Final Report for CENIIT Project 17.05

Project Title: Integration and Interoperability of Graph-Data Systems

Project Leader: Olaf Hartig <olaf.hartig@liu.se>

1 Most Important Scientific Results

The project set out to make contributions related to six different stages of graph-data system integration (see Figure 1). While the goals of the sixth stage turned out to be too ambitious and are planned to be actively worked on during the next few years, we made significant contributions in the first five stages. These contributions resulted in five peer-reviewed journal articles and 19 peer-reviewed conference publications (see Section 8). *Five* of these publications were honored with a *best research paper award*, respectively; another one earned a *peoples' choice best poster award*. Other tangible contributions of the project are several repositories[1] of Open Source software developed both by the PI and by students supervised on thesis projects related to the project. The remainder of this section summarizes the most important results of the research in the project.

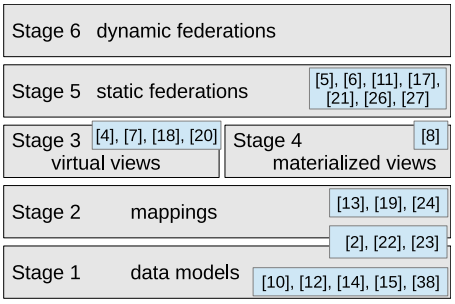

Fig. 1: Stages of graph-data system integration and related publications in the project.

RDF-star and SPARQL-star. The PI proposed an approach to extend the standard RDF graph data model and its query language SPARQL with features from the Property Graph model [22, 23]. This approach was honored with a best poster award at ISWC 2017 and became so popular that several RDF system vendors extended their products to support it (see Section 5). This adoption by the industry attracted further attention and a desire to develop the proposal into a standard. The first step towards this end was a specification [38] created by a W3C Community Group which the PI co-chaired. Since the end of 2022, a W3C Working Group is now in the process of producing the actual standard documents, with the PI as an invited expert.

Adaptations of RDF-star and SPARQL-star to streaming data. In a series of research papers we have shown that the RDF-star approach is suitable not only for static graph datasets but it can also increase the expressiveness and capabilities of systems that process streams of RDF-based graph data. To this end, we have extended the state-of-the-art data model and query language for such systems by adding the concepts of RDF-star [14]. Thereafter, we have shown how this extension can be employed to capture and to query different notions of uncertainty in RDF stream processing [12] and we have proposed query optimization techniques that are suitable in this context [10]. For two of these research papers [10, 14], we received a best paper award.

Interoperability based on RDF-star and SPARQL-star. Considering the RDF-star data model as a means to achieve interoperability between the RDF data model and Property Graphs, mappings between the data models have been introduced and studied. In particular, we have shown mappings that posses desirable properties such as information preservation and query result preservation [13] and we have compared the mappings to earlier mapping approaches in terms of query performance that RDF systems can achieve for data converted based on the mappings [8]. As an alternative to converting the data, we have also started working on an approach to translate SPARQL-star queries into the Property Graph query language Cypher (yet to be published).

Query-based interfaces to access RDF-based graph data. Another strand of research that is relevant to the integration of graph-data sources is to study relevant types of Web-based interfaces to access and to query such data sources. While the PI had been involved in some of this research already before the project, with the CENIIT funding he continued carrying out new research work in this area. Most notably, in collaboration with

[1] See https://github.com/LiUSemWeb (7 repos related to the project), https://github.com/LiUGraphQL (10 main repos related to the project), https://github.com/RDFstar (2 repos), and https://github.com/hartig/OptPlus4Jena

Jorge Pérez from the Universidad de Chile, we developed a theoretical model to compare different types of data access interfaces in terms of more fundamental properties rather than by experimentation. More specifically, we developed a formal framework for capturing possible client-server systems that may be implemented based on such interfaces, and we used the framework to conduct a comparison of several combinations of client and server capabilities [18, 20]. This work was honored with the Best Research Paper Award at ISWC 2017.

