ISSN 2614-0047

Knowledge graph completion for scholarly knowledge graph

Taufiqurrahman [a,1], Kemas Rahmat Saleh Wiharja [a,2,*], Gia Septiana Wulandari [a,3]

[a] School of Computing, Telkom University, Jalan Telekomunikasi No. 1, Terusan Buah Batu-Bojongsoang, Kabupaten Bandung 40257, Indonesia
[1] taufiqurrahmantopiq@student.telkomuniversity.ac.id; [2] bagindokemas@telkomuniversity.ac.id; [3] giaseptiana@telkomuniversity.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 11, 2023
Revised February 2, 2024
Accepted September 7, 2024

Keywords
Knowledge graph
Scholarly knowledge graph
Knowledge graph
Completion

ABSTRACT

Scholarly knowledge graph is a knowledge graph that is used to represent knowledge contained in scientific publication documents. The information we can find in a scientific publication document is as follows: author, institution, name of journal/conference, and research topic. A knowledge graph that has been built is usually still not perfect. Some incomplete information may be found. To add the missing information, we can use knowledge graph completion, which is a method for finding missing or incorrect relationships to improve the quality of a knowledge graph. Knowledge graph completion can be carried out on a scholarly knowledge graph by adding new entities and relationships to produce further information in the scholarly knowledge graph. The data added to the scholarly knowledge graph are only other papers of first author entity, the research field of first author entity, and a description of the conference/journal entity. The result shows that the scholarly knowledge graph was completed by adding 81% correct data for other papers of first author entity, 80.3% correct data for the research field of first author entity, and 53.9% correct data for the description of the conference/journal.

1. Introduction

A knowledge graph is a representation of structured relational information in the form of Entities and the relationships between them [1]. The graph is a multi-rational graph in which entities and relationships are connected. An entity is a real object or abstract information. Entity and relation are represented as triple [2]–[8]. For example, ("Bandung", "capitalOf", West Java) is a triple. "Bandung" and "West Java" are entities that are connected by a relation named "capitalOf". A knowledge graph is formed from a combination of knowledge contained in a knowledge base or the results of a reasoning system from the knowledge contained within it [9]–[12].

In 2012, Google introduced a knowledge graph on their finding machine as semantic evolved of it. The knowledge graph makes their machine able to find a more real-world object. Since 2012, similar applications have also used knowledge graphs, such as DBPedia, YAGO (Yet Another Great Ontology), Freebase, and Wikidata [1], [13]–[16]. Knowledge graphs can be constructed with Natural Language Programming (NLP) [17]–[21] and modules from graph databases such as Ontotext GraphDB [22]–[25], Neo4j [26]–[28], and Cayley.

A knowledge graph that has been constructed is usually still not perfect [29]. Some information still needs to be completed and can still be added. We can complete the information we have in a knowledge graph by using knowledge graph completion, a method for finding missing or wrong parts to improve the knowledge graph quality [29]. With this method, a knowledge graph can have more information on it.

Besides representing and managing common information and encyclopedias, a knowledge graph can also be used to manage scholarly data or bibliography, and it is called a scholarly knowledge graph

[30]–[32]. The scholarly data can be obtained from papers. Data that can be found are authors, institution, journal/conference, and topic of research [33]. Until now, there have been relatively few research endeavors that employ knowledge graphs for scholarly data [34]. Some of them are [33] and [35].

Scholar data at universities in Indonesia has been recorded by Directorate of Research and Community Service (known as Direktorat Penelitian dan Pengabdian Masyarakat or PPM) in the form of an Excel file. The data contains the title of the papers, the author of the papers, and the research group of the author. Research group data describes the research focus of the lecturers in conducting their research. However, there are areas for improvement in the data. Even though lecturers' research fields are described in terms of their research groups, some still conduct research that does not match the lecturers' research groups. So, it is still difficult to determine the actual focus of lecturers' research fields.

On the other hand, Google Scholar, one of the most complete services related to researchers' data and bibliography on the internet [36], [37], cannot accommodate the categorization of researchers' research fields and the focus of research fields applied at each university. These conditions make it difficult for PPM to identify suitable lecturer research fields to be included in a research scheme or grant. Therefore, this research aims to help PPM overcome this problem by implementing knowledge graph completion on the scholarly knowledge graph of universities in Indonesia.

There is usually some null data in scholarly data, so the data becomes meaningless after being converted into a knowledge graph. In the Excel data used, null data is data that only contains the "-" symbol. For example, not all papers have a third author. As a result, after being converted into a knowledge graph, an entity with the value "-" is defined as the third author who wrote a paper ("-", "authorOf", "paper"). The null data needs to be removed from the knowledge graph. Otherwise, the triple data, like the previous example, is useless and becomes incorrect information in the knowledge graph and affects the visualization of the knowledge graph. For example, a triple ("-", "authorOf", "paper") could mean that an author entity "-" has written a paper, even though the author entity does not actually exist. To avoid this kind of misleading information, null data and its relationship with other entities will be removed using a special graph framework, GP2 [38], [39], which can remove null data directly from graphs.

There are several approaches to knowledge graph completion, like exBERT [34], [40], [41], CoDEx [42]–[45], and TransE [46]–[49]. Of the three knowledge graph completion models above, there are several shortcomings in completing the scholarly data used in this research. First, the PPM data in Excel form does not match the input data type of the three models, namely RDF. Second, exBERT and TransE carry out completion by looking for new data from the entities in the scholarly knowledge graph itself, while for problems with the PPM data other data is needed to complete the existing gaps. Third, all these models also do not check for null data contained in the knowledge graph. Meanwhile, the PPM data used still contains null data, which has no meaning.

