowledge Graph for Science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018). https://doi.org/10.15488/3401
3. Berčič, K., Kohlhase, M., Rabe, F.: Towards a Unified Mathematical Data Infrastructure: Database and Interface Generation. In: Proceedings of the Twelfth Conference on Intelligent Computer Mathematics (CICM 2019), (2019). https://doi.org/10.1007/978-3-030-23250-4_3
4. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-independent Extraction of Scientific Concepts from Research Articles. CoRR (2020). http://arxiv.org/abs/2001.03067

[6] http://msc2010.org/mediawiki/index.php?title=MSC2010

5. Chaves, A. A., Miralles, C., Lorena, L. A. N.: Clustering search approach for the assembly line worker assignment and balancing problem. Proceedings of the 37th international conference on computers and industrial engineering (2007).
6. Chrapary H., Dalitz, W., Neun, W., Sperber, W.: Design, Concepts, and State of the Art of the swMATH Service. In: Math.Comput.Sci. 11, 469481 (2017). https://doi.org/10.1007/s11786-017-0305-5
7. Gutjahr, A. L., Nemhauser, G. L.: An Algorithm for the Line Balancing Problem. In: Management science, (1964). https://doi.org/10.1287/mnsc.11.2.308
8. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. Proceedings of the 10th International Conference on Knowledge Capture - K-CAP '19 (2019). https://doi.org/10.1145/3360901.3364435
9. Kasprzik, A.: Putting Research-based Machine Learning Solutions for Subject Indexing into Practice. In: Proceedings of the Conference on Digital Curation Technologies (Qurator 2020), (2020). http://ceur-ws.org/Vol-2535/paper_1.pdf
10. Oelen, A., Jaradeh, M. Y., Farfar, K. E., Stocker, M., Auer, S.: Comparing Research Contributions in a Scholarly Knowledge Graph. In: Proceedings of the Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), (2019). http://ceur-ws.org/Vol-2526/paper3.pdf