GraphQL interfaces. During the project, GraphQL has emerged as a tremendously popular approach to build Web-based data access interfaces. Since GraphQL provides a graph-like abstraction to query various forms of data sources, it became relevant for the project. While there existed numerous implementations of the GraphQL query language, a more fundamental understanding of the properties of the language was missing. Again in collaboration with Jorge Pérez, we have been the first to establish such an understanding [19, 24]. For instance, our work provides a formal semantics of the language, and we have shown that the combined complexity of the main decision problems for this language is NL-complete. Further GraphQL-related research that was conducted to some extend based on the CENIIT funding was the development of a performance benchmark for approaches to build GraphQL servers [7], an empirical analysis of GraphQL API schemas [16], and an adoption of the GraphQL schema language to define schemas for Property Graphs [15].

Query processing techniques for federations of graph data sources. As a basis for a virtual integration of graph-data sources in a federation, we also studied techniques for processing queries over such federations. As the most notable results of this work we developed a novel approach to index graph data of federation members and to leverage such indexes to achieve more efficient query execution plans [17, 21], and we developed a formal language to describe and reason about query execution plans [11]. For both of these works, we won a best paper award, respectively.

2 Degrees and Promotions

The project has contributed to the following degrees and promotions.

- Olaf Hartig became Docent in Computer Science in October 2018 and was promoted to the rank of a Senior Associate Professor (biträdande professor) in June 2022.
- Three PhD students graduated successfully with Olaf Hartig as co-supervisor, where the co-supervision was funded by the project. The three students are:
  - Huanyu Li (Department of Computer and Information Science, Linköping University) graduation in 2022, co-supervision since June 2017.
  - Robin Keskisärkkä (Department of Computer and Information Science, Linköping University) graduation in 2021, co-supervision since December 2017.
  - Jenni Reuben Shanthamoorthy (Department of Computer Science, Karlstad University) graduation in 2020, co-supervision since August 2017.

3 Master's Theses

During the time of the project, Olaf Hartig served as supervisor or examiner for a total of 19 Master's thesis projects where 12 of them have been directly related to the research topics of the project. For the complete list, refer to the appendix at the end of this document.

4 Persons Funded by the Project

The project funded varying percentages of the salaries of Olaf Hartig (PI), Sijin Cheng (research assistant, later PhD student in the project but then funded by CUGS), Musie Ebuy (research assistant), Rab Nawaz Jan Sher (amanuens), Robin Christensen (amanuens), Shahrzad Khayatbashi (PhD student), Sebastian Ferrada (technician, postdoc level), Simon Helgesson (amanuens).

5 Industrial Connections

The main industrial connections that the project had were with several vendors of RDF-based graph database systems. The most tangible outcome of these connections is the extension of several such systems with support for the RDF-star & SPARQL-star approach developed in the project. In particular, commercial products that already support the approach at the moment are the AnzoGraph system of Cambridge Semantics Inc. (contact: Barry Zane), Blazegraph of Systap, LLC (contacts: Bradley R. Bebee and Bryan Thompson), GraphDB of Ontotext (contact: Atanas Kiryakov), the neosemantics plugin of Neo4J (contact: Jesús Barrasa), and Stardog (contacts: Evren Sirin and Pavel Klinov). Additionally, the approach is supported in several open source systems and frameworks (e.g., Apache Jena, Eclipse RDF4J, Corese, EYE, Morph-KGC, Oxigraph, RDF.ex, RubyRDF) and is currently in the process of being integrated into the official RDF and SPARQL standards.

Further industrial connections are listed as follows.