The contributions of this paper are: (i) We develop a framework that can complete the scholarly knowledge graph. (ii) We identified that not all sources related to researchers' data and bibliography on the internet have complete data. (iii) We use a special graph framework GP2 to remove null data in knowledge graph.

2. Method

2.1. Scholarly Knowledge Graph Construction

The dataset used for this scholarly knowledge graph is scientific paper publication data within the Informatics Study Program at Telkom University in 2018. The type of the dataset is CSV file with 503 rows and 21 columns. The data contained in it includes the title of the paper, the author of the paper, the author's research group, and the conference/journal publication. The dataset is collected every year by the Telkom University Research Directorate. Beside this data, we also collect two other data such as follows: First, the final assignment paper submission data in the Telkom University Open Library application. Second, the files from all research groups at Telkom University that are filled by the lecturers every semester. This research takes data from the Telkom University Open Library application using the API. There is no biased data in the dataset because there is no omitted data or filtered data.

We build the knowledge graph using Ontotext GraphDB and Ontotext Refine applications [50], [51]. From the global point of view, there are two main steps as follows: First, using Ontotext Refine for mapping the data into triple format. Second, using Ontotext GraphDB for converting the triple format into a knowledge graph according to previously created mapping as show in Fig. 1.

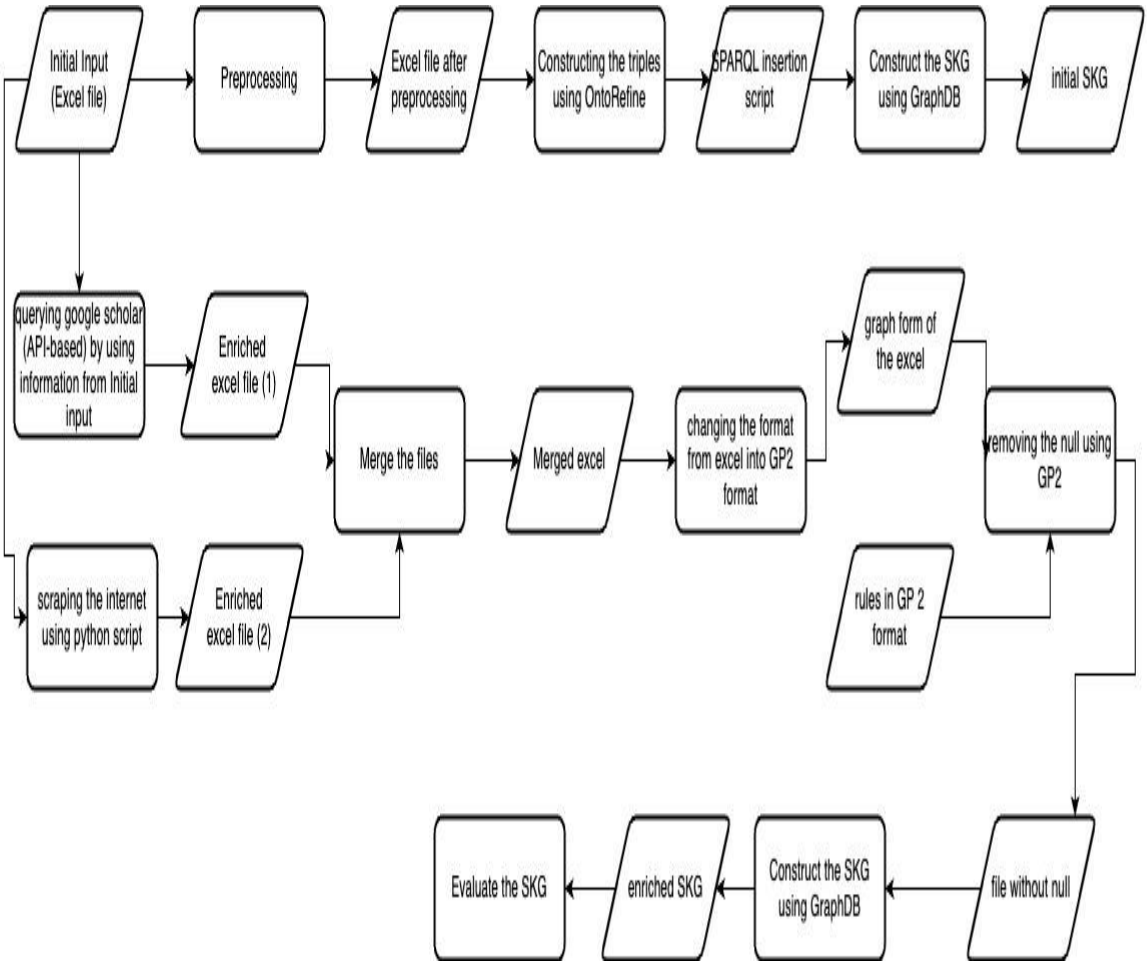

Fig. 1. Knowledge graph completion process

In more detail, the steps are as follows:

- Preprocessing

  In the preprocessing process, several things are done, such as removing unnecessary letters, punctuation marks and numbers. As well as replacing all existing spaces with "_" signs so that the data is easy to be read in Ontotext GraphDB. This is because Ontotext GraphDB will automatically change the spaces in the dataset into "%20".

- Construct Triple

  The first thing to do in this step is to use Ontotext Refine for mapping the dataset into triple format. Each table contained in the dataset will become an entity in the triple. Next, connect each table to another table that has a relationship with that table. Entities we used are named in Indonesian language. “Penulis” means author, “Makalah/Paper” means article, “Konferensi/Jurnal” means conference or journal, and “Kelompok Keahlian” means research group. For relations, “menulisMakalah” means author of, “memilikiKK” is part of research group, “judulMakalah” means title of the paper, then “nama”, “tempat”, “tanggal” mean name, place, and date (of conference or journal). For example, we connect the author entity with the article and research group entities because both can form a triple relationship, namely author of article and author is part of research group. The triples in scholarly knowledge graph as show in Table 1.