- In collaboration with Systap, LLC (contacts see above), a novel Web-based data access interface that was studied in the project was integrated into the Blazegraph graph database system.
- Three **Master's thesis projects** that focused on different aspects of applying graph database technology as a basis for the continuous integration framework Eiffel at Ericsson were jointly supervised by Daniel Ståhl (Ericsson) and Olaf Hartig.
- A collaboration with Juan Sequeda (data.world, Inc.) led to **two articles** that were co-authored for an "Encyclopedia on Big Data" [25, 26].
- Together with the Stockholm-based company MetaSolutions AB (contact: Matthias Palmér), we applied for funding from Vinnova for a joint project on a search-index-based data access interface for graph data. Unfortunately, the application was not successful.
- Together with the French start up Clotho AI (contact: Etienne Pillin), we applied for a Horizon Europe project on a secure data platform to support research, training, and testing for forensic and security applications. Unfortunately, the application was not successful.
- Olaf Hartig acted as a **technical advisor** on Semantic Web and Knowledge Graph technologies for Scania AB (2019–2020) and for Semcon AB (2021). For Scania, we also delivered two one-week courses on these technologies as **commissioned education** (2019 and 2022).
- Consulting for Systap, LLC, before the start of the project resulted in **three patents** (one of them still pending) during the time of project [39, 40, 41].
- In 2021, Olaf Hartig was selected for the prestigious **Amazon Scholar** program. In this role, he is now working at 10% with the Neptune graph database team at Amazon Web Services where his research on interoperability of graph data models directly affects the development of the Neptune system [2] and will become available to thousands of customers.

6 Connections with Other CENIIT Projects

We have collaborated with Eva Blomqvist (CENIIT project 12.10) to jointly supervise the PhD student Robin Keskisärkkä who successfully graduated in 2021 and to form a (virtual) group of researchers at IDA in the area of Semantic Web. In the context of this group, we established a biweekly PhD colloquium for PhD students in the area (which is still ongoing) and co-organized the 2017 edition of the annual "Linked Data in Sweden" day.

7 Creation of a New Research Group

The CENIIT funding has enabled Olaf Hartig to start building up a research group which currently consists of two PhD students and one postdoc-level technician. At some points during the project, the group also included 1–2 students who were employed as amanuens to work on some project-related tasks. In addition to using the CENIIT funding directly for some members of the group, the project also enabled Olaf Hartig to attract further funding for the group (in the form of a starting grant from the Swedish Research Council).

8 Publications and Patents

The following publications have been produced during the project.

Refereed Journal Articles

[1] Patrick Lambrix, Rickard Armiento, Huanyu Li, **Olaf Hartig**, Mina Abd Nikooie Pour, and Ying Li. "The Materials Design Ontology." In: *Semantic Web Journal* (2023)

[2] Ora Lassila, Michael Schmidt, **Olaf Hartig**, Brad Bebee, Dave Bechberger, Willem Broekema, Ankesh Khandelwal, Kelvin Lawrence, Carlos Manuel Lopez Enriquez, Ronak Sharda, and Bryan Thompson. "The OneGraph Vision: Challenges of Breaking the Graph Model Lock-In." In: *Semantic Web Journal* 14.1 (2023)

[3] Sherif Sakr et al. "The Future Is Big Graphs: A Community View on Graph Processing Systems." In: *Communications of the ACM* 64.9 (2021)

[4] Sijin Cheng and **Olaf Hartig**. "OPT+: A Monotonic Alternative to OPTIONAL in SPARQL." in: *Journal of Web Engineering* (2019)

[5] **Olaf Hartig** and Giuseppe Pirrò. "SPARQL with Property Paths on the Web." In: *Semantic Web Journal* 8.6 (2017)

Refereed Conference Papers

[6] Sijin Cheng and **Olaf Hartig**. "Source Selection for SPARQL Endpoints: Fit for Heterogeneous Federations of RDF Data Sources?" In: *Proceedings of the 6th Workshop on Storing, Querying and Benchmarking Knowledge Graphs (QuWeDa)*. 2022

[7] Sijin Cheng and **Olaf Hartig**. "LinGBM: A Performance Benchmark for Approaches to Build GraphQL Servers." In: *Proceedings of the 23rd Int. Conf. on Web Information Systems Engineering (WISE)*. 2022

[8] Shahrzad Khayatbashi, Sebastián Ferrada, and **Olaf Hartig**. "Converting Property Graphs to RDF: A Preliminary Study of the Practical Impact of Different Mappings." In: *Proc. of the Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA)*. 2022

[9] Sijin Cheng and **Olaf Hartig**. "Towards Query Processing over Heterogeneous Federations of RDF Data Sources." In: *Proc. of Posters and Demos at the 19th Extended Semantic Web Conference (ESWC)*. 2022

[10] Robin Keskisärkkä, Eva Blomqvist, and **Olaf Hartig**. "Optimizing RDF Stream Processing for Uncertainty Management." In: *Proceedings of the 16th SEMANTiCS Conference*. 2021
Note: This paper was awarded the **Best Paper Award** in this conference.

[11] Sijin Cheng and **Olaf Hartig**. "FedQPL: A Language for Logical Query Plans over Heterogeneous Federations of RDF Data Sources." In: *Proceedings of the 22nd Int. Conf. on Information Integration and Web-based Applications & Services (iiWAS)*. 2020
Note: This paper was awarded the **Best Paper Award** in this conference.

[12] Robin Keskisärkkä, Eva Blomqvist, Leili Lind, and **Olaf Hartig**. "Capturing and Querying Uncertainty in RDF Stream Processing." In: *Proceedings of the 22nd Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW)*. 2020

[13] **Olaf Hartig**. "Foundations to Query Labeled Property Graphs using SPARQL*." In: *Proceedings of the 1st Int. Workshop on Approaches for Making Data Interoperable (AMAR)*. 2019

[14] Robin Keskisärkkä, Eva Blomqvist, Leili Lind, and **Olaf Hartig**. "RSP-QL*: Enabling Statement-Level Annotations in RDF Streams." In: *Proceedings of the 15th SEMANTiCS Conference*. 2019
Note: This paper was awarded the **Best Paper Award** in this conference.

[15] **Olaf Hartig** and Jan Hidders. "Defining Schemas for Property Graphs by using the GraphQL Schema Definition Language." In: *Proc. of the 2nd Joint Int. Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) (GRADES-NDA'19) at ACM SIGMOD*. 2019

[16] Yun Wan Kim, Mariano P. Consens, and **Olaf Hartig**. "An Empirical Analysis of GraphQL API Schemas in Open Code Repositories and Package Registries." In: *Proceedings of the 13th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)*. 2019

[17] Muhammad Saleem, Alexander Potocki, Tommaso Soru, **Olaf Hartig**, and Axel-Cyrille Ngonga Ngomo. "CostFed: Cost-Based Query Optimization for SPARQL Endpoint Federation." In: *Proc. of the 14th SEMANTiCS Conference*. 2018
Note: This paper was awarded the **Best Paper Award** in this conference.

[18] **Olaf Hartig**, Ian Letter, and Jorge Pérez. "A Model of Distributed Query Computation in Client-Server Scenarios on the Semantic Web." In: *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI)*. 2018

[19] **Olaf Hartig** and Jorge Pérez. "Semantics and Complexity of GraphQL." in: *Proc. of the 27th International World Wide Web Conference (WWW)*. 2018

[20] **Olaf Hartig**, Ian Letter, and Jorge Pérez. "A Formal Framework for Comparing Linked Data Fragments." In: *Proc. of the 16th International Semantic Web Conference (ISWC)*. 2017
Note: This paper was awarded the **Best Research Paper Award** in this conference.

[21] Alexander Potocki, Muhammad Saleem, Tommaso Soru, **Olaf Hartig**, and Axel-Cyrille Ngonga Ngomo. "Federated SPARQL Query Processing via CostFed." In: *Proc. of the 16th International Semantic Web Conference (ISWC)*. 2017 (demo paper)

[22] **Olaf Hartig**. "RDF* and SPARQL*: An Alternative Approach to Annotate Statements in RDF." in: *Proc. of the 16th International Semantic Web Conference (ISWC)*. 2017 (poster paper)
Note: This work was the **People's Choice Best Poster Award** in this conference.