Table.1 The triples in scholarly knowledge graph

No	Entity	Relation	Entity
1	Penulis 1, Penulis 2, Penulis 3, Penulis 4,dan PenulisTambahan	menulisMakalah	Makalah
2	Penulis 1, Penulis 2, Penulis 3, Penulis 4,dan PenulisTambahan	memilikiKK	Kelompok Keahlian
3	Makalah/Paper	judulMakalah	Judul Makalah/Paper
4	Makalah/Paper	abstrak	Abstrak
5	Makalah/Paper	jenisMakalah	Jenis Publikasi Makalah/Paper
6	Makalah/Paper	dipublikasikanDi	Konferensi/Jurnal
7	Konferensi/Jurnal	namaKonferensi	Nama Konferensi/Jurnal
8	Konferensi/Jurnal	website	Alamat Website Konferensi/Jurnal
9	Konferensi/Jurnal	tempatKonferensi	Tempat Konferensi/Jurnal
10	Konferensi/Jurnal	tanggalKonferensi	Tanggal Konferensi/Jurnal

- Construct Scholarly Knowledge Graph

  Next, using Ontotext GraphDB, the dataset is converted into a knowledge graph according to previously created mapping. From Ontotext Refine, we will get a SPARQL query that is ready to run on Ontotext GraphDB. The results of the knowledge graph that has been built can be seen with visualization from Ontotext GraphDB. The visualization capabilities of Ontotext GraphDB helps us in seeing the data more clearly. Data and the relationships among them become more visible than just seeing the data in its original form after it becomes a graph. Some data is in the form of a URL, whereas in the visualization, the data is displayed directly in the form of a string. The visualization of knowledge graph as show in Fig. 2.

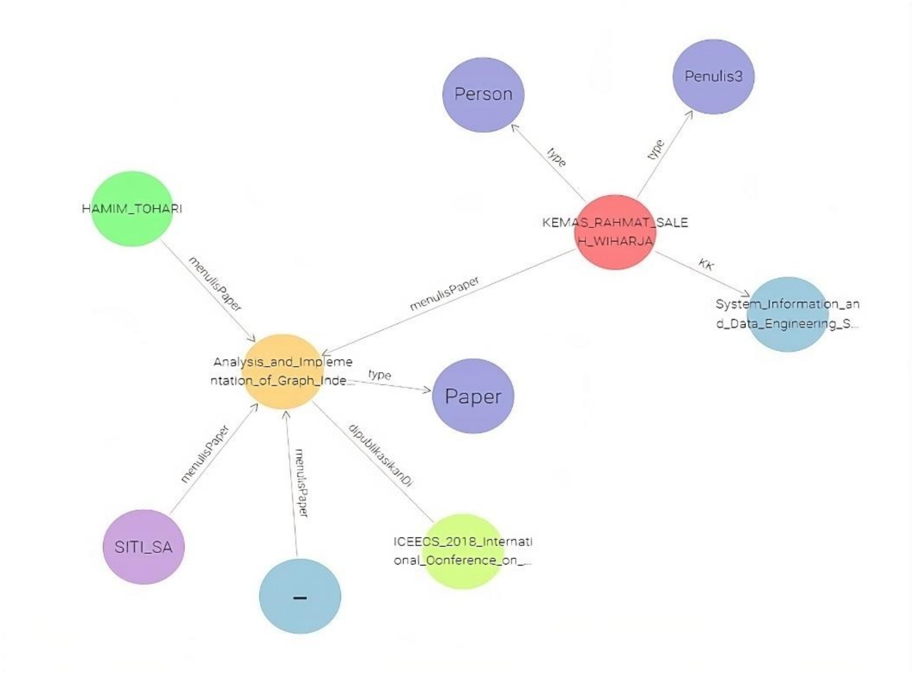

Fig. 2. The visualization of knowledge graph

2.2. Knowledge Graph Completion Process

Knowledge graph completion is carried out by adding information about other papers and research fields related to the first author from Google Scholar. Additional information added is a description of the journal or conference in the knowledge graph. This information is obtained from web scraping the journal or conference website address.

- Knowledge Graph Completion by Retrieving Information from Google Scholar

  In the Google Scholar search feature, retrieving data from Google Scholar is done by searching for the name of the author of each paper in the dataset. All data on the first author's name in the dataset is used as a search query with the words "Telkom University" added behind it. So, the query you are looking for is "name of first author Telkom University". The data taken is the paper's title and the first author's research field that appears from the search results. So, in the knowledge graph, the entity of the first author will have new information about other papers and the author's research field. The data retrieval process uses go language programming code. This stage's output is a dataset equipped with search result data from Google Scholar.

- Knowledge Graph Completion by Web Scraping

  In this process, the data that we want to retrieve is the description of a conference or journal contained in the dataset so that later, the conference or journal entity will have new information regarding the description of the entity. The web scraping process is carried out by using website address data for each conference or journal in the dataset. The web scraping technique used is HTML parsing by using Python programming code with the BeautifulSoup library. In the HTML parsing process, data is taken from the HTML tags 'h1', 'div', 'p', 'text', and 'span'. This tag was chosen because, generally, the text contains a description of the conference or journal website address in that tag. Then from the data in the tag, only data that contains the words "journal" or "conference" will be retrieved. The output from this stage is a dataset that has been equipped with web scraping data.