[23] **Olaf Hartig**. "Foundations of RDF* and SPARQL* – An Alternative Approach to Statement-Level Metadata in RDF." in: *Proc. of the 11th Alberto Mendelzon Workshop on Foundations of Data Management (AMW)*. 2017

[24] **Olaf Hartig** and Jorge Pérez. "An Initial Analysis of Facebook's GraphQL Language." In: *Proc. of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW)*. 2017

[25] **Olaf Hartig** and Olivier Curé. "Semantic Data Management in Practice." In: *Companion Volume of the Proceedings of the 26th International Conference on World Wide Web (WWW)*. 2017 (tutorial)

Book Chapters

[26] **Olaf Hartig**, Katja Hose, and Juan F. Sequeda. "Linked Data Management." In: *Encyclopedia of Big Data Technologies.* Springer International Publishing, 2019

[27] Maribel Acosta, **Olaf Hartig**, and Juan F. Sequeda. "Federated RDF Query Processing." In: *Encyclopedia of Big Data Technologies.* Springer International Publishing, 2019

Edited Proceedings

[28] **Olaf Hartig** and Oshani Seneviratne, eds. *Proceedings of the Doctoral Consortium at ISWC 2022, co-located with the 21st International Semantic Web Conference (ISWC)*. CEUR Workshop Proceedings, Vol-3165. 2022

[29] Ladjel Bellatreche, Marlon Dumas, Panagiotis Karras, Raimundas Matulevicius, Ahmed Awad, Matthias Weidlich, Mirjana Ivanovic, and **Olaf Hartig**, eds. *New Trends in Database and Information Systems - ADBIS 2021 Short Papers, Doctoral Consortium and Workshops: DOING, SIMPDA, MADEISD, MegaData, CAoNS, Tartu, Estonia, August 24-26, 2021, Proceedings*. Vol. 1450. Communications in Computer and Information Science. Springer, 2021. ISBN: 978-3-030-85081-4

[30] Andreas Harth, Valentina Presutti, Raphaël Troncy, Maribel Acosta, Axel Polleres, Javier D. Fernández, Josiane Xavier Parreira, **Olaf Hartig**, Katja Hose, and Michael Cochez, eds. *The Semantic Web: ESWC 2020 Satellite Events - ESWC 2020 Satellite Events, Heraklion, Crete, Greece, May 31 - June 4, 2020, Revised Selected Papers*. Vol. 12124. Lecture Notes in Computer Science. Springer, 2020. ISBN: 978-3-030-62326-5

[31] Chiara Ghidini, **Olaf Hartig**, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, eds. *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*. vol. 11778. Lecture Notes in Computer Science. Springer, 2019. ISBN: 978-3-030-30792-9

[32] Chiara Ghidini, **Olaf Hartig**, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, eds. *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*. vol. 11779. Lecture Notes in Computer Science. Springer, 2019. ISBN: 978-3-030-30795-0

[33] Pascal Hitzler, Sabrina Kirrane, **Olaf Hartig**, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, eds. *The Semantic Web: ESWC 2019 Satellite Events, Portorož, Slovenia, June 2-6, 2019, Revised Selected Papers*. Vol. 11762. Lecture Notes in Computer Science. Springer, 2019. ISBN: 978-3-030-32326-4

[34] Matthias Wauer, Mohamed Ahmed Sherif, Muhammad Saleem, **Olaf Hartig**, Ricardo Usbeck, Ruben Verborgh, and Axel-Cyrille Ngonga Ngomo, eds. *Joint Proceedings of the 3rd International Workshop on Geospatial Linked Data (GeoLD) and the 2nd Workshop on Querying the Web of Data (QuWeDa), co-located with the 15th Extended Semantic Web Conference (ESWC)*. CEUR Workshop Proceedings, Vol-2110. 2018

[35] Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and **Olaf Hartig**, eds. *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*. vol. 10249. Lecture Notes in Computer Science. 2017. ISBN: 978-3-319-58067-8

[36] Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and **Olaf Hartig**, eds. *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part II*. vol. 10250. Lecture Notes in Computer Science. 2017. ISBN: 978-3-319-58450-8

[37] Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and **Olaf Hartig**, eds. *The Semantic Web: ESWC 2017 Satellite Events*. Vol. 10577. Lecture Notes in Computer Science. 2017

Other Publications

[38] **Olaf Hartig**, Pierre-Antoine Champin, Gregg Kellogg, and Andy Seaborne, eds. *RDF-star and SPARQL-star*. W3C Community Group Report. Dec. 2021

[39] **Olaf Hartig**, Maria-Esther Vidal, and Johann-Christoph Freytag. "Federated Semantic Data Management (Dagstuhl Seminar 17262)." In: *Dagstuhl Reports* 7.6 (2017), pp. 135–167

Patents

[40] Bradley R. Bebee, Bryan B. Thompson, Thomas James Lewis, and Olaf Hartig. *Acceleration Techniques for Graph Analysis Programs*. U.S. Patent 20220050666, status: pending

[41] Bradley R. Bebee, Bryan B. Thompson, Thomas James Lewis, and Olaf Hartig. *Acceleration Techniques for Graph Analysis Programs*. U.S. Patent 11200032, December 14, 2021

[42] Bradley R. Bebee, Bryan B. Thompson, Thomas James Lewis, and Olaf Hartig. *Acceleration Techniques for Graph Analysis Programs*. U.S. Patent 10409560, September 10, 2019

Appendix: List of Master's Theses

The Master's theses that are marked with an asterisk (*) have been directly related to the research topics of the project.

1. * Nuha Batool Taqi *"Evaluation of Tools to Generate a GraphQL-Accessible Database System"* (ongoing).
2. * Juliette Winneroth *"Using Vocabulary Mapping for Federated RDF Query Processing"* (ongoing).
3. * Albin Pettersson *"An Investigation of Source Selection Approaches for Federated Query Processing"* (ongoing).
4. * Johan Lindberg *"Server-side Batching and Caching for GraphQL Schema Delegation"* (ongoing).
5. * Wiktor Karlsson *"Integrating GraphQL APIs into an RDF Federation"* (2023).
6. * Markus Gustafsson and Pontus Jarnemyr *"A Declarative Approach to GraphQL Schema Wrapping"* (2022).
7. Alexandra Goltsis *"A Performance Comparison of SQL and NoSQL Databases for 5G Radio Base Station Configuration"* (2022).
8. * Muhammad Ismail *"Evaluation of Generic GraphQL Servers for Accessing Legacy Databases"* (2022).
9. Tim Hellberg and Filip Ström *"An Analysis of Cloud Services for an IoT Scenario"* (2021).
10. Karl Söderbäck *"Organizing HLA Data for Improved Navigation and Searchability"* (2021).
11. * Linn Mattsson *"Implementing Facebook's GraphQL Interface on top of a Graph Database"* (2021).
12. * Lukas Lindqvist *"A Study of Existing Techniques for Building an Efficient GraphQL Server"* (2020).
13. * Robin Christensen *"An Analysis of Notions of Differential Privacy for Edge-Labeled Graphs"* (2020).
14. * Jonatan Pålsson *"Querying Federations of Eiffel Event Data Repositories"* (2020).
15. * Andreas Lundquist *"Combining Result Size Estimation and Query Execution for the GraphQL Query Language"* (2020).
16. Rickard Hellenberg *"Implementing a Database Abstraction Layer and Evaluating Databases for the Eiffel Event Persistence Solution"* (2019).
17. Joakim Ericsson *"Object Migration in a Distributed, Heterogeneous SQL Database Network"* (2018).
18. John Bengtson and Christoffer Nilsson *"Combining MongoDB and Elasticsearch to Achieve a Highly Available Search Engine"* (2017).
19. Gustav Stenhag *"Defining a power system's historical operating states – The design, implementation and use of a computer-based framework"* (2017).