2.3. Removing Null Data

The process of removing null data is carried out by utilizing graph transformation in GP2. First, the data in the dataset is converted into a syntax format acceptable to GP2. The dataset that is converted into GP2 syntax is only entity data that has null data. When converting a dataset to GP2 syntax, the value of an entity is changed to the text “Entity name is not null”. This is done because of GP2's limitations in storing text and entity values, which are also not needed in the process of removing null data. The replaced value after the null data has been removed will be changed back to its original value.

The syntax format for entities (nodes) is "(node id, data)", and the syntax format for relationships (edges) is "(edge id, node id, node id, relationship name)". The output of this process is a host-type file whose data is a graph of entities and their relationships. This file becomes one of the inputs for the running process on GP2. Another input is a file of type gp2, which contains rule syntax to transform the input graph into a graph without null data. The rule applied to remove null data is from the input graph contained in the previous host type file, looking for all nodes that have "-" data. Then, all edges connected to that node are removed. After that, all nodes that have "-" data are removed from the knowledge graph.

Next, after the running process is carried out on GP2, we will get an output file of text type which contains a graph that no longer has null data. The data in the file is then converted back from GP2 syntax format into a CSV type dataset so that it can be processed again by Ontotext Refine and Ontotext GraphDB.

2.4. Feading the Data into the Scholarly Knowledge Graph

The dataset after the completion process that no longer contains null data is returned to Ontotext Refine. Then, the data is mapped again by importing the mapping results that were created during the previous stage of developing the scholarly knowledge graph, plus new mapping data from the completion results to make it a triple. Triples added from data completion as show in Table 2.

Table.2 Triples added from data completion

No	Entity	Relation	Entity
1	Penulis 1	makalahLain	Journals
2	Penulis 1	memilikiBidang	Research Fields
3	Konferensi/Jurnal	memilikiDeskripsi	Deskripsi Konferensi/Jurnal

Next, the mapping results are feeded into Ontotext GraphDB to be constructed into a scholarly knowledge graph, which has been completed. The new data resulting from knowledge graph completion can be seen in Fig. 3 and Fig. 4. The graph also no longer has null data, as in Fig. 2.

	nama	paperLain	researchFields
1	"ABDUL_AZIZ"	"["Wireless-powered sensor networks: How to realize","Distributed wireless power transfer system for Internet of Things devices","Simultaneous wireless information and power transfer (SWIPT) for Internet of Things: Novel receiver design and experimental validation","Theory and experiment for wireless-powered sensor networks: How to keep sensors alive","Toward realization of long-range wireless-powered sensor networks","Battery-less location tracking for I	"["RF Wireless Power Transfer","RF Circuit Design","Internet of Things"]"

Fig. 3. The example of data completion results from google scholar

	KonferensiatauJurnal	Deskripsi
1	"International_Conference_on_Data_and_Information_Science"	"['All participants and parties who supported this conference']"
2	"International_Journal_on_Advanced_Science,_Engineering_and_Information_Technology"	"['International Journal on Advanced Science, Engineering and Information Technology', 'International Journal on Advanced Science, Engineering and Information Technology (IJASEIT) is an international peer-reviewed journal dedicated to interchange for the results of high quality research in all aspect of science, engineering and information technology. The journal publishes state-of-art papers in fundamental theory, experiments and simulation, as well as applications, with a systematic proposed method, sufficient review on previous works, expanded discussion and concise conclusion. As our commitment to the advancement of science and technology, the IJASEIT follows the open access policy that allows the published articles freely available online without any subscription.']"

Fig. 4. The example of data completion results from web scraping

The visualization of the knowledge graph after the completion process looks like in Fig. 5.

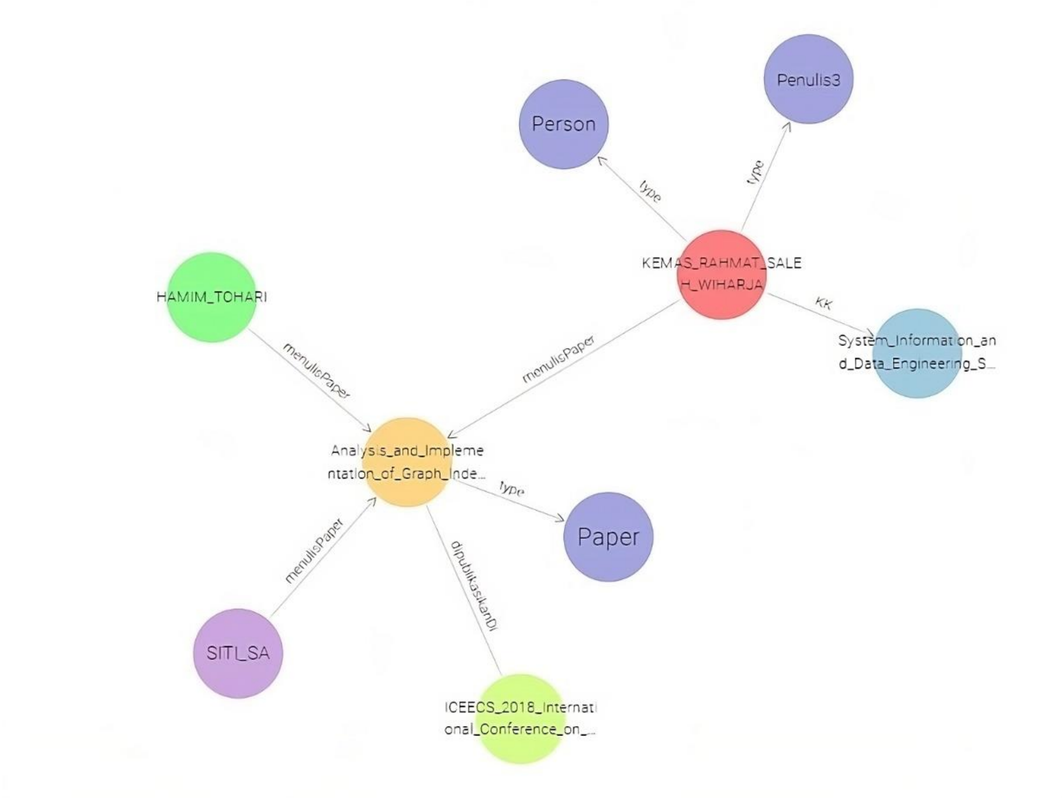

Fig. 5. The visualization of knowledge graph after completion

2.5. Evaluation of Knowledge Graph Completion

The knowledge graph completion process carried out will be evaluated by measuring the coverage and correctness of the completed scholarly knowledge graph [52], [53]. Evaluation of coverage is carried out by comparing the completed scholarly knowledge graph data and the data source. Next, the correctness evaluation is carried out by comparing the knowledge graph completion data with comparable external data.

3. Results and Discussion

3.1. Data Coverage

The evaluation of data coverage is carried out by comparing the results of the scholarly knowledge graph data query after the completion process with the data source CSV file. In the experiment, a random sample of data was taken, each with triple data showing the relationship between The first author and Journals containing other papers related to The first author (“Penulis 1”, “makalahLain”, “Journals”); The relationship between The first author and the Research field (“Penulis 1”, “memilikiBidang”, “Research Fields”); and the relationship between the conference/journal and the description of the conference/journal (“Konferensi/Jurnal”, “memililkiDeskripsi”, “Deskripsi Konferensi/Jurnal”). Each triple data mentioned before has 50 random data. The process of querying the scholarly knowledge graph is carried out in the Ontotext GraphDB application using SPARQL as show in Fig. 6.

	NamaPenulis1	MakalahLain
1	"ARIE_ARDIYANTI_SURYANI"	"['Word2vec model analysis for semantic similarities in english words','Cancer detection based on microarray data classification using pca and modified back propagation','NusaCrowd: Open Source Initiative for Indonesian NLP Resources','The rule-based sundanese stemmer','Experiment on a phrase-based statistical machine translation using PoS Tag information for Sundanese into Indonesian','Analisis Sentimen Berbahasa Indonesia dengan Pendekatan Lexicon-Based pada Media Sosial','Enriching English into Sundanese and Javanese translation list using pivot language','Using lstm for context based ap

(a)

	NamaPenulis1	ResearchFields
1	"RINI_HANDAYANI"	"['signal processing','embedded system','computer network']"
2	"FHIRA_NHITA"	"['Data Mining','Artificial Intelligence']"
3	"AHMAD_MUSNANSYAH"	"['Software Engineering','Cloud Computing/Manufacturing','Machine Learning']"

(b)

	KonferensiatauJurnal	Deskripsi
1	"ICCSI"	"['The 7th International Conference on Computer Science and Computational Intelligence (ICCSCI) is annual forum for researchers, and scientist to disseminate their knowledge and research on Computer Science, Computational Intelligence and Information Technology. The conference warmly welcomes prospected authors to submit their research and idea to ICCSCI 2022 and share the valuable experiences with the scientist and scholars around the world. ICCSCI 2022 is organized by School of Computer Science, and will be held on:', 'All accepted papers in ICCSCI will be published in the conference proceedings and will be submitted for publication index by Scopus']"

(c)

Fig. 6. (a) Example of random data for The first author and Journals containing other papers related to The first author, (b) Example of random data for The first author and the Research field, (c) Example of random data for The conference/journal and The description of the conference/journal

The data samples that have been taken from the knowledge graph are then compared with the data in the CSV file data source. The coverage measured by calculating the amount of data found that is the same, compared to the number of data samples.

We can see from Table 3, after comparing the data, the data obtained from the knowledge graph query results can also be found in the CSV file data source. So, from the test results, the data completeness (coverage) value is 100%. This means all data from the CSV file data source is already contained in the scholarly knowledge graph.

Table.3 Data coverage testing results

No	Triple	Result	Percentage
1	(Penulis 1, makalahLain, Journals)	All data is same	100%
2	(Penulis 1, memilikiBidang, Research Fields)	All data is same	100%
3	(Konferensi/Jurnal, memililkiDeskripsi, Deskripsi Konferensi/Jurnal)	All data is same	100%

3.2. Data Correctness

- *Other Related Papers from the First Author*
  Other related papers from The first author that have been included in the scholarly knowledge graph are searched for the title of the paper on dblp (https://dblp.uni-trier.de) and Scopus (https://scopus.com) using the API service from both sites. The comparison method carried out with dblp is to take author data contained in the paper search results data. The author's data is compared with the paper author's data in the scholarly knowledge graph. If the author's data is the same, then the completion data is correct. From the results of the comparison with dblp, it was found that 25% of the data was correct and that the added paper data had the same author data in the scholarly knowledge graph and in dblp. The factor for the very low percentage is the large number of papers that have yet to be recorded on the dblp site, while the completed data comes from Google Scholar, which has more paper data as show in Table. 4.

Table.4 Correctness testing results of other related papers from the first author data

No	Comparison	Percentage
1	dblp	25%
2	Scopus	81%

Because the percentage could have been better, the data was also compared with other sources, namely Scopus. The method used is almost the same. In the results of searching for papers in Scopus, the data of the paper's author is taken and then compared with the author's data in the scholarly knowledge graph. However, the data obtained from Scopus only has the data of the first author of the paper being searched, while the data of the authors to be compared could be the second author, third author, and so on. So, another condition is added here, namely, if the search result paper is affiliated with Telkom University, then it is considered correct. From the comparison results with Scopus, 81% of the data was correct, and the paper data added had the same author data in the scholarly knowledge graph and in Scopus. The better factor in the comparison percentage with Scopus is because more paper data overlaps between Google Scholar and Scopus. So, comparing data on scholarly knowledge graphs taken from Google Scholar is more effective with Scopus than dblp.

- Research Field from the First Author

  The research field data from each first author, that is already contained in the scholarly knowledge graph, is compared with the research field data for each author contained in Sinta Kemdikbud (https://sinta.kemdikbud.go.id). The comparison was carried out using web scraping to the search results page for the author's name. Furthermore, from the scraping results, data containing the author's research field was taken. The data is then compared with data from the scholarly knowledge graph. If the author's research field on the scholarly knowledge graph is the same as that obtained from the Sinta Kemdikbud website, then the data is considered correct. The comparison results found that 80.3% of the data was correct that the first author in the scholarly knowledge graph had a research field like the one on the Sinta Kemdikbud website. This large amount of correct data is because the author's research field data in Sinta Kemdikbud quite closely overlaps with that in Google Scholar, which is a source of completion data for field research.

- Description of Journal/Conference

  The conference/journal description data obtained from the website data in the scholarly knowledge graph is compared with the conference/journal name data in the scholarly knowledge graph. The comparison process is carried out using the API service from OpenAI. The comparison is done by querying OpenAI to check whether the description text truly describes the relevant conference/journal. OpenAI will give “yes” if correct and “no” if not. The comparison results found that 53.9% of the description data matched the name of the conference/journal. Several factors cause some descriptive data that does not match the conference or journal. First, the limitations of data retrieval through web scraping, which is only based on HTML tags from conference/journal sites, make the data retrieved inappropriate. Second, there is a conference/journal website address in the scholarly knowledge graph that does not match the name of the conference/journal. The second factor is that the name of the existing conference/journal does not match the website address of the conference/journal. The third factor, OpenAI, which is used to compare, is also partially correct. The use of OpenAI as a tool to compare whether the description data matches the name of the conference/journal was carried out because we had not found comparable data that had a similar form to the description data in this scholarly knowledge graph. Retrieving data by scraping directly to the conference/journal site is still not optimal. It would be better if scraping could directly identify data that is actually a description of the conference/journal, not just based on html tags from the scraping results.

- Null Data

  Evaluation of null data is carried out by querying the scholarly knowledge graph to check whether there is still null data. The way to check this is to look for triples in the scholarly knowledge graph that have the entity "-" in the subject or object position. The data query process for the scholarly knowledge graph is carried out in the Ontotext GraphDB application using SPARQL. After carrying out the query, the result obtained is "no result". This means that there are no triples with the entity "-" in the subject or object position. Thus, the percentage of null data that is successfully removed is 100%. Based on the percentage results, it was found that GP 2 can be used to remove null data in the scholarly knowledge graph. But, the condition is that a knowledge graph needs to be changed first according to the GP 2 syntax, and if necessary, it is changed back to the previous form so that queries and visualization of the scholarly knowledge graph can be carried out.

4. Conclusion

This research develops a framework that can complete the scholarly knowledge graph. We successfully get 100% for the value of the coverage parameter. This means that the scholarly knowledge graph contains all data from CSV files of scientific paper publications within the Informatics Study Program at Telkom University in 2018, Google Scholar, and web scraping. Furthermore, for the correctness parameter, it was found that for knowledge graph completion data from Google Scholar, in the data from papers related to The first author, 25% of the data was correct based on comparison with dblp and 81% of the data was correct based on comparison with Scopus. So, comparing paper data taken from Google Scholar is more effective with Scopus than dblp. In research field data, which was also taken from Google Scholar, 80.3% of the data was correct based on comparison with Sinta Kemdikbud. Furthermore, for conference/journal description data taken from the web scraping process, each site obtained 53.9% correct data based on comparing the description data with the name of the conference/journal using OpenAI. Retrieving data by scraping directly to the conference/journal site is still not optimal because it is based on the html tags in the scraping results. For the process of eliminating null data using GP2, it was found that 100% of the null data was successfully removed. For the future work, we are planning to incorporate machine learning methods into our knowledge graph completion to obtain more accurate data.

References

[1] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, *Knowledge Graphs: Opportunities and Challenges*, vol. 56, no. 11. Springer Netherlands, 2023, doi: 10.1007/s10462-023-10465-9.

[2] S. Choudhary, T. Luthra, A. Mittal, and R. Singh, "A Survey of Knowledge Graph Embedding and Their Applications," p. 11, 2021. [Online]. Available at: https://arxiv.org/abs/2107.07842.

[3] K. Wiharja, J. Z. Pan, M. J. Kollingbaum, and Y. Deng, "Schema aware iterative Knowledge Graph completion," *J. Web Semant.*, vol. 65, 2020, doi: 10.1016/j.websem.2020.100616.

[4] M. Zamini, H. Reza, and M. Rabiei, "A Review of Knowledge Graph Completion," *Inf.*, vol. 13, no. 8, pp. 1–19, 2022, doi: 10.3390/info13080396.

[5] A. Hogan *et al.*, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, 2021, doi: 10.1145/3447772.

[6] D. Fensel *et al.*, "Introduction: What Is a Knowledge Graph?," *Knowl. Graphs*, pp. 1–10, 2020, doi: 10.1007/978-3-030-37439-6_1.

[7] B. Zhang, J. Zhu, and H. Su, "Toward the third generation artificial intelligence," *Sci. China Inf. Sci.*, vol. 66, no. 2, pp. 1–19, 2023, doi: 10.1007/s11432-021-3449-x.

[8] M. Jovanovic, M. Campbell, and M. Campbell, "Connecting AI: Merging Large Language Models and Knowledge Graph," *Computer (Long. Beach. Calif).*, vol. 56, no. 11, pp. 103–108, 2023, doi: 10.1109/MC.2023.3305206.

[9] K. R. S. Wiharja, D. T. Murdiansyah, M. Z. Romdlony, T. Ramdhani, and M. R. Gandidi, "A Questions Answering System on Hadith Knowledge Graph," *J. ICT Res. Appl.*, vol. 16, no. 2, pp. 184–196, 2022, doi: 10.5614/itbj.ict.res.appl.2022.16.2.6.

[10] I. Mondal, Y. Hou, and C. Jochim, "End-to-End NLP Knowledge Graph Construction," *Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021*, pp. 1885–1895, 2021. [Online]. Available at: https://arxiv.org/abs/2106.01167.

[11] I. Muhammad, A. Kearney, C. Gamble, F. Coenen, and P. Williamson, "Open Information Extraction for Knowledge Graph Construction," *Commun. Comput. Inf. Sci.*, vol. 1285 CCIS, pp. 103–113, 2020, doi: 10.1007/978-3-030-59028-4_10.

[12] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, and H. Sack, "AI-KG: An Automatically Generated Knowledge Graph of Artificial Intelligence," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12507 LNCS, pp. 127–143, 2020, doi: 10.1007/978-3-030-62466-8_9.

[13] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.

[14] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an Architecture for Never-Ending Language Learning," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence, AAAI 2010*, 2010, pp. 1306–1313, doi: 10.1609/aaai.v24i1.7519.

[15] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, and E. Motta, "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain," *Knowledge-Based Syst.*, vol. 258, pp. 1–41, 2022, doi: 10.1016/j.knosys.2022.109945.

[16] R. Clancy, I. F. Ilyas, J. Lin, and D. R. Cheriton, "Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond," *Proc. Second Work. Fact Extr. Verif.*, pp. 39–46, 2019, doi: 10.18653/v1/D19-6607.

[17] D. Flocco *et al.*, "An Analysis of COVID-19 Knowledge Graph Construction and Applications," *Proc. - 2021 IEEE Int. Conf. Big Data, Big Data 2021*, pp. 2631–2640, 2021, doi: 10.1109/BigData52589.2021.9671479.

[18] D. Buscaldi, D. Dessì, E. Motta, F. Osborne, and D. Reforgiato Recupero, "Mining Scholarly Publications for Scientific Knowledge Graph Construction," *Eur. Semant. Web Conf.*, pp. 8–12, 2019, doi: 10.1007/978-3-030-32327-1_2.

[19] M. Nayyeri *et al.*, "Trans4E: Link prediction on scholarly knowledge graphs," *Neurocomputing*, vol. 461, pp. 530–542, 2021, doi: 10.1016/j.neucom.2021.02.100.

[20] M. Färber and L. Ao, "The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings," *Quant. Sci. Stud.*, vol. 3, no. 1, pp. 51–98, 2022, doi: 10.1162/QSS_A_00183.

[21] O. Irrerra, A. Manocci, P. Manghi, and G. Silvello, "A Novel Curated Scholarly Graph Connecting Textual and Data Publications," pp. 1-24, 2023, doi: 10.1145/3597310.

[22] F. Faheem, Z. Li, and S. Husung, "Analysis Of Potential Errors In Technical Products By Combining Knowledge Graphs With," no. September, pp. 4–8, 2023. [Online]. Available at: https://www.researchgate.net/profile/Stephan-Husung/publication/374086951.

[23] M. Färber, D. Lamprecht, J. Krause, L. Aung, and P. Haase, *SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples*, vol. 14266 LNCS. Springer Nature Switzerland, 2023, doi: 10.1007/978-3-031-47243-5_6.

[24] M. Daquino *et al.*, "The OpenCitations Data Model," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12507 LNCS, pp. 447–463, 2020, doi: 10.1007/978-3-030-62466-8_28.

[25] S. Boytcheva, A. Kiryakov, and P. Gyurov, "Similarity Search in Knowledge Graphs : Vector Space Model and Graph Embedding," *Knowledge, Lang. Model.*, no. September, pp. 20–37, 2020. [Online]. Available at: https://lml.bas.bg/~svetla/Publications/Similarity.

[26] C. Kemper, *Beginning Neo4j*. pp. 1-153, 2015, doi: 10.1007/978-1-4842-1227-1.

[27] G. Campbell, J. Romo, and D. Plump, "The Improved GP 2 Compiler," no. October, pp. 1-11, 2020. [Online]. Available at: https://arxiv.org/pdf/2010.03993.

[28] D. Fernandes and J. Bernardino, "Graph databases comparison: Allegrograph, arangoDB, infinitegraph, Neo4J, and orientDB," *DATA 2018 - Proc. 7th Int. Conf. Data Sci. Technol. Appl.*, no. Data, pp. 373–380, 2018, doi: 10.5220/0006910203730380.

[29] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge graph completion: A review," *IEEE Access*, vol. 8, pp. 192435–192456, 2020, doi: 10.1109/ACCESS.2020.3030076.

[30] S. Verma, R. Bhatia, S. Harit, and S. Batish, "Scholarly knowledge graphs through structuring scholarly communication : a review," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 1059–1095, 2023, doi: 10.1007/s40747-022-00806-6.

[31] J. Liu, J. Ren, W. Zheng, L. Chi, I. Lee, and F. Xia, "Web of Scholars: A Scholar Knowledge Graph," *Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 2153–2156, 2020, doi: 10.1145/3397271.3401405.

[32] A. Oelen, M. Y. Jaradeh, M. Stocker, and S. Auer, "Generate FAIR Literature Surveys with Scholarly Knowledge Graphs," *JCDL '20 Proc. ACM/IEEE Jt. Conf. Digit. Libr. 2020*, pp. 97–106, 2020, doi: 10.1145/3383583.3398520.

[33] M. Färber, "The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11779 LNCS, pp. 113–129, 2019, doi: 10.1007/978-3-030-30796-7_8.

[34] M. Y. Jaradeh, K. Singh, M. Stocker, and S. Auer, *Triple Classification for Scholarly Knowledge Graph Completion*, vol. 1, no. 1. Association for Computing Machinery, pp. 225 - 232, 2021, doi: 10.1145/3460210.3493582.

[35] M. Y. Jaradeh *et al.*, "Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge," *K-CAP 2019 - Proc. 10th Int. Conf. Knowl. Capture*, pp. 243–246, 2019, doi: 10.1145/3360901.3364435.

[36] E. Delgado López-Cózar, E. Orduña-Malea, and A. Martín-Martín, "Google scholar as a data source for research Assessment," *Springer Handbooks*, pp. 95–127, 2019, doi: 10.1007/978-3-030-02511-3_4.

[37] A. Martín-Martín, E. Orduna-Malea, M. Thelwall, and E. Delgado López-Cózar, "Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories," *J. Informetr.*, vol. 12, no. 4, pp. 1160–1177, 2018, doi: 10.1016/j.joi.2018.09.002.

[38] G. Campbell, B. Courtehoute, and D. Plump, "Linear-time graph algorithms in GP 2," *Leibniz Int. Proc. Informatics, LIPIcs*, vol. 139, no. 16, pp. 1–16, 2019. [Online]. Available at: https://drops.dagstuhl.de/storage/00lipics/lipics-vol139 calco2019/LIPIcs.CALCO.2019.16.

[39] C. M. Poskitt and D. Plump, "Monadic second-order incorrectness logic for GP 2," *J. Log. Algebr. Methods Program.*, vol. 130, 2023, doi: 10.1016/j.jlamp.2022.100825.

[40] X. Wang, Q. He, J. Liang, and Y. Xiao, "Language Models as Knowledge Embeddings," *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2291–2297, 2022, doi: 10.24963/ijcai.2022/318.

[41] M. Y. Jaradeh, "Thinking Outside the Graph: Scholarly Knowledge Graph Construction Leveraging Natural Language Processing," p. 189, 2022. [Online]. Available at: https://repo.uni-hannover.de/items/6cf1b1f7-49be-42a2-9394-dd7a92a8ee1d.

[42] T. Safavi and D. Koutra, "CODEX: A comprehensive knowledge graph completion Benchmark," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 8328–8350, 2020, doi: 10.18653/v1/2020.emnlp-main.669.

[43] P. Betz, C. Meilicke, and H. Stuckenschmidt, "Supervised Knowledge Aggregation for Knowledge Graph Completion," *Eur. Semant. Web Conf.*, pp. 74–92, 2022, doi: 10.1007/978-3-031-06981-9_5.

[44] A. Khatiwada, S. Shirai, K. Srinivas, and O. Hassanzadeh, "Knowledge Graph Embeddings for Causal Relation Prediction," *CEUR Workshop Proc.*, vol. 3342, pp. 1-12, 2022. [Online]. Available at: https://ceur-ws.org/Vol-3342/paper-8.pdf.

[45] K. Liang *et al.*, "A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multimodal," pp. 1–20, 2022. [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/10577554.

[46] M. Nayyeri, S. Vahdati, J. Lehmann, and H. S. Yazdi, "Soft Marginal TransE for Scholarly Knowledge Graph Completion," pp. 1-10, 2019. [Online]. Available at: https://arxiv.org/abs/1904.12211.

[47] M. Nayyeri, C. Xu, Y. Yaghoobzadeh, H. S. Yazdi, and J. Lehmann, "On the Knowledge Graph Completion Using Translation Based Embedding: The Loss Is as Important as the Score," 2019. [Online]. Available at: https://www.researchgate.net/publication/335599514.

[48] A. Bordes, N. Usunier, and A. Garcia-Durant, "Translating Embeddings for Modeling Multi-relational Data," *Adv. Neural Inf. Process. Syst.*, 2013, doi: 10.1109/NAVITEC.2014.7045139.

[49] M. Nayyeri, C. Xu, Y. Yaghoobzadeh, H. S. Yazdi, and J. Lehmann, "Towards Understanding The Effect Of Loss Function On The Performance Of Knowledge Graph Embedding," pp. 1-14, 2019. [Online]. Available at: https://arxiv.org/abs/1909.00519.

[50] Ontotext, "Ontotext Refine." [Online]. Available at: https://www.ontotext.com/products/ontotext-refine/.

[51] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Capturing Semantic Relationships in Electronic Health Records Using Knowledge Graphs: An Implementation Using MIMIC III Dataset and GraphDB," *Healthc.*, vol. 11, no. 12, 2023, doi: 10.3390/healthcare11121762.

[52] F. Wang, A. Bundy, and X. Li, "Schema-aware Iterative Completion for Knowledge Graphs Revisited," *WWW J.*, pp. 1-64, 2023. [Online]. Available at: https://knowledge-representation.org/j.z.pan/pub/SICKLE2023.pdf.

[53] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semant. Web*, vol. 8, no. 3, pp. 489–508, 2017, doi: 10.3233/SW-160218.